International Series in
Operations Research & Management Science

Bruce W. Patty *Editor*

Handbook of Operations Research Applications at Railroads

Springer

International Series in Operations Research & Management Science

Volume 222

Series Editor

Camille C. Price
Stephen F. Austin State University, TX, USA

Associate Series Editor

Joe Zhu
Worcester Polytechnic Institute, MA, USA

Founding Series Editor

Frederick S. Hillier
Stanford University, CA, USA

More information about this series at http://www.springer.com/series/6161

Bruce W. Patty
Editor

Handbook of Operations Research Applications at Railroads

Springer

Editor
Bruce W. Patty
Veritec Solutions
San Rafael, CA, USA

ISSN 0884-8289 ISSN 2214-7934 (electronic)
International Series in Operations Research & Management Science
ISBN 978-1-4899-7570-6 ISBN 978-1-4899-7571-3 (eBook)
DOI 10.1007/978-1-4899-7571-3

Library of Congress Control Number: 2015933378

Springer New York Heidelberg Dordrecht London
© Springer Science+Business Media New York 2015
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

This book is dedicated to my beautiful wife, Paula, whose courage, intelligence and determination inspire me every day. Without her support and encouragement, this book would still be a dream for me.

Bruce W. Patty, Editor

Preface

I am very pleased and proud to have been able to serve as the Editor for this book. I first started working on railroad problems over 20 years ago while a member of American Airlines Decision Technologies, a consulting group within American Airlines. I quickly learned how operationally complex railroad problems are compared to airline problems. For example, passengers can transport themselves from one gate to another when they need to make a connection between flights. To connect a railcar from one inbound train to an outbound train requires several steps involving many people, tracks, and locomotives. Airlines rarely have to be concerned about the capacity *between* two airports. Railroads do not have the luxury of traveling *over* another train moving between the same two terminals. However, there are some areas where railroads have a definite advantage. Airlines would love to have the ability to add another engine to a plane, allowing it to carry more passengers. They also would love to have the ability to have two planes, each with three engines arrive at a terminal, and then have two planes depart the terminal, one with four engines and one with two. As I continued to work with railroads, I became more and more engaged with figuring out how to apply Operations Research models and approaches to address these complex rail problems.

The topics covered by the chapters in this publication have been specifically selected to give readers the complete spectrum of the role that Operations Research has played and can play in the improvement of North American freight railroads. Not only have the topics been specifically chosen to provide this spectrum, but the authors of the chapters are recognized award-winning scholars and practitioners with a deep knowledge and understanding of their specific topics. The chapters have been written in a diverse manner so that readers who are looking for an understanding of how decisions are made at railroads will find what they are looking for, as will readers who are looking for examples of mathematical programming formulations to complex problems.

The team of Carl Van Dyke and Dr. Marc Meketon have authored three chapters: Train Scheduling, Car Scheduling and Railway Blocking Process, and teamed with me on a fourth, Network Analysis and Simulation. Carl and Marc have worked with railroads for over three decades to apply Operations Research tools, most recently while at Oliver Wyman. The tool they developed, Multi Rail, is in use at all of the Class I railroads in North America. Their work was recognized in 2003 by INFORMS as key members of the Franz Edelman prize-winning team for the work at Canadian Pacific Railway. Carl was recently awarded the 2014 Distinguished Member award of the Railway Applications Section of INFORMS. Their chapters provide readers with deep insights as to how decisions are made at railroads, the kinds of tools used, and the IT challenges that must be overcome.

The team of Ravindra Ahuja and Bala Vaidyanathan have authored two chapters: Locomotive Scheduling and Crew Scheduling. Dr. Ahuja is a renowned optimization expert, having won the Koopman Prize and the Lanchester Prize, and he has been named an INFORMS Fellow. He and his colleagues at Optym, formerly Innovative Scheduling, have applied his optimization approaches to problems at railroads for several years, including CSX Transportation, BNSF, and Norfolk Southern. Their chapters provide a detailed look at mathematical programming formulations and solution approaches to their topics.

Dr. Michael Gorman, currently a Professor at the University of Dayton, has authored two chapters: Empty Railcar Distribution and Pricing/Revenue Management. Dr. Gorman is especially well qualified to expound on these two topics. His work in the area of Empty Railcar Distribution for CSX Transportation resulted in their team being named Finalists for the Franz Edelman Award. His work in the area of Pricing for the Hub Group resulted in their team being named a Finalist for the Wagner Prize. Both of these competitions are focused on application and practice of Operations Research. Dr. Gorman has also received awards for his teaching of Operations Research. Prior to joining the faculty at the University of Dayton, Dr. Gorman led the Operations Research groups at Santa Fe Railroad and then BNSF after the merger. His chapters provide a combination of mathematical and algorithmic insights as well as insights into real-world applications.

Roger W. Baugher has worked to apply Operations Research tools to railroads for over 40 years. He has authored two chapters: Simulation of Line of Road Operations and Terminal Simulation. He was instrumental in the development of Algorithmic Blocking and Classification (ABC) while at Norfolk Southern and has also worked at BNSF. He was the first recipient of the RASIG Award for his contributions to OR in the railroad industry. Mr. Baugher has also contributed to the book, "The Railroad, What it Is, What it Does." His chapters provide significant insights into rail and terminal operations.

I have applied my understanding of network optimization approaches to various forms of transportation, including airlines, freight railroads, and intermodal. My chapter on Intermodal Rail is based on insights from my years of working in the rail industry, especially those spent in Intermodal. I helped to found the Railroad Applications Special Interest Group (RASIG) and have worked to apply Operations Research techniques at railroads for over 20 years as both a railroad employee and a consultant. I am a Franz Edelman Laureate and served on the INFORMS Board of Directors. While at American Airlines, I led the project team for the work done at Conrail during which time the Conrail Network Analysis Model (CNAM) was developed. I later spent 7 years at Pacer Stacktrain as AVP of Equipment Strategy during a time when Pacer Stacktrain had the largest domestic intermodal container fleet in the United States, pioneering the use of sophisticated analytical approaches to chassis management.

We all hope that you enjoy the book and that it provides you with insights regarding the application of Operations Research at Freight Railroads!

San Rafael, CA, USA Bruce W. Patty

Contents

Chapter 1
Train Scheduling

Carl Van Dyke, Marc Meketon, and Problem Solving Competition Committee

1.1 Introduction and Background

In traditional railroad operations, sets of railcars are grouped together on a temporary basis into blocks. These blocks are moved by trains, where each train may carry a single block, or may carry multiple blocks. In this manner the cars are relayed from their origin to their destination by being placed in a series of blocks, which are moved by a series of trains. This overall process is often called trip planning or car scheduling and is described in a separate Chap. 4. Blocking is the grouping of cars that may have disparate origins and destinations, but will be moved together from one point to another before being broken apart and formed into another block. See the separate Chap. 5 on the blocking problem for further discussion of this topic. See Ireland et al. (2004) for one perspective of all of the components of the operating plan design problem.

This chapter focuses on the role of the train schedules, and describes the data elements making up a train schedule, the process of designing the train schedules, and managing these schedules on a real-time basis. This chapter provides the definitions for the following OR train design problems:

C. Van Dyke (✉)
TransNetOpt, Princeton, NJ, USA
e-mail: carl@cvdzone.com

M. Meketon
Oliver Wyman, Princeton, NJ, USA
e-mail: marc.meketon@oliverwyman.com

Problem Solving Competition Committee, Railway Applications Section, INFORMS

© Springer Science+Business Media New York 2015

B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3_1

- *Train routing*: how best to generate the routes of each train such that all blocks will be moved, and total train miles will be minimized. Minimizing total train miles also tends to maximize train size subject to a requirement that minimum train frequencies be observed.
- *Block-to-train assignments*: which blocks will be placed on each train, minimizing overall train complexity and the need to swap blocks en-route from one train to another.
- *Train timing and connections*: setting the timing of each train such that the overall transit times for all shipments will be minimized, taking into account the connections of railcars from one train to another, and the associated minimum processing times for such connections. Timing must also take into account the effective numbers of trains per hour that can be processed at each yard and can travel over each line segment.

1.2 Role of Trains in the Railroad Operations Research Landscape

Along with the blocking plan, train design plays one of the most critical roles in determining the operational efficiency and effectiveness of a railroad. These roles include:

- *System costs*: a significant amount of the operating cost is driven by the train design. Minimizing the total number of trains operated, while maximizing their velocity, tends to minimize overall costs through maximizing use of available line capacity, minimizing crew requirements, and minimizing locomotive requirements. Train design can also impact fuel requirements, often the single largest expense for a railroad. However, minimizing the number of trains can result in excessive dwell time for railcars, which can have a countervailing impact on system costs and customer service. Other elements of the train design that can impact system costs include:
  - *Circuity*: in some cases multiple route choices exist for a train. For various reasons trains may use the less direct routes of the options available, causing some increase in railcar circuity, and driving up costs related to distance traveled (crews, fuel, locomotives, asset velocity-related costs). This is done for a number of reasons, including managing the capacity utilization on each of the available routes, and a need to provide service to specific intermediate locations.
  - *Balance*: this is the idea that the number of trains operated in each direction over a line or between yard pairs should be the same. Balance ensures that equal capacity to move railcars exists in each direction, and ensures that crews and locomotives have a natural flow that keeps them in balance and minimizes deadheads. It is often a specific goal of the train plan to be balanced both overall and by train type.

- *Line capacity*: each rail line has a finite capacity to handle trains. This capacity is determined both by the physical characteristics of the line and by the trains that are designed to traverse the line. The impact of the trains comes from the mix of trains to be operated (long versus short trains, fast versus slow trains, etc.), the total number of trains to be operated, and any peaking in the number of trains. Other influencers include issues such as the use of "fleeting" to operate many trains in a single direction over a line, and the operation of over length trains that cannot fit in all of the passing sidings. See the separate Chapter 3 on line capacity modeling for further discussion on this topic.
- *System capacity*: each train has a limit as to how many railcars it can transport. This limit can be determined by the pulling power of the available locomotives, and by the characteristics of the line (length of passing sidings, constraints on train length due to grades, etc.). The total number of trains in the design traversing each line determines the total carrying capacity of the plan, and how much spare capacity exists to handle peaks. While extra trains can be operated, these tend to be disruptive to operations, and thus not desirable. Thus, the effective throughput of the plan is determined by the overall train plan design. See Chapter 4 on car scheduling and Chap. 8 on simulation for a more detailed discussion of train capacities and their role in plan evaluation.
- *Crew requirements*: in North America, each freight train that operates represents at least one crew job. For longer distance trains, multiple crews may be required to advance the train across the network. Thus, the total number of trains that operate, and their relationship to where the crew bases are located, directly impacts crew requirements. See the separate Chapter 6 on crew requirements for further discussion of this topic.
- *Locomotive requirements*: as with crews, the design of the trains can directly impact locomotive requirements. Key drivers include the number of trains to be operated, the specific locomotive type requirements for each train, the expected performance characteristics of each train (power to weight ratio requirements), the overall balance of the trains by direction, total distance travelled and transit times for the trains, and the timing of trains relative to the required time for locomotives to connect from one train to another. See the separate chapter on locomotive planning for further discussion of this topic.
- *Yard requirements/balance*: most yards have a limited capacity to handle inbound trains and makeup outbound trains. If the train plan tries to arrive or depart too many trains in a short period of time, this can overload the yard or drive up costs in order to have the capacity to handle the peak. As a consequence, a design goal is to ensure relatively even patterns of train arrival and departure times. See the Chapter 9 on terminal simulation for more information on this topic.
- *Service levels*: the train schedules impact service in a number of ways. The speed of trains directly impacts the time railcars spend moving from one location to another. The train design impacts velocity through the number of intermediate work events each train undergoes, and the overall design of the train in terms of its physical performance (power to weight ratio, handling of speed restricted railcars, etc.). If multiple routes exist, then the route choice also impacts speed.

The frequency with which each block is handled directly impacts the average time railcars spend in yards (a block that has two departures per day will yield lower yard dwell times than a block that departs only once per day). The timing of the trains, and the number of times per week each train operates, also determines the connection times for railcars at yards, and thus the overall transit time for the railcars. See Chapter 4 on car scheduling for further details on the determinants of shipment transit times. The service levels have a direct impact on railcar requirements:

  - *Railcar velocity/fleet size*: transit time or velocity ultimately translates into total cycle times for railcars, which directly determines fleet size requirements. See the Chapter 8 on simulation for a discussion on how to estimate railcar fleet requirements based on an operating plan.

- *Reliability*: train plan design influences railcar transit reliability in two major ways. One impact is on the reliability of the individual trains to achieve their designed schedules. The other is on the consequences of connection failures at yards. While many factors impact the achievability of train schedules, the most critical design factors are ensuring that the train design does not overly tax the capacity of the lines that trains traverse, and minimizing the complexity of any en-route work that a train must do (including connections with other trains to swap blocks). When a railcar misses its planned connection, the length of time it must wait for the next train directly impacts its transit time reliability. For example, the train design can determine if this railcar has only one movement opportunity per day, or more than one such opportunity. See Chapter 3 on line capacity simulation for a discussion on how to determine schedule achievability, and the Chapter 4 on car scheduling for a detailed discussion of the role of dwell times and train connections in the determination of shipment transit times.

Ideally, each of these considerations should be factored into the train design process, and into any optimization or heuristic process for the design of a train operating plan. In general, this problem is treated as a cost minimization problem, not a profit maximization problem. This is because in most formulations, the traffic to be moved (and its associated revenue) is treated as a fixed constraint. That is, the solution must move all of the traffic specified in the traffic database for the design period. Given this constraint, with a fixed traffic database and thus a fixed amount of revenue, the minimization of costs becomes the same as profit maximization. This assumption also can result in constraints on minimum service requirements, with the implication that a failure to achieve these service constraints could result in a loss of traffic and/or revenue. It is the author's understanding that in the short term, railroad shipment volumes are relatively inelastic to both price and service, while in the long term there may be greater elasticity through modal shifts, sourcing changes, and carrier substitutions. However, such relationships do not appear well enough understood to incorporate in current train design processes, and thus the design process is treated as a cost minimization problem, subject to service constraints.

1.3 Types of Trains and Related Definitions

Trains are generally broken into several types:

- *Road trains*: these are the "classic" definition of a longer haul train. In general they carry traffic between a pair of yards, perhaps picking up or setting off blocks of railcars at a small number of intermediate stations. They generally do not directly serve customers, but instead only serve yards of various sizes where cars are processed and formed into blocks. These trains typically handle general merchandise traffic, but also include specialized trains such as intermodal or automotive trains.
- *Unit trains*: these trains typically carry a single block of traffic directly from a single customer origin, and deliver directly to a single customer at destination. From a definitional perspective they look much the same as a road train, except that they have no intermediate pick-ups or set-offs of railcars, and carry only a single block. Unit trains have more flexibility in the routes they can take, and can change the exact route on a day-to-day basis if parts of the network are congested.
- *Local trains*: these trains provide direct service to customers, placing cars on customer sidings, and picking up cars from these sidings. Locals come in many flavors including trains that serve only a small area, trains that start and end at the same terminal while traversing a significant distance (turn trains), and trains that start at one terminal and end at another (through locals). Locals can also carry through blocks of the same sort as those carried by road trains, and of course, some road trains can do small amounts of local service.

There are a number of key definitions that need to be understood before we discuss the specification of train schedules in detail (see the Chapter 4 on car scheduling for further information on a number of these definitions):

- *Block*: A block is a group of cars that may have disparate origins and destinations, but will be moved together as a group from a common assembly point to a common disassembly point. At the disassembly point the block will be broken apart and the railcars will be formed into new blocks along with other railcars arriving from other locations. Thus, for an individual railcar, the origin and destination of a block may be either the same as the ultimate origin or destination of the railcar, or may be intermediate points in the railcar's route where the car is to be marshaled.
- *Yard-blocks/train-blocks*: Perhaps for historic reasons, most blocking systems do not provide a definition of a car to yard-block assignment in terms of a block origin, destination, and block name. Instead, they provide a "yard-block code," which is variously referred to as a "tag" or "class code." In most systems, trains specify a separate concept called a "train-block" that provides the pick-up location for the train-block, the set-off location, and a train-block name. Yard-blocks (class codes/tags) are then associated with the train-block. More than one yard-block can be assigned to the same train-block. This is done to provide visibility

to subsets of the traffic in a train-block (both codes are displayed by most systems), and to allow sets of traffic to be easily shifted from one train or destination yard to another for capacity management purposes. Since the yard-blocks (class codes) do not have a destination, the destination becomes the location where the train-block is set-off. On the one hand, this makes it very hard to validate that appropriate class codes have been assigned to a particular train-block; on the other hand, it also provides flexibility to send the same class code/yard-block to different locations by day-of-week or based on other factors related to the available train service. See Chapter 5 on blocking and Chapter 4 on car scheduling for further discussion of this topic.

- *Block swaps*: A block swap is defined as the movement of a group of cars (a block) from one train to another on an intact basis without intermediate classification. For example, if a block is made at A, destined to C, but the train sets off this block at B instead, for pick-up by a second train, the activity at B is called a block swap. The benefit of a block swap is that it can help reduce intermediate switching work at a yard and the associated delays, but it can also create:

  – More complex train operations
  – A potential loss of capacity for the line or yard where the swap occurs
  – Additional delays and costs at the block swap location

- *Connections*: when shipments (railcars) move from one train to another, this is called a connection. In most cases the cars making a connection at a yard come from a variety of sources such as local originations, other inbound trains, and in some cases from other railroads. These cars then must be processed (switched or marshaled) and placed into an appropriate outbound block, which is then placed into an outbound train. The connection process is driven by the blocking plan (see Chapter 5). Typically, a minimum processing time is specified for a connection at a yard. Cars can only connect to outbound trains that depart after this minimum processing time has elapsed.
- *Pick-ups/Set-offs*: a pick-up is the placement of a block of cars into a train. A set-off is the removal of a block of cars from a train. The blocks on a train are often ordered to minimize the amount of work that is required to perform a pick-up or set-off by minimizing the number of places along the length of a train that must broken to insert or remove blocks from the train. Further, in some cases blocks are picked-up at intermediate points that have the same characteristics as a block already on the train. Such blocks are typically merged as part of the pick-up process.
- *Work events*: The act of picking-up or setting off blocks at an intermediate point in a train route is called a (intermediate) work event. Work events represent an overall activity of the train, and thus the number of work events for a train does not change if more than one block is picked-up or set-off at the same route location. Work events are important not only because they represent time delays for the train and switching work that must be performed, but also because they represent the consumption of network capacity. The consumption of network capacity for a work event can be different than for a train origination or termination because the train must be kept intact and thus may need to use different tracks at a location than would be used by originating or terminating trains.

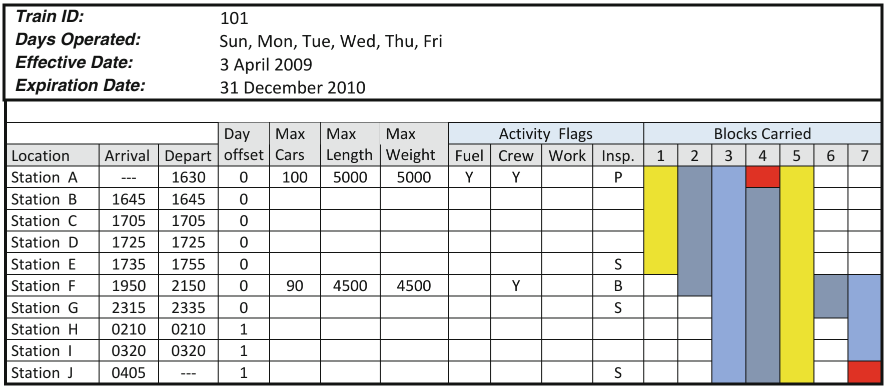
- *Crew segments/districts*: train crews are assigned based on specific rules that are a function of both labor agreements and safety rules. The safety rules relate to maximum work and rest times requirements for crews, and the need for a crew to be qualified to operate over a specific line. Qualification typically means that the crew is familiar with a line's physical characteristics and operating rules, where such familiarity is achieved through a structured training process. The result is that a particular crew will only be qualified to operate over specific parts of a network. To manage this process, railroads are typically broken into a set of crew segments or districts, where crews hold qualifications to operate over the rail lines associated with a specific segment or district. On a North American freight railroad, operating a train over a single segment typically represents a full day's work. Some trains may go faster than others, and thus use longer segments. Most crews are based at a specific location, and work one or more segments originating from that location. They typically work a train outbound from their home location on the first day of a duty cycle, rest for 8-24 hours at the "away" location, and then work a train back to their home location on the second day of a duty cycle. While it is easiest to think of a crew segment as a pair of locations (home and away terminal), in practice each end of a segment can be a cluster of stations.

1.4 Specifying Road Trains

Each train has a route, timing information, and may carry a number of blocks. For each block the pick-up location, set-off location, and block attributes are specified. Thus, a great deal of information can be contained within the specification of each road train, which includes the following core elements:

- *Overall train attributes*: This typically includes the train symbol, the days operated, effective/expiration dates for the schedule, the train type, and whether the train is a regularly scheduled train or an "as-required" train. Beyond this, a variety of other information may be present such as locomotive requirements in terms of both unit types/count and power to weight ratios, operating divisions responsible for the train, train size limits, train notes, special instructions, etc.
- *Train route*: The train route specifies the locations (stations) the train will pass through, the arrival and departure times for each location, and any required dwell times. Not every station is included in the main train route, so in some cases there is additional information listing the more detailed stations in the route. A great deal of other information may be found in the route such as crew changes, en-route inspection indicators, work location designations, fueling locations, size limits for the train in terms of weight, length, or railcars, changes in the power to weight ratio or locomotive requirements, etc. A common decomposition of the train design problem is to generate the train routes first, ensuring that there is both the necessary coverage to move all of the traffic, and sufficient capacity. Block-to-train assignments (see below) are then used to fill out these trains. In some cases the routing process may be driven by the existence of specific "anchor blocks" that are identified by the user as forming the foundation of specific trains.

- *Block-to-train assignments*: The block-to-train assignments specify each train-block in terms of its name, where it will be picked-up, and where it will be set-off. This information can also include weight and length limits for each block, whether the block is a primary block or a fill block, the standing order of the blocks in the train, and in some cases the connecting train for block swaps. This information can also specify if a block picked-up at an intermediate route location should be merged with a block that is already on the train. At many railroads, there is a second part to the specifications detailing the yard-block to train-block assignments. This is typically simply a list of yard-block codes or class codes that the train-block is to be composed of. In some cases, to support local blocking, station ranges may be associated with the train-block or yard-block—this idea is discussed below in Section 1.9 on local services.



Train ID:	101
Days Operated:	Sun, Mon, Tue, Wed, Thu, Fri
Effective Date:	3 April 2009
Expiration Date:	31 December 2010

			Day	Max	Max	Max	Activity Flags				Blocks Carried						
Location	Arrival	Depart	offset	Cars	Length	Weight	Fuel	Crew	Work	Insp.	1	2	3	4	5	6	7
Station A	---	1630	0	100	5000	5000	Y	Y		P	🟨	🟦	🟦	🟥	🟨		
Station B	1645	1645	0								🟨	🟦	🟦	🟦	🟨		
Station C	1705	1705	0								🟨	🟦	🟦	🟦	🟨		
Station D	1725	1725	0								🟨	🟦	🟦	🟦	🟨		
Station E	1735	1755	0							S	🟨	🟦	🟦	🟦	🟨		
Station F	1950	2150	0	90	4500	4500		Y		B		🟦	🟦	🟦	🟨	🟦	🟦
Station G	2315	2335	0							S			🟦	🟦	🟨		🟦
Station H	0210	0210	1										🟦	🟦	🟨		🟦
Station I	0320	0320	1										🟦	🟦	🟨		🟦
Station J	0405	---	1							S			🟦	🟦	🟨		🟥

Representative Train Schedule with Block Display
(yellow and blue colors represent different block categories, red represents a block swap)

Representative Train Schedule with Block Display (yellow and blue colors represent different block categories, red represents a block swap)

- *Connection standards or cut-offs*: At most railroads the connection standards or cut-offs specify the timing rules for cars connecting to the train. The role of the connection times is discussed in detail in Chapter 4 on car scheduling, but can be summarized as specifying the minimum time allowance required for a railcar to successfully connect from a specific inbound train to a specific outbound train. While these connection standards can be specified at a location level, many railroads also specify these connection times by inbound or outbound train, or at the route or train-block level of each train. As a result, each train may own one or more connection standards that play a critical role in the car scheduling process. The standards consist of the cut-off time and generally seven optional data elements: the in-bound train, in-bound train-block, the in-bound yard-block, the out-bound train, the out-bound train-block, the out-bound yard-block, and the current location. The most commonly used optional elements are the specific out-bound train and either outbound train-block or route location. The cut-off is either an elapsed time before the train departs the location, or a specific clock time. The elapsed time is converted to a clock time by subtracting the elapsed time from the departure time of the train. In either case, to connect to a specific train a car must arrive at the yard earlier than the cut-off time when expressed as a clock time. Some railroads also specify specific connection types, and restrict the connection standard to apply only to a specific type. Typical connection types are regular classifications, to/from industry, and to/from interchange. While important when managing the detailed car scheduling processes, and used in a number of simulation type models, these connection standards are typically replaced by global or location-specific connection times in most optimization type models.

- *Capacities*: The capacities of trains and train-blocks are typically expressed in terms of a maximum weight and length for the train or train-block, and are important to understand when assessing if an overall train plan will be feasible in moving the available traffic. As a result, overall train capacity must be considered in any optimization solution, and is often taken into account in simulations. While such capacities are often considered to be a soft constraint, they nonetheless are real, and need to be understood. In general, they exist at two levels within the train specification. One is at the overall route location leve and the other is by individual train-block. The overall train capacity is typically a function of the physical characteristics of the line being traversed and the make-up parameters for the train (number of locomotives assigned, design of the cars being moved, use of mid-train power, etc.). The capacity by train-block is used to manage the allocation of space on the train to different blocks, ensuring that the needs of all of the customers assigned to the train are managed in a structured way that protects both operational needs and customer service commitments. For example, consider a train that has a route of A–B–C, which picks-up an A–C block at A, and a B–C block at B. The train design might limit the size of the A–C block in order to ensure that sufficient space is available to protect the B–C block. See Chapter 4 on car scheduling and Chapter 8 on network simulation for a more extensive discussion of specifying train capacities.

There are a number of complexities and special considerations that must be taken into account when designing a train plan. Some of the key ones are described below.

- *Fill blocks, extras, and annulments*: Most railroads support the designation of block-to-train assignments as either primary blocks or fill blocks. The concept behind a fill block is that it will only be used if the train is below capacity after first being loaded with its preferred traffic. See Chapter 4 on car scheduling for further discussion on this topic. Field operations may also add extra trains or annul trains. An extra train is typically a train that was not in the base plan, but is needed to carry excess traffic due to a peak in volume. Annulling is the act of cancelling a train, which may be done due to operational problems such as the lack of locomotives or crews, or for tactical reasons such as insufficient traffic for the train. When this happens, the date-specific train database used by the car scheduling system is updated to reflect these actions. While annulments will always be reflected in the updated trip plans, use of extras will depend on how

they are designated, and if their block-to-train assignments are designated as primary or fill. Capacitated simulation models often take advantage of fill blocks as well.
- *Interchange blocks/run-through trains*: Railroads often enter into agreements with other railroads to build blocks for each other (called "pre-blocks"), and in some cases to operate "run-through" trains with the other railroad. Run-through trains are cases where entire single or multi-block trains are created and passed to the other railroad on an intact basis. In some cases, special logic is required to specify these trains since their routes extend off of the railroad's home network. This topic is discussed in more detail in Chapter 5 on blocking and Chapter 4 on car scheduling. Because the design of such trains are negotiated between pairs of railroads, they are generally considered fixed, and either not allowed to be changed by train plan optimizers, or only allowed to be changed in very limited ways.

Beyond the above, many other data elements may be found in the specification of a train schedule such as:

- Locomotive requirements and assignments
- Crew assignments
- Consist details (cars assigned to the train)
- Information required by specialized trains, such as intermodal trains

We will not explore these additional data elements in this chapter. See Chapter 2 on locomotive planning and Chapter 6 on crew planning for more information on these topics.

1.5 OR Challenges: Designing the Road Train Plan

In an idealized world, one would attempt to optimize the train plan and the blocking plan at the same time, while also optimizing the crew and locomotive plans. All of this would be done in a manner to also optimize the velocity and handling costs of the railcars, ensure even and feasible workloads at each yard, and that sufficient line capacity was available to handle the proposed trains.

In the current state of the art, this holistic problem is generally decomposed into a number of separate sub-problems:

- Blocking plan optimization (see Chapter 5 on blocking)
- Crew planning/optimization (see Chapter 6 on crews)
- Locomotive planning/optimization (see Chapter 2 on locomotives)
- Train scheduling (largely holding blocking plan as fixed and treating locomotives and crews as dependent sub-problems)

As part of addressing the train scheduling problem, one also needs to take line capacity into account. While some solutions attempt to do this by developing a line-specific slot plan as part of the train scheduling process, in our discussion we will assume that the most common practice of setting limits on the number of trains that can be operated over a line during a specific period of time will be sufficient for developing the base train plan. This train plan is then adjusted using a line capacity model as a separate exercise. See Chapter 3 on line capacity for more information on the interactions between line capacity and train scheduling.

As discussed earlier, there are a number of different types of trains, such as road trains, unit trains, and local trains. This chapter's discussion on train design algorithms will focus on road trains. The authors are not aware of any significant work with respect to algorithms for generating local train plans, and this issue will not be addressed here.

The base unit train problem is fairly straight forward in the case of shuttle train type operations where the train sets are kept intact, and will not be addressed here. Unit train planning/optimization tends to focus heavily on the cycling plans for the train sets as a driver of total throughput and fleet size. In the real-time environment, the problem statement tends to focus on order management in the deployment of the train sets against the traffic volumes that must be moved.

Other unit train plan design problems exist that are of higher complexity. One example is the grain train scheduling problem, where sets of cars representing between 25 and 100 % of a full train are released from grain elevators, and must be combined into full trains for movement to ports or other unloading points. This class of problem is largely handled manually at present, but might lend itself to the use of a real-time scheduling algorithm.

See Section 1.11 on opportunities below for further discussion of unit train scheduling issues.

1.5.1 Road Train Design Problem

The road train design problem is often decomposed into three sub-problems:

- Train route design
- Block-to-train assignment
- Train scheduling or timing (including frequency)

The train route design and block-to-train assignment problems are described in Section 1.6, and the characteristics of the train scheduling (timing) problem are described in the subsequent section.

1.5.2 Single Versus Multi-Block Trains

It is important to note that there are a number of business policies, and operating practices that can factor into the design of the train plan, and as a result may need to be incorporated into any OR solution to the design problem. Perhaps the two most important such factors are the use of anchor blocks and restrictions on the number of blocks a train can carry.

Trains can carry one or more blocks. As the number of blocks increases, the complexity of operating the train also increases. Furthermore, the more blocks one makes, the smaller the blocks tend to be in size, which means that it takes more blocks to fill out a train to its logical limits of length and weight. However, making more blocks avoids intermediate handlings, so this may be worth it in a trade-off against train complexity, particularly where the delays associated with handlings are long.

The longer trains become (i.e. the more cars that are carried), the more likely it is that trains will have multiple blocks. In short train environments, such as Europe where trains are often only 20–40 railcars long, single block trains can make much more sense. The authors have seen single train operations in other settings as well, even with fairly long train lengths. This typically happens where the number of smaller long distance blocks is limited, and instead blocks are primarily made only as far as the next major yard. This tends to drive up block size, as well as the number of intermediate handlings. However, it also may permit the operation of multiple trains per day to carry each block, which can act as a countervailing force by driving down the delays associated with each handling. For example, in Europe, dwell times in yards can be as little as ±6 hours due to expeditious handlings, and multiple departures per day for each block, compared to times of ±24 hours at large North American rail yards with only one departure per day for each block.

The end result is that some railways design their train plans so that most of their road trains are hub-to-hub with no intermediate stops. They tend to have many major yards (hubs) and run trains between consecutive hubs. The hub-to-hub trains have a single block, and at the termination of the train the cars are switched to other outbound trains. Even if most cars on the train are meant for a set of destinations a thousand miles away, the cars would still be switched several times en-route to their destination.

In some railways, these single-block trains do not have a schedule—rather they run whenever they reach their maximum length or weight. This creates long trains that on the surface seem to be very efficient by reducing the number of trains operated. However, this also tends to make efficient use of locomotives or crews difficult due to the randomness of train departure times and the number of trains operated. It may also drive up overall transit times for railcars as well, increasing the total amount of equipment needed to operate the railroad.

The alternative to this single block strategy is to allow multi-block trains. One methodology that is often employed in the design of multi-block trains is to drive the process using “anchor blocks.” An anchor block represents a key block that is the foundation for the operation of the train. An anchor block is typically a block that carries critical shipments from a volume or customer perspective. However, the anchor block may not be large enough to fill out the train, and thus using only the anchor block the train may not reach its limits on length and weight. As a consequence, other blocks are assigned to the train to “fill it out” to the limits of its carrying capacity.

The building of a train using an anchor block as the starting point has challenges. The additional blocks added to the train may not be a perfect fit to the anchor block—that is their origin/destination may not be on the same route that the train would take if it only carried the anchor block, and the additional blocks may delay the train as they are picked-up and set-off. In some cases to accommodate the additional blocks, the train's route may need to be extended to include a different origin or destination. Multi-block trains also tend to introduce work events that may be disruptive to other trains if these events tie-up the mainline, especially if they are setting out blocks for a block swap.

1.6 Train Routing/Block-to-Train Assignment Problems

The Railroad Application Section of INFORMS sponsors an annual problem solving competition, which in 2011 focused on the train route design and block-to-train assignment problems. The following is largely a slightly modified extract of the problem description provided for the 2011 competition (Railroad Applications Section 2012).

While the freight railroad industry has been in existence for over a century, the fundamental concept of aggregating freight railcars based on different attributes to create blocks and subsequently combining blocks to create trains has not changed. Freight railroads receive requests from customers to transport cars. Upon receiving the request, based on each car's attributes (such as physical dimensions, freight type, etc.), the railway generates a trip plan detailing the movement of the car from the customer's origin location to the requisite final destination.

Train routing design includes identifying the origin, destination and route for each individual train, such that these routings are consistent with the rail network and the blocks to be transported. Along its route, a train can visit different yards to either (a) pick-up block(s), (b) set-off block(s), or (c) both set-off and pickup blocks. Both the train routes and the block-to-train assignments are generally designed in advance of it being operated, and the plan is then followed and adjusted as necessary during actual operation.

In this problem description it is assumed that the blocks made at each of the yards have already been determined and cannot be changed. Hence, the block attributes such as origin, destination, number of cars, length and tonnage is treated as a fixed input to the process.

Thus, this problem description will focus on Block-To-Train Assignment (BTA) and Train Routing, which will be collectively referred to as "Train Design."

Train design is one of the most fundamental and difficult problems encountered in the railroad industry. A Class I railroad can operate around 200 merchandise or road trains per day (excluding locals), which follow a predetermined schedule. These trains can transport close to 1,000 blocks by picking up or setting off blocks at 180–200 locations. Approximately, 400–500 crews are involved in moving the merchandise trains between corresponding origin and destination locations.

This problem has huge potential for benefiting from the application of Operations Research. Identifying the optimal routes for the trains, and associated block-to-train assignments, subject to different capacity and operational constraints, is called Train Design Optimization. Operational and capacity constraints involved in this problem include:

(a) *Blocks per train:* A train is constrained by the maximum number of blocks it can carry. Assigning too many blocks to a train can result in too complex a train, which increases the chances for errors (impacting reliability), and increases the time and yard capacity required to make up the train at origin, and switch it at intermediate points. It also can increase the number of work events (see below).

(b) *Block swaps per block:* Each block is constrained by the maximum number of times it can be block swapped. Even though theoretically block swaps are more efficient than a classification event, from a practical perspective they require additional time and resources, introduce dwell time, and increase the chances of an operational failure.

(c) *Work events per train:* Each time a train is stopped en-route to either pickup or set-off blocks, it is called a work event. If a train performs both pickups and set-offs at an intermediate yard, it is still considered a single work event. Work events are costly in terms of carrying out the tasks of adding and removing the blocks, in terms of train delay (to the cars, locomotives, and crew that are on the train) and in terms of potential consumption of network capacity while the train is stopped. Work events as defined here are only the intermediate stops, and do not include the origination or termination events for the train.

(d) *Train length and tonnage restrictions by link:* Depending on geographical and track attributes, each section of the railroad has limitations on maximum train length and tonnage. Train tonnage refers to the weight of the train.

(e) *Number of trains passing over a link:* In order to avoid congestion on certain links of the rail network, links are constrained by the maximum number of trains that can traverse the link either by direction or for both directions on a combined basis. This can be expressed as a limit in trains/day, or on a more refined basis by shorter periods of time, possibly broken out by train type.

(f) *Crew originating and terminating yards.* In North America freight crews can only travel on predetermined crew segments and every train has to be assigned to a crew on each crew segment. As a result, all trains must originate at the start of a crew segment, and terminate at the end of a crew segment, even if this means that they have to move part of the way along a crew segment without carrying any blocks or railcars. In more complex versions of the train design problem, complex crew segments can be reflected, where the ends of each segment are made up of a cluster of relatively closely spaced locations.

Different crew segments are governed by different union agreements in the railroad industry. At times, these union agreements can get very complicated. To simplify the problem, optimization strategies typically assume fairly basic crewing rules using a version of the crew segments called single-ended territories. In a single ended territory, all crews have one end of a crew segment as their home terminal, and the other end as their away terminal. They can move a train in either direction, but must take at least 8–12 hours of rest between each move, and cannot stay at their away terminal more than a certain number of hours. See Chapter 6 crew scheduling for more information on the crew planning problem. Trains can travel across multiple crew segments. Crew imbalance on a crew segment is considered as the absolute difference between number of trains going from A to B and number of trains going from B to A. Crew imbalance results in additional expense for repositioning the crews using an over-the-road taxi service. In the simplest formulation of the train design problem, promoting train balance through the cost function is used as a proxy for ensuring minimization of overall crew requirements and minimization of crew deadhead moves.

In railroad operations, the number of locomotives required to transport a train is dependent on the power of the locomotives, weight of the freight (tonnage) on the train and the geography of the route. Locomotive requirements estimation, and the interactions between train size limits and locomotive assignments can become quite complex. As a result, trying to fully accommodate the locomotive planning problem within the train design problem may not be feasible given current solution techniques. Instead, the train design problem presented here includes objectives focused on train balance that tend to drive toward efficient use of the locomotives, but do not fully address the locomotive problem. The basic concept is to have the same number of locomotive trips terminating and originating at each location. If all trains use the same number of locomotives, then this can be represented by a cost function that promotes balance in the number of originating and terminating trains by location. In a somewhat more complex approach, the number of locomotives used by each train can be determined based on train size, locomotive attributes, and other business rules, and these numbers can be used in the locomotive balance tests. See Chapter 3 on locomotive scheduling for more details on this subject.

The objective of the Train Design Optimization problem is to minimize the sum of:

(a) Train start cost—Product of the number of trains created and the train start cost. This cost can be viewed as the cost of making up a unique train and the costs of managing the train. This cost tends to minimize the total number of unique trains, and tends to drive toward trains traveling longer distances.
(b) Train travel cost—Product of train travel distance and train travel cost per mile (this assumes all trains are largely identical in terms of speed, and thus does not consider the time-related elements of train travel cost to be a separate factor). Buried in this cost are the crew costs, the locomotive costs, fuel costs, track utilization costs, and other costs related to the operation of a train. This factor tends to minimize the total number of train-miles operated, and maximize train size.
(c) Railcar travel cost—Product of car travel distance and railcar travel cost per mile (railcars to be based on the number of railcars specified to be in each block, again ignoring any time factors).
(d) Work event cost—Pickup and set-off costs for a block varies depending on the yard/location of the activity. The sum of these individual activity costs at all the yards for all the trains is the total work event cost. This can have a number of

different approaches to how it is structured, with the costs being driven by an event cost for the overall train, and event costs by activity type for each block picked-up or set-off. How these costs are structured can be used to change the complexity of the trains, the number of en-route work events, and the desirability of block swaps. In the example problem given below this is strictly an overall cost for stopping a train at an intermediate location.

(e) Block swap cost—sum of all block swap costs across all block swap events. This is a cost per swap, not a cost per railcar, and can be used to minimize the use of block swaps. It could include a cost for the typical time that railcars dwell at a location due to a block swap operation. This is separate from the work event cost to provide an incentive to limit the use of block swaps for individual blocks.

(f) Crew imbalance cost—Product of number of imbalanced crews and crew imbalance penalty (difference in number of crews required by direction by crew segment).

(g) Train (locomotive) imbalance cost—In the simplest version of this problem formulation, this is the imbalance in the number of trains originating and terminating at each location times a cost per train for each train that is out of balance. In more complex versions, this is based on the number of locomotives used on each train and the imbalance in the number of locomotives originating and terminating at each location (if the number of locomotives is the same on all trains there is no difference between these two approaches).

(h) Missed car (block) cost—this represents the case of a block not being moved from its origin to its destination. It could be a cost or a constraint depending on the problem formulation. One could weight this cost by the number of railcars in each block, driving solutions to ensure that at least all of the largest blocks are moved.

(i) Car hire cost—this represents the time cost of the railcars being moved by the plan. If the problem is being decomposed into a phase that focuses on train routing and the BTA problem, and a separate phase to address train timing, then this cost can only be approximated in the first phase. In general this is the total transit time for the cars from shipper release to placement at the consignee multiplied by an hourly rate. In the train design problem, the variable portion of this can be approximated by applying an average velocity to each train, plus standardized time allowances for dwell times by trains at each work event location and for each block swapped block.

The train design problem is highly combinatorial in nature and a very complex optimization problem. Several attempts have been made in the past to solve special cases of the problem (Assad 1980a, b; Carpara et al. 2002; Crainic and Rousseau 1986; Dorfman and Medanic 2004; Gorman 1998a, b; Haghani 1987, 1989; Huntley et al. 1995; Jha et al. 2008; Keaton 1989, 1992; Kraft 1998, 2000; Newman and Yano Candace 2000, 2001). Recent work includes the four finalists of the train-design competition sponsored by the Railroad Applications Section (2012). These approaches vary in terms of cost and business constraints considered and the size of the underlying problem instances.

As noted earlier in this chapter, it is assumed that the traffic to be moved is fixed, and that an underlying constraint in this problem formulation is the movement of all of the available traffic. This is reflected in the missed car or block cost described above. As a consequence, the traffic volumes, and hence revenue, become effectively fixed, and the overall train design problem becomes one of cost minimization, rather than profit maximization.

1.6.1 Example Problem

To better understand the nature of the problem, it is helpful to look at the method by which a specific solution to a sample problem would be evaluated. In this example, which is adapted from the RAS problem solving competition cited above, we consider a railroad network with four nodes as depicted in Fig. 1.1. Block pickup and set-off cost information is provided for each of the nodes in Table 1.1.

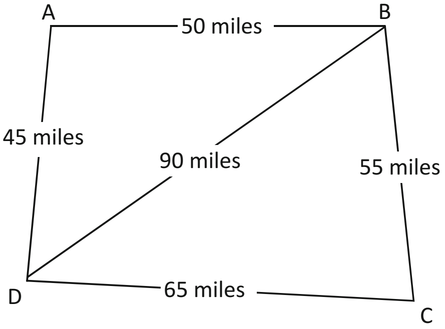


Fig. 1.1 Railroad network

Table 1.1 Pickup and set-off cost ($) at different nodes in the network

Node name	Block pickup cost	Block set-off cost	Block swap cost
A	20	10	30
B	30	20	50
C	30	20	50
D	40	30	70

Since all blocks must be picked-up at their origins and set-off at their destinations, these costs are not variable unless the number of trains that carry the block can be changed. Thus, only the block swap cost is influenced by the train design in many cases.

In this example, five blocks are made and their corresponding information is presented in Table 1.2.

Table 1.2 Block information

Block ID	Origin	Destination	# of cars	Total length (feet)	Total tonnage (tons)	Shortest distance (miles)
Block 1	A	C	50	3,000	2,500	105
Block 2	A	D	25	1,500	1,250	45
Block 3	B	D	40	2,400	2,000	90
Block 4	D	A	28	1,680	1,400	45
Block 5	D	B	16	960	800	90

Table 1.3 Network and capacity information

Origin	Destination	Distance (miles)	Max train length (feet)	Max tonnage (tons)	Max # of trains
A	B	50	8,000	10,000	3
A	D	45	5,000	11,000	4
B	C	55	9,000	9,000	5
B	D	90	8,500	10,000	4
C	D	65	9,200	11,000	4

Network and link capacity restrictions are provided in Table 1.3. All the distances are assumed to be symmetrical and all links bidirectional. For example, link B to A is 50 miles and subject to capacity constraints the same as link A to B.

Crew segment information is presented in Table 1.4. If a train's route is A → B → C, then a crew from crew segment (B–A) is assigned to the train from A → B at A and subsequently a crew from crew segment (B–C) is assigned to the train from B → C at B. When a train crosses over from one crew segment to the next, the onboard crew gets off the train and a new crew gets onboard. Further, crew segments are bidirectional. Hence, crews in crew segment A–D can take a train from either A to D or D to A. Each crew has to either travel on the shortest path between its on and off points, or at most take a route with only a limited amount of circuity compared to the shortest path for the crew segment. For our purposes we will limit the circuity to 15 %, though in reality it would be a function of the territories for which the crew is qualified and the relevant labor agreements. Also, there is a limit for the total amount of time that a crew can be on duty, which we will treat for plan design purposes as being 10 hours.

Table 1.4 Crew segments information

Node1	Node2
A	D
B	A
B	D
B	C
D	C

Other input parameters for this optimization problem are provided in Table 1.5.

Table 1.5 Other optimization parameters

Parameters	Values
Crew imbalance penalty per imbalance	$600
Train (locomotive) imbalance penalty per imbalance	$1,000
Maximum blocks per train	8
Maximum block swaps per block	3
Train travel cost per mile	$10
Car travel cost per mile	$0.75
Maximum intermediate work events per train	4
Cost per work event	$350
Cost per train start	$400
Cost per crew start	$200
Missed cost per railcar (blocks not moved penalty)	$5,000
Car hire cost per hour	$0.75
Time required for block pick-up	40 min
Time required for block set-off	20 min
Average speed of the trains (miles/h)	20

1.6.2 *Feasible Solution*

Table 1.6 presents a feasible solution in which three trains are created to transport the bloclks. Train 1 travels from yard A to yard D after picking up 75 cars at yard A. Later, Train 1 arrives at yard D, drops off 75 cars and picks up 28 cars. Subsequently, Train 1 travels from yard D to yard A with the 28 cars. Similarly, Train 2 and Train 3 travel between the rail yards to transport the cars. The total train miles in this example are 335 miles.

Table 1.6 Train routes solution. Note that times are in the form d/hh:mm, so a day 1 departure at 10:00 is 1/10:00

Train name	Seq.	Node	Scheduled arrival	Scheduled departure	Cumulative miles	Pick-up cars	Set-off cars	Out-bound cars	Crew change flag
Train 1	1	A		1/10:00	0	75	0	75	No
	2	D	1/12:15	1/13:15	45	28	75	28	No
	3	A	1/15:30		90	0	28	0	No
Train 2	1	B		1/11:00	0	40	0	40	No
	2	D	1/15:30	1/16:30	90	50	40	50	Yes
	3	C	1/20:15		165	0	50	0	No
Train 3	1	D		1/17:00	0	16	0	16	No
	2	B	1/21:30		90	0	16	0	No
Total train miles					345				

Train 1 and Train 2 stop at the common intermediate node D. At node D, both the trains either pickup and/or set-off blocks, where this activity for each train is collectively called a work event. Hence, the total number of work events done by all the trains is 2.

It is assumed that the same trains run on all days of the week. Hence, Train 1 departs yard A at 10:00 on day 1 (represented as 1/10:00) and arrives yard D at 1215 on the same day. Based on the average train speed input parameter of 20 miles/h, it takes 2 hours 15 min to travel between yards A and D. Subsequently, Train 1 has to wait for 60 min at yard D as one set-off (20 min) and one pick-up (40 min) work event happens. A train's journey can span over multiple days.

Block-To-Train Assignment information is provided in Table 1.7. For example, Block 1 travels on Train 1 from yard A to yard D. Car miles (2,250) for A to D segment for Block 1 is the product of A to D segment miles (45) and the number of cars (50) in Block 1. In other words, car miles for a block is the product of the block travel distance and the number of cars in the block. The total car miles is the sum of individual car miles for each of the blocks. In addition, this Block-To-Train Assignment solution satisfies the maximum number of block swaps constraint as presented in Table 1.5. For example, Block 1 travels on two different trains resulting in one block swap. This feasible solution also satisfies the constraint that a train can carry at most eight blocks.

Block swap costs at the intermediate nodes for a block are presented in Table 1.7. For example, Train 1 sets-off Block 1, which is subsequently picked-up by Train 2 at node D. Hence, a block swap cost at Node D is assigned to Block 1. Note that the block swap cost is not applied to the origin or destination of the block. Because each block is carried by only one train on any of its legs, we have elected to not include the block pick-up or set-off costs.

Table 1.7 Block-to-train assignment solution

Block	Seq.	Train	Start node	End node	Block swap cost	Segment miles	# of cars	Car miles
Block 1	1	Train 1	A	D	70	45	50	2,250
	2	Train 2	D	C	0	75	50	3,250
Block 2	1	Train 1	A	D	0	45	25	1,125
Block 3	1	Train 2	B	D	0	90	40	3,600
Block 4	1	Train 1	D	A	0	45	28	1,260
Block 5	1	Train 3	D	B	0	90	16	1,440
Totals					70			13,425

Table 1.8 Crew imbalance information

Crew district	Train ID	Forward	Reverse
A–D	Train 1	1	1
B–D	Train 2	1	0
B–D	Train 3	0	1
D–C	Train 2	1	0

Table 1.8 presents the crew assignment information. For Train 1, we assume that a crew is assigned from A to D, and the same crew then takes Train 1 from D back to A. This is an example of a turn-around crew that starts and ends at the same location. This is possible providing the crew stays within a single crew district and does not violate any time or distance constraints on the amount of work a single crew can do. Hence the forward and reverse direction crew balance values for Train 1 are both 1. For Train 2, one crew is assigned from B to D, and another crew is assigned from D to C, resulting in only the forward direction column being set to 1 for this train. Train 3 operates in the opposite direction on crew district B–D, so the reverse direction gets flagged for this train on this crew district. If one sums across all trains on each crew district one sees that the A–D and B–D districts are balanced, while the D–C district is not balanced.

Table 1.9 Locomotive imbalance information

Yard	Originating trains	Terminating trains	Train imbalance
A	1	1	0
B	1	1	0
C	0	1	1
D	1	0	1
Total train imbalance			2

Table 1.9 presents the train (locomotive) imbalance information, which is extracted from Table 1.6. In Table 1.6 it can be observed that one train originates at each of the yards A, B and D. The intermediate stops of the trains are not considered in this calculation. Similarly, one train terminates at each of the yards A, B and C. As one train terminates at yard C but no train originates there, yard C has a train surplus imbalance, which implies a locomotive imbalance if all trains have the same number of locomotives. Similarly, yard D has a one train deficit or imbalance.

While not shown, one could also estimate the car hours associated with each train plan. In most formulations the exact timing of trains, and the connection patters of traffic between trains is not known during the solution of the train design problem. As a result, the car hours estimation focuses primarily on the variable components associated with the average velocity of each train over the identified route, the dwell time for cars that remain on the train during work events based on standardize time allowances, and allowances for time delays for blocks being block swapped. The next section addresses the train scheduling or timing problem, and more directly examines the car hour issue.

The objective function for our example problem can be computed based on a number of different components as follows:

(a) Train start cost is 3 (number of trains) * $400 (cost per train start) = $1,200
(b) Crew start cost is 5 (number of crews used) * $200 (cost per crew) = $1,000
(c) Total train travel cost is 345 (total train miles) * $10 (cost per train mile) = $3,450
(d) Total car travel cost is 13,425 (total car miles) * $0.75 (cost per car mile) = $10,068.75
(e) Work event cost is 2 (number of train work events) * $350 (cost per work event) = $700
(f) Block swap cost is 1 (number of swapped blocks) * $70 (cost per swap) = $70
(g) Crew imbalance cost is 1 (crews out of balance) * $600 (cost per crew) = $600
(h) Train (locomotive) imbalance cost is 2 (trains out of balance) * $1,000 (cost per train) = $2,000
(i) Missed block (cars) cost is 0 (number of missed cars) * $5,000 (cost per miss) = $0

The final objective function value is $19,088.75. Obviously, this is only a small "toy" problem that has been created so the reader can follow along with the calculations. Many other solutions could be created for even this very simple problem, and thousands of solutions are possible for full scale versions of the problem.

1.7 Train Scheduling (Timing) Problem

Each train has a specific set of times associated with it. This includes the departure time from its origin point, running times between stations, intermediate dwell times, and the days of the week each train operates (frequency). Assuming a complete train plan, the train scheduling problem is focused on fixing the departure times, dwell times, frequency, and potentially the running times, with the objective of minimizing costs related to railcars, crews, and locomotives. This process must respect both line capacity and yard capacity constraints.

Most known solutions use various forms of iterative search techniques to find improvements to a train schedule. The basic idea is to adjust each train, one at a time, finding the best timing for that train, keeping all other trains fixed. This is repeated for all trains until no further improvements can be found. Some solutions do this first on the assumption that all trains operate every day of the week, and then make a second pass to adjust the frequencies. We will not be presenting specific solution techniques in any detail in this chapter, but instead will focus on defining the variables and constraints that make up the problem.

1.7.1 Key Assumptions

- *Fixed train routes*: the setting of the train times will not in any way alter the physical route taken by the trains. In general this is not an issue, with the primary exception being a case where there are alternate routes that do not impact the operational requirements of the train, but may allow line capacity to be better balanced. While one could conceive of a search algorithm that could check such alternate routes, the set-up and management of the process of identifying suitable alternate routes for trains would add significant complexity to the problem.
- *Fixed block-to-train assignments*: the block-to-train assignments are assumed to be fixed and will not be changed by the scheduling process.
- *Fixed crew change points*: in general, most scheduling algorithms assume the crew change points are fixed. In theory, changes in transit times (running times), or changes in dwell time at locations falling between crew change points, could impact how far a train could go with a single crew under the hours of service regulations. However, to simplify the problem, this factor is generally not considered in the scheduling process, and is addressed as a dependent problem that takes the train schedules as an input.
- *Fixed locomotive characteristics*: the running time of a train between locations is determined in part by the line characteristics, and in part by the train make-up including the type and number of locomotives used. Transit or running time can be changed by changing the train's locomotive characteristics. However, as a simplifying assumption, this is generally not considered a variable in the scheduling process, but instead is treated as an input.
- *Fixed weekly frequency*: trains may run daily, or less than daily. In the block-to-train assignment process, and the train route design process, the volumes expected to use each train are determined, and based on those volumes and other business requirements, the weekly train frequency is set. While the scheduling algorithm can change which days of the week a train operates, it is generally assumed that the scheduling algorithm cannot change the number of times per week each train runs.
- *Consistent operating times*: an overarching scheduling principal is that the same train will operate at the same times on each day of the week that it is run. While this is not an absolute requirement, and there can be some variations, most railroad operating plans strive to maximize the consistency of the operating times of each train by day of the week.
- *Fixed shipment release times*: the times that shipments are released by customers for movement, and the times that shipments are received at interchanges from other railroads are usually an input to the scheduling process, and are treated as fixed. This is important as these times can be leveraged by the scheduling algorithm to set the timing of at least some trains that carry a large proportion of originating shipments.
- *Fixed minimum connection times*: while the plan can call for shorter or longer connection times at yards, this is a design decision that is generally not made algorithmically. Thus, the minimum connection times are usually an input to the scheduling process, and not changed by the scheduling algorithm.

1.7.2 Scheduling Variables

- *Departure time*: this is the time the train departs its origin.
- *Transit (running) times*: these are generally treated as fixed. While there might be benefit to extending running times to improve connections, this benefit is generally achieved through adjustments to dwell times instead.
- *Dwell times*: based on the en-route work activities, there are generally minimum dwell times for specific locations. These include minimum time allowances for picking up or setting off blocks, changing crews, inspections, and fueling activities. In some cases extending these dwell times to delay the departure of the train may prove valuable if it raises the number of shipments that can connect to the train, or better balances the volumes at the yards or across the lines.
- *Frequency*: as discussed earlier, the number of times per week each train operates is typically treated as fixed, but the specific days of the week that the train operates is often a variable. For some types of trains, such as local trains, even the days operated may be fixed.

1.7.3 Scheduling Constraints

- *Line capacity*: ideally, a detailed line capacity analysis would be used to ensure the feasibility of each scheduling option. From a practical perspective, this is not possible as a scheduling algorithm examines thousands of possible scheduling options. As a consequence, most solution strategies take a higher level approach to line capacity by simply limiting the total number of trains that can traverse a specific line during a time increment (e.g., no more than X trains per hour may traverse a line in each direction).
- *Yard capacity*: from a train scheduling perspective, the primary constraint is on the number of trains per hour that a yard can receive or originate. Implicitly there is also a limit on the number of railcars that can be processed, but by limiting the number of trains, the number of railcars tends to also be limited. There could also be a limit on the number of trains that can be made up (originated) at the yard per hour.
- *Locomotive and crew availability*: in principal, the trains should be distributed over time in such a manner as to ensure that the associated crew and locomotive requirements can be met, where peaking and other timing factors can impact total locomotive and crew requirements. However, in most solution strategies this constraint is ignored, or simplified to trying to ensure a relatively even distribution of trains over time. Instead, separate sub-problems are solved to determine locomotive and crew requirements. These sub-problems may suggest further refinements to the schedules. See Chapter 6 on crew planning and Chapter 2 on locomotive planning.
- *Minimum/maximum frequency*: as discussed earlier in this section, the frequency of each train is generally treated as fixed. An alternative approach treats the fixed

frequencies as a minimum required frequency, and allows the scheduling algorithm to consider higher frequencies. In most cases the maximum frequency of a train is set at a daily frequency. If additional frequencies are required, then a separate train should be created to support the additional train runs.

- *Shipment service commitments*: in some cases specific shipments must be delivered within specific overall transit times, or by specific arrival times. When such constraints exist, the scheduling algorithm must attempt to satisfy these service commitments. See Chapter 4 on car scheduling for a discussion of how end-to-end transit times are computed.

1.7.4 Cost Parameters

Overall, most of the costs that apply to the general train routing design and block-to-train assignment problem apply to the train scheduling problem. However, if we assume that the train routes and frequencies are fixed, then the costs addressed in the route design and block-to-train assignment problem are no longer variable in the scheduling problem, and do not need to be factored into the solution (instead these scheduling requirements are treated as constraints). The two exceptions are (a) if the train frequency is allowed to vary, the costs of running additional trains must be accounted for, and (b) if route variations are allowed, the relative costs of the different routes must be taken into account.

Assuming fixed train frequency and routes, time-based costs tend to be the primary drivers of the train scheduling process

- *Railcars*: while adjustments to the train schedules, particularly en-route dwell time, can impact the transit time of the railcars, the largest impact on railcars is the dwell time cars spent in yards waiting for trains to depart. Thus, a dock-to-dock view of overall transit times for railcars should be considered in the cost function, tying the train schedules to the time cost of the railcars. The unit cost for the railcars is typically either a representative per diem or car hire rate, or in some cases it is the car hire rate plus an allowance for the carrying cost of the goods within the railcars. Since railroads tend to focus more on their direct costs, the carrying costs are generally not included in the calculation. See Subsection 1.7.5 for a discussion on how the railcar time factors are calculated.
- *Train hours*: this is a time-based cost for the train from the time it is made-up to the time it terminates. It often includes cost components for the crews and locomotives associated with the train, as well as the costs for the time railcars spend in the train (with the dwell time for the railcars being calculated separately). If the running time between stations is treated as fixed, then the primary variable in this cost is en-route dwell time.

Train scheduling can also impact the efficiency with which locomotives and crews can be used. For locomotives, this impact is primarily on the idle time (dwell time) for locomotives waiting on train departures. For crews, it is primarily on the extent to which crews must be deadheaded to their home terminal. An example of a locomotive impact would be a case of a location with one terminating and one originating train. Depending on the timing of these trains, and the time it takes to service the locomotives and put them on the originating train, the exact timing of the terminating and originating trains will directly impact the dwell time for the locomotives at this location. For crews, there are a number of rules related to minimum rest times between assignments, and how long they can spend away from home. Depending on the timing of the trains, some crews may need to be taxied (deadheaded) to their home terminals if the away from home time limits are exceeded. Adjustments to the train schedules have the potential to reduce the amount of deadheading required.

Due to the complexities of the crew and locomotive scheduling problems, they are generally not addressed in any detail in the train scheduling process, but instead are treated as a separate sub-problem that can provide suggestions for schedule adjustments. See Chapter 2 on locomotives and Chapter 6 on crews for a more detailed discussion of this topic.

1.7.5 Observations on Solution Strategies

Solution strategies for the train scheduling problem of which the authors are aware generally examine three primary variables for each train: the origin departure time, the length of intermediate dwell times, and the days of the week that the train should operate. This process can be decomposed into two or three phases, with the first phase focusing on the best time to originate trains, the second on dwell time adjustments, and the third on the days operated. Various forms of heuristic search strategies are typically used, adjusting one train at a time while keeping all others fixed.

The primary drivers of these adjustments are the dwell times experienced by railcars connecting to the train, and the total time that equipment and crews spend in the train. Given the assumption that the number of railcars in each train will not change with changes in the train timing, and that the number of locomotives is fixed, the in-train time for equipment and crews is a straight forward calculation (this assumption may not be correct in the case of a block moving on more than one train, but is still used to simplify the scheduling algorithm). Thus, the efficient computing of the dwell times for the connecting railcars becomes one of the key focuses of any scheduling algorithm.

The principles of car scheduling or trip planning are used to compute the dwell times. See Chapter 4 on car scheduling for an extensive discussion of this process, as well as the examples provided below. Shipments connecting to a train at a specific location come from one of two sources: local originations at the location (including railcars received through interchange with other railroads), or arrivals on in-bound trains at the location. Under a strategy that adjusts one train at a time, all of the origination and arrival times of the connecting railcars are known. Thus, one can apply the car scheduling logic, including the minimum processing times for each railcar at a location, to compute the dwell times for each connecting railcar given a specific departure time. Using this approach, each departing train can be tested for a variety of possible departure times to find the time that will produce the lowest amount of total car dwell for all cars connecting to the train at all locations in the train route.

Using the above framework, the algorithm needs to employ a strategy that determines in what order the trains should be tested, and the extent to which dwell time adjustments are tested in addition to adjusting the overall train forwards or backwards in time. Any dwell time adjustments must include the costs of the railcars, crews and locomotives already on the train at the connecting location, in addition to the railcars connecting to the train at that location.

A primary factor to consider in this process is that initially railcar arrival times at a yard are known for some shipments, and not for others. In particular, cars that originate at a location have fixed times, while railcars that arrive on trains at a location could experience changes in their arrival times as the schedules are adjusted. Furthermore, if the times for a train have not yet been set, then the arrival times for cars traveling on that train are effectively unknown. As a consequence, there is a benefit to adjusting the schedules of trains with a high proportion of traffic that has known arrival times first, and trains where the arrival times of some cars are not known later in the process. The scheduling algorithm will likely not consider the traffic with unknown arrival times carried by a particular train when it sets that train's timing. As train schedules are fixed, the proportion of traffic with known arrival times will steadily increase. Overall it is likely that any such algorithm will take an iterative approach, and some trains will be adjusted multiple times as greater proportions of their traffic have known arrival times.

Testing changes to the days operated for a train uses the same approach as the train timing adjustments, examining the overall dwell time for the railcars connecting to the train to determine the best days for the train to run.

1.7.6 Special Cases

There are many potential complexities and special cases that must be considered in any scheduling process. These include the handling of unit trains, intermodal traffic, addressing customer commitments, handling of "anchor blocks," local train scheduling, line capacity modeling, crew and locomotive requirements analysis, and handling of special operations such as the gathering of grain traffic to make up solid trains. A few of these special cases are addressed below:

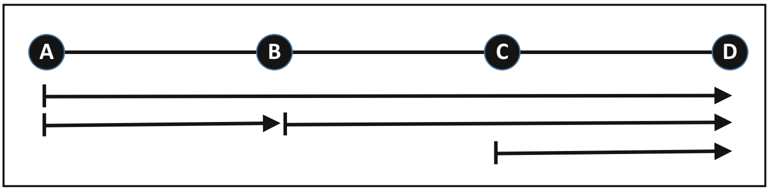
- *Unit train scheduling*: unit trains come in many flavors, but in most scenarios there is an assumption that each unit train consists of a fixed is composed of railcars that cycles through sequential loaded and empty movements. Under such a scenario, the scheduling of unit trains becomes very dynamic, and is not so much focused on meeting specific timing goals as ensuring the efficient assignment of the train consist to a series of loads, ensuring sufficient time is allowed for the necessary empty repositioning movements. As a consequence, the scheduling strategies discussed in this section generally would not apply to most unit train operations.

- *Grain scheduling*: grain operations have evolved in North America to comprise two core types of operations: shuttle trains and gathering networks. Shuttle trains are unit trains that are dynamically scheduled to move a series of loads from grain elevators to ports or other points of consumption. As such, the unit train scheduling principles apply. Smaller lot grain is typically handled through a gathering process, where groups of railcars are loaded at grain elevators and then processed for movement to destination. These groups of railcars are brought to gathering points, and depending on the available volume they are then either forwarded through the regular manifest network, or made up into unit or solid trains for movement to destination. Again, this becomes a dynamic scheduling process, and would not typically be addressed by a fixed scheduling process such as that discussed in this section.
- *Customer commitments*: there are many flavors of commitments. Some promise that shipments will be delivered within a maximum amount of time from when the shipments are released at origin. Others specify that if shipments are released by a specific time, they will be delivered by a specific time at destination. The overall process of minimizing total railcar hours in the scheduling process described above may or may not satisfy a specific customer commitment. As a consequence, the scheduling algorithm may need to be modified if specific customer commitments are to be met. There are a number of strategies that can be employed. The simplest is to minimize dwell times for commitment traffic, typically by placing a higher cost per hour on the railcars with commitments. Back testing at the end of the process can determine if any shipments are out of compliance with the final solution, possibly causing further adjustments in the schedules. More complex solutions will attempt to fix the timing of some trains based on the commitment requirements. This is particularly true of intermodal, where there can be very tight time windows for the departure and arrival of trains.
- *Anchor blocks*: some railroads have the concept of anchor blocks, where an anchor block is the most important block or group of railcars on the train. These anchor blocks typically represent the primary commercial reason for the train's existence, and may have specific scheduling requirements that must be treated as taking precedence over the needs of any other traffic on the train. In effect, only the traffic on the anchor blocks will be considered when setting the timing of the trains carrying the anchor blocks.
- *Intermodal*: the service requirements for intermodal can be very specific. A typical requirement might be something like stating that shipments will depart no earlier than 10 pm from a loading ramp, and must be available at destination no later than 8 am, 2 days later. If only one train is used to move these shipments, then the ability to adjust the train's schedule is determined by the amount of slack that exists between the overall running time for the train and the amount of time in the service commitment. Further, there may be a bias in how the trains are

scheduled to provide further protection against service failures (e.g., try to have the train arrive as early as possible to have some allowance for unplanned delays). If the shipments must connect between trains, then the scheduling parameters become more complex as each train in the shipment routing must take the overall service commitment into account.

1.7.7 Problem Examples

To understand the train scheduling process, we need to understand the scheduling of an individual train. For this purpose, we will use a train that has four route locations, and carries four blocks as follows:



As depicted above, this train progresses from location A to location D, via locations B and C. It carries the following blocks: A to D, A to B, B to D, and C to D.

As discussed earlier, each block has a set of shipments that connect to it, and these shipments have specific arrival times at the location where the connection is being made. For example, we might have the following arrival pattern at location A for the block A to D:

Arrival group	Arrival time	Number of railcars
1	02:00	4
2	04:00	8
3	08:00	12
4	13:00	4
5	15:00	12
6	23:00	8

As discussed in the Chapter 4 on car scheduling, the dwell time for a railcar at a yard is a combination of the minimum processing time for the railcar to be switched and placed in the outbound train, and the waiting time between the end of processing and the departure of the train. For example, consider the case where the minimum processing time allowance at a yard is 8 hours, and a railcar arrives at the yard at 0200. This would mean that the railcar could depart the yard at any time from 1000 onwards. If the train the car is assigned to does not depart until 1600, then the total dwell time will be 14 hours (8 hours to process, and 6 hours of waiting time).

While the processing time for each railcar could differ based on its priority and other factors, for simplicity we will assume that the processing time for all railcars is always 8 hours. This tells us that the optimal departure time for arrival group 1 in the above table would be 1000, and for group 2 it would be 1200, etc. In this simple example, this gives us six possible departure times to test for this particular block. The results of such testing would be as follows:

			Dwell time in hours by train departure time for block A to D					
Arrival group	Arrival time	Number of railcars	10:00 departure	12:00 departure	16:00 departure	21:00 departure	23:00 departure	07:00 departure
1	02:00	4	8	10	14	19	21	29
2	04:00	8	30	8	12	17	19	27
3	08:00	12	26	28	8	13	15	23
4	13:00	4	21	23	27	8	10	18
5	15:00	12	19	21	25	30	8	16
6	23:00	8	11	13	17	22	24	8

The dwell time for the 10:00 departure and the 02:00 arrival time is 8 hours because the train departs at exactly the point when the processing is complete. The railcars for the 04:00 arrival time will not be ready to depart until 12:00, which is 2 hours after the 10:00 departure time, so these cars would need to wait 22 hours once processing time is complete to depart, resulting in a 30 hours dwell time. Using this approach, each of the dwell times can be computed.

If we multiply the dwell times by the number of cars, we can compute the total car dwell associated with each departure time:

			Total car hours by train departure time for block A to D					
Arrival group	Arrival time	Number of railcars	10:00 departure	12:00 departure	16:00 departure	21:00 departure	23:00 departure	07:00 departure
1	02:00	4	32	40	56	76	84	116
2	04:00	8	240	64	96	136	152	216
3	08:00	12	312	336	96	156	180	276
4	13:00	4	84	92	108	32	40	72
5	15:00	12	228	252	300	360	96	192
6	23:00	8	88	104	136	176	192	64
Total car hours			984	888	792	936	744	936

What this shows is that having the train depart at 23:00 would minimize the total car hours for the A to D block. However, there are three other blocks being assigned to this train, so this testing process needs to be expanded to include the impact on total car hours for all blocks carried by the train.

To understand this, let us add consideration of the shipments that join the train at location B on the B to D block. To keep things relatively simple, our example has this traffic arriving in only four groups at location B as follows, and that the same 8 hours of processing time applies:

Arrival group	Arrival time	Number of railcars
7	04:00	6
8	09:00	8
9	11:00	12
10	19:00	4

Based on a formula, we would determine the elapsed time from when the train leaves A to the time when the train leaves B. This would typically be the running time from A to B, plus the dwell time at B. The dwell time at B would be a minimum time based on the activities that take place at B (crew changes, inspections, locomotive changes, setting off of cars, picking up of cars). The running time would be based on the expected speed of the train over each route segment, which might vary by train type. For our example, we will assume a 3 hours running time, plus a 1 hour dwell time, so that the departure time from B will be 4 hours after the train departs from A.

The “ideal” departure times for B would be 12:00, 17:00, 19:00, and 03:00 based on the 8 hours processing time allowance, which would imply departure times from A of 08:00, 13:00, 15:00, and 23:00 (4 hours earlier). The 23:00 departure time matches one already tested for A. Based on the other tested departure times for A, this would yield the following additional times from B: 14:00, 16:00, 20:00, 01:00, and 11:00.

We can view this as introducing three more times at A, plus giving us nine times to test at B. The additional times at A yield the following:

			Total car hours by train departure time for block A to D		
Arrival group	Arrival time	Number of railcars	08:00 departure	13:00 departure	15:00 departure
1	02:00	4	120	44	52
2	04:00	8	224	72	88
3	08:00	12	288	348	372
4	13:00	4	76	96	104
5	15:00	12	204	264	288
6	23:00	8	72	112	128
Total car hours			984	936	1032

The times at B yield the following results in terms of dwell hours at B:

			Dwell time in hours by train departure time for block B to D								
Arrival group	Arrival time	# of railcars	01:00 dept.	03:00 dept.	11:00 dept.	12:00 dept.	14:00 dept.	16:00 dept.	17:00 dept.	19:00 dept.	20:00 dept.
7	04:00	6	21	23	31	8	10	12	13	15	16
8	09:00	8	16	18	26	27	29	31	8	10	11
9	11:00	12	14	16	24	25	27	29	30	8	9
10	19:00	4	30	8	16	17	19	21	22	24	25

This translates to the total car hours shown below for B. Also shown are the corresponding car hours at A, and the total car hours for both locations:

			Total car hours by train departure time for block B to D								
Arrival group	Arrival time	# of railcars	01:00 dept.	03:00 dept.	11:00 dept.	12:00 dept.	14:00 dept.	16:00 dept.	17:00 dept.	19:00 dept.	20:00 dept.
7	04:00	6	126	138	186	48	60	72	78	90	96
8	09:00	8	128	144	208	216	232	248	64	80	88
9	11:00	12	168	192	288	300	324	348	360	96	108
10	19:00	4	120	32	64	68	76	84	88	96	100
Total car hours (B to D)			542	506	746	632	692	752	590	362	392
A to D departure times			21:00	23:00	07:00	08:00	10:00	12:00	13:00	15:00	16:00
Total car hours (A to D)			936	744	936	984	984	888	936	1,032	792
Total car hours for both blocks (A to D and B to D)			1,478	1,250	1,682	1,616	1,676	1,640	1,526	1,394	1,184

As can be seen from the above, adding in consideration of the block from B to D changes the best departure time from A to be 16:00, instead of the time of 23:00 when the A to D block was only considered. One would expect further changes as the other blocks carried by the train are considered.

A more advanced strategy would also consider adding extra time to selected dwell times to see if that would improve overall dwell times. As an example, consider adding 2 hours to the dwell time at B for the 13:00 departure time from A. To do this, we must take into account the car hours associated with the A to D block that must wait the additional time at B:

Departure time at A	13:00	Car hours at A	936
Departure time at B	19:00	Car hours at B	362
Added dwell time at B	2 h for 48 cars	Extra car hours at B	96
Total car hours	For revised schedule		1394

While this example yields no benefit to the overall timing of the train relative to a 15:00 departure time from A and a 1 hours dwell time, it does show the type of analysis that can be used to explore such alternatives. It also shows the ability to produce alternative schedules that can be equally optimal, which may prove valuable in balancing the departure times of the trains against the capacities of the yards and lines.

There are several important considerations in this process:

Outbound train perspective: the process typically only considers the impact on dwell time of the cars connecting to the train being evaluated. Changing this train's times might increase the dwell for cars connecting to other downstream trains. While these downstream impacts are typically not considered during the processing of the current train, they will likely be captured in later iterations of the process when these other trains have their schedules adjusted.

Prioritization of blocks: a number of approaches can be taken to make sure that commercially important traffic on a train is treated preferentially in the scheduling process. The two most common strategies are to weight the car hours differently based on the priority of the traffic, or to only consider selected traffic during the schedule setting process.

Balancing yard workloads: as trains are scheduled, limits may need to be observed on the number of trains originating or terminating at a yard during a particular time of day. Such limits could be treated as either hard constraints (not allowing train departures during those times) or soft constraints (by penalizing car hours for trains departing during congested periods).

Line capacities: as with the yards, as trains are scheduled, limits may need to be placed on the number of trains traversing a line during certain times of the day, where such limits could be imposed through hard or soft constraints.

1.8 Specifying Unit Trains

Most railways specify unit trains similarly to specifying road trains and represent them using the same data elements: effective/expiration dates, day-of-week frequency, route information with arrival and departure times, and usually a single train-block. Usually they are marked "as-required" which indicates that the train will run only when operations specifically designates it to run.

However, there are several operational and data specification attributes for unit trains that should be noted. Typically the day-of-week frequency is all 7 days of the week, even if the train runs only once a week or once a month. The timing information is considered to have the correct run times (times between stations), but origin start time is considered to be fictitious in the system and set to a specific value for a specific train instance when it is manually specified to actually run by a railroad's operations group. There may be several versions of the same train, but with different routes.

As part of day-to-day management of the railroad, the train master will select a unit train, and specify it to operate on a specific day with a specific start time.

The overall concept is that it takes specialized talent to correctly enter a train schedule into the system: the route, the crew change locations, the desired locomotive power, and the run times require careful analysis and management approval to be adopted for actual use. By having the unit trains in the system, even marked “as-required,” all that information is already there and approved. Operations Department only needs to allow that train to be run, and give it a designated start time. So the representation of the unit trains in the planning system is essentially a template that has been preapproved.

In some cases, notably mine-to-port operations, many trains per day are created. Often the planning system has 24 or 48 of these trains represented for each hour or each half-hour of possible train departure times. These trains each have a different train symbol. Operations Department only needs to instantiate the subset of trains that will run each day. The importance is that most systems do not allow two trains with the same train symbol to originate on the same day, and by having enough predefined trains it allows operations to choose the train with the best fit to reality, especially in regard to train origin departure time.

Grain trains are typically the hardest to implement: Even though they are unit trains, there are many combinations of silos and destinations for them.

From an analysis view, the unit train specification makes it very difficult to estimate train sizes, locomotive power needed and so on since the trains as represented in the system have a much different frequency of operation when compared to real life. There are two paths to dealing with this. To get average train sizes, often only a “runs per week” value is needed. So if a unit train that is depicted as running all 7 days of the week has a “runs per week” of 0.5, the planning system will mathematically account for it as if it runs once every two weeks.

However, trip plans and detailed locomotive and crew models need a more specific train schedule. This is done by having the unit trains be modeled as a typical week in history. This means that some unit trains that run less than once per week will be modeled, and some will not. While not perfect, this approach does ensure that a typical week of schedules is part of the planning analysis.

1.9 Local Service Specification Strategies

The movement of railcars to/from industry often poses special challenges that require an alternate set of specifications for trains and blocks. This is caused by several factors, including the nature of how local switching services are provided, the ”addresses” for customers, and the large number of unique customers that must be served.

One factor to consider is the number of customers that need to be served. The car scheduling process must have a means of generating a solution to every station and customer that might generate a railcar movement, not just the ones that consistently generate such movements. A local train that serves all of the customers along a line might have 50, 100, or even more potential customers within its service area. If we were to generate a block to and from each and every one of these customers, this could result in a set of local trains that had dozens, or even hundreds of blocks on them. To avoid this, many railroads have systems that allow local trains to be defined as serving a range of stations or customers, without specifying specific blocks. While effective from a data management perspective, this results in the need for special logic in the blocking system and trip planning system to handle these alternate train definitions and "implicit" blocks, as well as alternate ways of specifying the trains.

A second factor is that some local trains do not leave the area covered by a single station. For example, customers and interchanges that are located within a yard area are all specified with a single station number. Most train scheduling systems require that the train goes to more than one station. To specify yard switching operations that serve customers and interchanges at the yard, special trains must be designated that do not match the pattern of other trains and require special logic for blocking and trip planning purposes. Furthermore, these single station operations create the need to assign a yard-block to the car movement at the destination of the movement. Normally, once one has reached the destination for a trip, there is no further action required. However, in the case of customers or interchanges located at the yard, it is likely the yard will need to switch these cars into specific blocks for delivery to these customers even though the destination has been reached. The result is the need to support the designation of a final yard-block for each shipment, and special logic to determine when these yard-blocks are required.

A final factor to consider is that a single station may contain multiple customers. Most of the train schedule, block, and trip planning processes are built around the concept of the station. However, when providing local services one must operate at the level of the specific customer, and in some cases for large customers a specific siding at the customer's site. This results in a second addressing system below the level of the station. Often called the zone-track-spot (ZTS) system, it goes by many names across the industry. Most blocking systems need to have overrides of some form to assign block codes by customer and/or ZTS type information, and train services must have ways of specifying the timing of services to be provided at the ZTS level. Generally this is handled within the process of generating final yard-blocks, and the specification of local train services. Other complications may arise when road trains provide local switching services en-route, raising the need to specify the specific customers to be served by these trains.

The result of the above is that there does not exist within the industry a consistent manner for specifying local services. Furthermore, many railroads use different methods for operations within a single station and for trains that move between stations. Common methodologies for local service specification include:

- *Local trains with explicit blocks*: Under this scheme, local blocks are treated like any other block and are assigned to trains as conventional train-blocks. All railroads to varying extent have some trains that carry local blocks and are specified in this manner. As noted earlier, one big issue with this approach is that some trains could end up with dozens of separate blocks on them, making them very complicated and making the train specification difficult to maintain. On a given day, a local train with 15 blocks and 30 unique route points might only have traffic to use 3 blocks and 5 route points. Hence, the timing and the precise route of the local train is not known, and would certainly be different than a specification that has all 15 blocks and 30 route points. Finally, there is the need in some cases to have single station trains, which may result in requiring special specifications for yard switcher type operations.

- *Local trains with partially explicit blocks*: One partial solution to the above that has been adopted by at least one Class I railroad is to only put a few representative blocks on each train. A station range is then associated with each block. While the block goes to only one place, the station range implies that the train could set-off the block at any of the locations in the station range. For example, a train might carry a block from A to F, with a station range of D to H. This means that the train is also implicitly serving stations D, E, G, and H with that block, in addition to F. The train may only have timing information for location F. As a result, logic has been developed to choose a time for the other stations in the range, typically using a "best guess" time from among the times appearing in the train route. This approach greatly simplifies the train definition, and is fairly easy to maintain and understand. However, it also complicates the trip planning and block generation logic because both must accommodate the station range concept and the implicit creation of the other blocks.
- *Local blocks on road trains*: This is the case of a road train serving selected local customers en-route. It is a fairly common scenario, and occurs on essentially all railroads. In general, this situation is handled by allowing trains to carry a combination of regular blocks and local blocks, using one of the two strategies outlined above.
- *Local trains with implicit blocks*: In some cases, local trains are specified without the existence of any local blocks. In this scenario the train is given a route, but no blocks are designated. Instead the train contains specifications of the customers that may be served at each location in the train route, timing information related to how that service will be supplied, and information potentially down to the zone-track-spot level on exactly what traffic may be handled at each location and whether the train can deliver cars, pick-up cars, or do both. In general these types of schedules assume that all cars being delivered by the train are put on the train at the train's origin, and all cars being picked-up by the train are moved to the train's destination. The train may also have a set of yard-block codes associated with it, with no specific pick-up or set-off locations specified. This type of train in effect has implicit blocks from the train origin to each station/customer it serves, and implicit blocks from each station/customer it serves to the train destination. This type of train is straight forward to specify, likely adequate for specifying most purely local trains, and as a result greatly simplifies the train definition and maintenance process. However, it does complicate the blocking system and trip planning system logic, and this must be carefully addressed.
- *Service area local specifications without routes*: Most railroads have some form of single point or terminal area service specification. These are typically trains that never leave the area covered by a single station designation, and provide switching to customers and interchanges located within that station. Typically these "schedules" contain a set of timing parameters, in most cases a set of yard-block codes they will handle, and some additional rules related to their purpose and how to carry

out that purpose. In terms of timing data, this might include a cut-off time by which terminating cars must be available to be handled by the switch job, an on-duty time, a delivery time by which cars should be placed by the switch job, a cut-off time by which originating cars must be released to be handled by the switch job, a return time by which cars should be available for classification at the on-duty yard, and an off-duty time. Sometimes other timing parameters are used, but the above captures the essence of the time factors. The rules related to the switch job generally specify if the train can deliver cars, pick-up cars, or both, and if the train is serving an interchange or customers. In the case of an interchange, the connecting railroad will be specified, and in the case of customers, the customer codes or zone-track-spot information will be specified. These types of specifications do not set a specific work order and can cover a wide variety of customers. Essentially, they define an open-ended duty assignment where the exact duties and timing can vary from day to day. They are straight forward to specify and maintain, but again must be specially handled by the trip planning and blocking systems since they have both implicit blocking and do not contain conventional train routes. Most of the Class I railroads have some variation of this train type. There can be variations of this type of specification that also include multiple stations and thus serve a larger area.

- *Non-train-based specifications*: In some cases the exact nature of the local services are not known, or are not maintained centrally. As a result, some trip planning systems provide timing parameters for the pick-up and delivery of cars to customers without reference to specific trains or switch jobs. These typically look like some variation of the service area specification approach or the local trains with implicit blocks approach, though there may be separate entries for each customer or local station as there is no need to bundle the specifications into jobs. While this approach can be adequate for generating trip plans, it provides little insight into the nature of how local services are provided, no support for analyzing local workloads and costs, and at best may only provide an approximation of the timing factors for providing local services. This type of data is often difficult to maintain because it does not relate directly to how the services are delivered, and as a result tends to be ignored and poorly maintained.

As noted, each of the above approaches has its strengths and weaknesses, which vary depending on the circumstances and business practices of each individual railroad.

1.10 Train Plan Design Versus Real-Time Operations

The train plan design represents a catalog of trains that the railroad may operate. Some of these trains will operate all of the time, and others will only be operated when appropriate traffic exists to justify their operation. We can thus view the base train plan as a set of template trains, which are then converted to date-specific instances of these trains during actual operation of the railroad.

The result is that the planning process and the real-time operations have somewhat different focuses. In the planning process, the focus is on designing a set of regularly operated trains that meet the expected traffic volumes the railroad will experience, and ensuring that templates exist for use in the real-time operations to meet the needs of any as-required trains (unit trains being the most common example of such as required trains). In the real-time environment, the focus is on managing date-specific schedules, and generating these trains from the templates found in the base train plan.

From an OR perspective, the consequence is that during the design process, the planner needs to understand the expected volumes on each train to ensure the plan is appropriately sized, and whether the plan as designed is complete in terms of being able to move all available traffic. The estimation process for train volumes is discussed in detail in Chapter 4 on simulation. Completeness is generally tested by checking that all of the blocks in the blocking plan have a way to be moved from their origins to their destinations. This is generally done by testing each block against the train plan, where the testing process must take block swaps into account (separate tests are performed to ensure that the blocking plan can move all of the expected shipments).

Even if only one train in the base train plan can pick-up a specific block, in the real-time environment, different, date-specific versions of that train will operate on each day of the week as one looks ahead. These date-specific trains are viewed as independent from each other for trip planning and train schedule management purposes, and each is considered a separate, eligible train to carry the block, and thus the various railcars.

It is important to note that in the real-time environment the near term trains are likely known with greater precision than the trains to be operated further in the future. Thus, most car scheduling systems use “dated” or actual trains in the near term, and planned or “template” trains further out in time. As train schedules change, trains are added, annulled, etc., the near term dated train schedules are updated so that these changes are reflected in the trip plans. Thus, in the real-time environment the train schedules can still be somewhat dynamically created, as long as an up-to-date, complete, forward view of the plan is maintained in the computer system with a 7- to 14-day planning horizon.

There are a number of different types of trains in the real-time environment from a data management perspective:

Type of train	Template train schedules	Dated train schedules
Auto-add road trains	These are the base schedules from planning for the regularly scheduled trains that are expected to always operate	These are date-specific versions of the regularly scheduled trains and include any changes that occur during operation of the trains
Manual add trains	These are templates for trains that may be called at the discretion of operations, including unit trains	These are date-specific versions of trains called at the discretion of operations, and include the actual times of operation and traffic to be carried
Local trains	These are templates for the planned local services, and are generally regularly scheduled, but in some cases may be operated only at the discretion of operations	These are the date-specific versions of the local trains, both those that are regularly scheduled and those called at the discretion of operations, and include any changes that occur once called

To expand on the above, let us more carefully define a template and dated train schedule:

- *Template train schedule*: A template train schedule is simply a prototype train that may or may not actually be operated. These are the train schedules typically designed and maintained by planning. Some trains are regularly scheduled, and might specify that they run on specific days of the week. Others are "as-required" trains that will only be operated if needed during actual operations. This second group of templates exists to make it easier for operations to add trains to the dated train schedule if and when they want to operate an additional train.
- *Dated train schedule*: Railroads typically maintain a database containing the actual trains currently being operated, and the trains the railroad anticipates it will operate in the next several days. Where the template database might have a single train schedule that says it operates Monday through Friday, the dated train schedule database will have a separate entry for each day that the train actually operates. Most databases require that there only be one instance of a particular train symbol originating on a specific date. If today was the 23rd day of the month, and train 409 operated every day of the week, the database might contain the trains 409-21, 409-22, 409-23, 409-24, and 409-25. These would represent copies of the same train both in the recent past, and in the near future. By having separate copies of the trains, we can record the actual operating times for each train, make adjustments to the train plan, and uniquely associate traffic movements with each train.

Using the approach described above car schedules can be generated for any time period in the future. Many of the Class I railways maintain template databases with multiple copies of the same train, where the expiration dates on the trains cover various time periods to reflect special situations anticipated to happen in the future, such as maintenance of way (MOW) activities. In general, the template database returns to the "standard" version of each train once one goes out past the period of time for which MOW changes are reflected (typically 1–3 months).

Template trains are divided into two groups, "auto-add" (also called regularly scheduled trains) and "manual adds." In general, the auto-add trains are put into the dated train database on an automatic basis. Typically, a process runs once or twice per day that inserts either 12 or 24 hours worth of auto-add trains based on the template database. Once added, if some of these trains will not be operated, they must be manually annulled by operations within the dated train schedule database. All other trains must be manually added by operations as the need arises.

The process for manually adding trains to the dated train schedule database typically starts by a user selecting a specific train schedule in the template database to use as a model for the train to be added. This template could be a regularly scheduled train, or a manual add train. The user then makes a few potential adjustments to the train. Common adjustments are to change the train symbol, possibly truncate part of the train route, add or drop a train-block, adjust a connection standard, or mark a train-block as primary or fill. The user then provides the specific date the train will originate on, and the specific time it will depart its origin. The train is then added to the dated train database using the template, plus the changes provided by the user. Typically, the times at the route locations are set as offsets from the origin departure time supplied by the user based on the times found in the template schedule.

Once in the dated schedule database, significant amounts of data may be associated with each train, such as the actual cars that will be carried by the train, locomotive data, crew data, etc. The schedule is then further updated to reflect changes on when the train will actually operate, changes to the work the train will perform, etc., as such changes become known. In the best versions of these databases, if the train deviates from the planned route, these deviations are captured, and as actual train timings are received, the remainder of the schedule is updated to reflect the expected downstream effects of the train being ahead or behind schedule.

1.11 Opportunities

Opportunities abound in the area of train scheduling. Unlike the blocking problem, which has been extensively studied and well established procedures for optimization are in active use, the train schedule design problem is much less advanced. Optimization tools have been designed for train schedule design, and applied in a number of areas, and have been effective in identifying incremental improvements to train plans. Through direct experience, the authors are aware of tools developed for use at three different North American railroads, and at least one European railroad. While these tools produced useful results, they also need to be further refined to be truly effective. Further, circumstances differ enough from one railway to the next that there may not be a “one size fits all” type solution. Instead, different tools may need to be created for each unique business environment and type of operation. For example, the European environment that uses fairly short duration/distance trains with a limited number of blocks, and drivers that can operate more than one train in a day will need a different approach than that required in North America with its multi-day runs and multiple crews per train run.

Specific opportunities include:

- Road train optimization tools for the carload business that reflect the local business/operating environment of each railroad. Assuming a fairly static train plan design, this is likely a tool that operates at the level of a monthly planning cycle, with some support for shorter timeframes.
- Intermodal planning also needs to be addressed. Due to its similarities to the carload problem, a carload solution might also serve the needs of intermodal planning, or there might need to be separate tools to tackle intermodal train design.

- Unit train planning tools at the long-term, short-term, and day-of-operation levels. As discussed earlier, a focus must be on equipment cycling and matching train sets to demand/specific orders. This likely is primarily a tool that can be used with a short planning horizon of less than 2 weeks.
- Grain train planning and management tools that can both manage the matching of supply to demand, and the decision process of when to run grain in dedicated trains, and when to move it in the carload network. As with the unit train problem, this probably represents a tool operating at a planning horizon of 1–14 days.
- Tactical evaluation and repair tools for evaluating the impact of short-term plan changes, and determining the best actions to return to plan.

While there are likely many other opportunities, such as local service planning, the authors believe that solutions to the above list would be a great place to start and provide a significant advance for the industry.

References

Assad AA (1980a) Modelling of rail networks: toward a routing/makeup model. Transport Res Part B 14(1–2):101–114

Assad AA (1980b) Models for rail transportation. Transport Res 14A:205–220

Carpara A, Fischetti M, Toth P (2002) Modeling and solving the train timetabling problem. Oper Res 50:851–861

Crainic TG, Rousseau JM (1986) Multicommodity, multimode freight transportation: a general modeling and algorithmic framework for the service network design problem. Transport Res 208:225–242

Dorfman M, Medanic J (2004) Scheduling trains on a railway network using a discrete event model of railway traffic. Transport Res B 38:81–98

Gorman MF (1998a) The freight railroad operating plan problem. Ann Oper Res 78:51–69

Gorman MF (1998b) An operating plan model improves service design at santa fe railway. Interfaces 28(4):1–12

Haghani AE (1987) Rail freight transportation: a review of recent optimization models for train routing and empty car distribution. J Adv Transport 21:147–172

Haghani AE (1989) Formulation and solution of a combined train routing and makeup, and empty car distribution model. Transport Res 23B:433–452

Huntley CL, Brown DE, Sappington DE, Markowicz BP (1995) Freight routing and scheduling at CSX transportation. Interfaces 25(3):58–71

Ireland P, Case R, Fallis J, Van Dyke C, Kuehn J, Meketon M (2004) The Canadian pacific railway transforms operations by using models to develop its operating plans. Interfaces 34(1):5–14

Jha KC, Ahuja RK, Sahin G (2008) New approaches for solving the block-to-train assignment problem. Networks 51:48–62

Keaton MH (1989) Designing optimal railroad operating plans: Lagrangian relaxation and heuristic approaches. Transport Res 23B:415–431

Keaton MH (1992) Designing optimal railroad operating plans: a dual adjustment method for implementing Lagrangian relaxation. Transport Sci 26:262–279

Kraft ER (2000) Implementation strategies for railroad dynamic freight car scheduling. J Transport Res Forum 39(3):119–137, jointly with Transportation Quarterly

Kraft ER (1998) A reservations-based railway network operations management system, Ph. D. Dissertation, Department of Systems Engineering, University of Pennsylvania, Philadelphia, PA, UMI Order # 9829930

Newman AM, Yano Candace A (2000) Direct and indirect trains and containers in an intermodal setting. Transport Sci 34:256–270

Newman AM, Yano Candice A (2001) Scheduling trains and containers with due dates and dynamic arrivals. Transport Sci 35:181–191

Railroad Applications Section (2012) 2011 RAS Problem Solving Competition, Train Design Optimization, INFORMS. https://www.informs.org/Community/RAS/Problem-Repository

Chapter 2
Locomotive Scheduling Problem

Balachandran Vaidyanathan and Ravindra K. Ahuja

2.1 Introduction

Every day, railroad managers need to assign hundreds of locomotives to hundreds of different trains. Locomotive scheduling involves optimally assigning a set of locomotives to each train so that the assignment satisfies a variety of business constraints and minimizes the total scheduling cost. Major US railroad companies have several billions of dollars of capital investment in locomotives. Thus, solving this problem effectively is of critical importance for railroads.

Locomotive Scheduling Problems (LSPs) are difficult problems to solve because of several operational complexities. One such complexity is called *consist-busting*. The set of locomotives assigned to a train is called a *consist*. When a train arrives at the destination, its consist is either assigned to an outbound train in its entirety, or the consist goes to a pool of locomotives where it is broken down and new consists are formed. The former case is referred to as a *train-to-train connection* between the inbound and outbound trains, and the latter case is referred to as a *consist-busting*. The cost of assembling and disassembling consists must be controlled by developing plans that minimize *consist-busting*. Locomotives which provide power to the train are referred to as *active* locomotives. Due to difference in power demand at different stations, locomotives are also *repositioned* from one station to another. Locomotives can be repositioned by simply *deadheading* on an existing train, or by traveling

B. Vaidyanathan (✉)
FedEx Corporation, 1000 Ridgeway Loop Road, Suite 500,
Memphis, TN 38120-4045, USA
e-mail: bala.vaidyanathan@gmail.com

R.K. Ahuja
Optym, 2153 SE Hawthorne Boulevard, Gainesville, FL 32641, USA
e-mail: ravindra.ahuja@optym.com

© Springer Science+Business Media New York 2015

B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3_2

independently from one station to another, also referred to as *light travel.* These kind of operational considerations make the problem hard to solve, and generating implementable locomotive schedules is therefore a challenge.

The paper Cordeau et al. (1998b) presents an excellent survey of existing locomotive scheduling models and algorithms for this problem. The models described in existing literature can broadly be classified in two categories: *single locomotive type* and *multiple locomotive type models. Single locomotive scheduling models* assume that there is only one type of locomotive available for assignment. These models can be viewed as minimum cost flow problems with side constraints; some papers on single locomotive scheduling models are due to Wright (1989), Forbes et al. (1991), Booler (1980), Booler (1995), and Fischetti and Toth (1997). Single locomotive assignment models are appropriate for some European railroad companies but are not suited for US railroad companies since most trains are assigned multiple types of locomotives. *Multiple locomotive assignment models* have been studied by Florian et al. (1976), Ramani (1981), Smith and Sheffi (1988), Chih et al. (1990), Nou et al. (1997), Cordeau et al. (1998a), Ziarati et al. (1997), and Ziarati et al. (1999). The most recent and comprehensive multiple locomotive assignment models are due to Ahuja et al. (2005) which has been further refined by Vaidyanathan et al. (2008). While Ahuja et al. (2005) develop the initial framework to solve this problem, Vaidyanathan et al. (2008) improve the initial effort on several dimensions leading to the development of a practical locomotive scheduling approach.

2.2 Background on Locomotive Scheduling

This section gives an overview of the LSP and defines the terminology needed to understand and define the problem.

Locomotive information: A railroad has different types of locomotives with different pulling and cost characteristics and number of axles (often ranging from four to nine). We denote the set of all the locomotive types by K and the index k represents a particular locomotive type. The following attributes are associated with each locomotive type k: (1) h^k: the horsepower provided by a locomotive of type k; (2) λ^k: the number of axles in a locomotive of type k; (3) G^k: the weekly ownership cost for a locomotive of type k; and (4) B^k: fleet size of locomotives of type k, that is, the number of locomotives available for assignment.

Train information: Locomotives pull trains from their origins to their destinations. We denote the set of all trains by L. Trains have different weekly frequencies; some trains run every day, while others run less frequently. If the same train runs on different days, we consider it as different train entities; that is, if a train runs 5 days a week, it is considered as five different trains. The index l is used to denote a specific train. Each train has the following associated information: (1) *dep-time*(l): the departure time for the train l; (2) *arr-time*(l): the arrival time for train l; (3) *dep-station*(l): the departure station for train l; (4) *arr-station*(l): the arrival station for train l; (5) T_l: tonnage requirement of train l; (6) β_l: horsepower per tonnage needed for train l; (7) H_l: horsepower requirement of train l, which is defined as $H_l = \beta_l T_l$; and (8) E_l: the penalty for using a single locomotive consist for train l.

Locomotive–train combinations: For each locomotive type assigned to a train, we consider the following attributes: (1) c_l^k: the cost incurred in assigning an active locomotive of type k to train l; (2) d_l^k: the cost incurred in assigning a deadheaded locomotive of type k to train l; and (3) t_l^k: the tonnage pulling capability provided by an active locomotive of type k to train l.

2.2.1 Hard Constraints

Hard constraints are mandatory constraints which have to be satisfied for a locomotive schedule to be feasible.

Power requirement for trains: Each train must be assigned locomotives with sufficient tonnage and horsepower to pull it from origin to destination.

Locomotive class to train type: Each train type (e.g., auto train, or merchandise train, or intermodal train) can only be pulled by certain locomotive types.

Geographic: Each geographic region permits the use of only specific locomotive types. For example, it may be specified that the Atlanta area can only use: CW40, AC44, and AC60 locomotives.

Locomotive balance constraints: The number of incoming locomotives of each type into a station must be equal to the number of outgoing locomotives of that type at that station.

Active axle constraints: Each train must be assigned locomotives with at most 24 active axles because exceeding 24 powered axles may overstress the couplers and cause a train separation.

Consist size constraints: Each train can be assigned at most 12 locomotives including both the active and deadheading locomotives. This policy reduces risk exposure if the train were to suffer a derailment.

Fleet size constraints: The number of utilized locomotives of each type is at most the number of available locomotives of that type.

Repeatability of the schedule: The routing of locomotives should be such that the number of locomotives of each type at each station at the end of the week should be equal to the number of locomotives of each type at each station at the beginning of the next week (so that the locomotive plan is repeatable every week).

2.2.2 Soft Constraints

Soft constraints define characteristics of a solution which are preferred but not mandatory.

Consistency in locomotive assignment: A train should be assigned the same consist each day that it runs. Railroads believe that crews will perform more efficiently and more safely if they operate the same equipment on a particular route and train.

Consistency in train connections: When locomotives assigned to a train connect from one train to another, then they should preferably make the same connection on each day that both the trains operate.

Avoid consist busting: Consist busting involves the use of additional resources to break consists and put together new consists. It is therefore preferable to avoid consist busting.

Avoid single locomotive consists: If a single locomotive is assigned to a train and this locomotive breaks down, then the train will get stranded.

2.2.3 Objective Function

The objective function for locomotive scheduling is to minimize the sum of: (1) cost of ownership, maintenance, and fueling of locomotives; (2) cost of active and deadheading locomotives; (3) cost of light traveling locomotives; (4) penalty for consist-busting; (5) penalty for inconsistency in locomotive assignment; and (6) penalty for using single locomotive consists.

2.3 Mathematical Models for Locomotive Scheduling

The LSP can be formulated as a multi-commodity flow problem with side constraints on a network called the *weekly space–time network*. Each locomotive type defines a commodity in the network. In this section, we describe the network and the formulation.

2.3.1 Space–Time Network Construction

Each node in the network is associated with two attributes: time and place. The network contains a *train arc* (i_l, j_l) for each train l. The tail node i_l of the arc corresponds to the departure of train l at *dep-station*(l) and is called a *departure node*. The head node j_l corresponds to the arrival of train l at *arr-station*(l) and is called an *arrival node*. For each arrival, an *arrival-ground node* is created, and for each departure, a *departure-ground* node is created. *Time*(*i*) denotes the time attribute of node *i* in the weekly space–time network. The sets of departure, arrival, and ground nodes are denoted by the sets *DepNodes*, *ArrNodes*, and *GrNodes*, respectively, and the set $AllNodes = DepNodes \cup ArrNodes \cup GrNodes$.

The network contains four types of arcs. The first is the set of train arcs, which is denoted by the set *TrArcs* and is described above. Each train arrival node is connected to the associated arrival-ground node by a directed arc called the *arrival-ground connection arc*. Each departure-ground node is connected to the associated train departure node through a directed arc called the *ground-departure connection arc*. All the ground nodes at each station are sorted in the chronological order of their time attributes and each ground node is connected to the next ground node through directed arcs called *ground arcs*. The ground arcs allow inbound locomotives to stay in an inventory pool as they wait to be connected to the outbound trains. The last ground node in the week at a station is connected to the first ground node of the week at that station through a ground arc; this ground arc ensures that the ending inventory of locomotives for a week becomes the starting inventory for the following week, which ensures the repeatability of the schedule. The possibility of an inbound train sending its entire consist to an outbound train is modeled by creating *train–train connection arcs* from train arrival nodes to train departure nodes whenever such a connection can be feasibly made. Thus, the four kinds of arcs are: train arcs (*TrArcs*), connection arcs (*CoArcs*), and ground arcs (*GrArcs*). Let al$lArcs = Tr Arcs \cup CoArcs \cup GrArcs$. Figure 2.1 displays a part of the weekly space–time network at one location.

The LSP can be formulated as a flow of different types of locomotives in the weekly space–time network. Locomotives flowing on train arcs are either active or deadheading; and those flowing on connection and ground arcs are idling (that is, waiting between two consecutive assignments). The following additional notation is used in the mathematical formulation: (1) $I[i]$: the set of incoming arcs into node $i \in AllNodes$; (2) $O[i]$: the set of outgoing arcs from node $i \in AllNodes$; (3) d_l^k: defined for every arc $l \in AllArcs$ (for a train arc *l*, d_l^k denotes the cost of deadheading of locomotive type *k* on train arc *l*, and for every other arc it denotes the cost of traveling for a non-active locomotive of locomotive type *k* on arc *l*); (4) *CB*: the set of all connection arcs from arrival nodes to ground nodes; alternatively, $CB = \{(i, j) \in AllArcs: i \in ArrNodes \text{ and } j \in GrNodes\}$; (5) *CheckTime*: a time instant of the week used to count the number of locomotives used; and (6) *S*: the set of arcs that cross the *CheckTime* [that is, $S = \{(i, j) \in AllArcs: time(i) < CheckTime < time(j)\}$].

2.3.2 Problem Size and Stage-Wise Solution Approach

The mathematical formulation of the LSP contains around 200,000 variables and 100,000 constraints and cannot be solved to optimality or near optimality using commercial state-of-the-art software. Additionally, the formulation does not capture

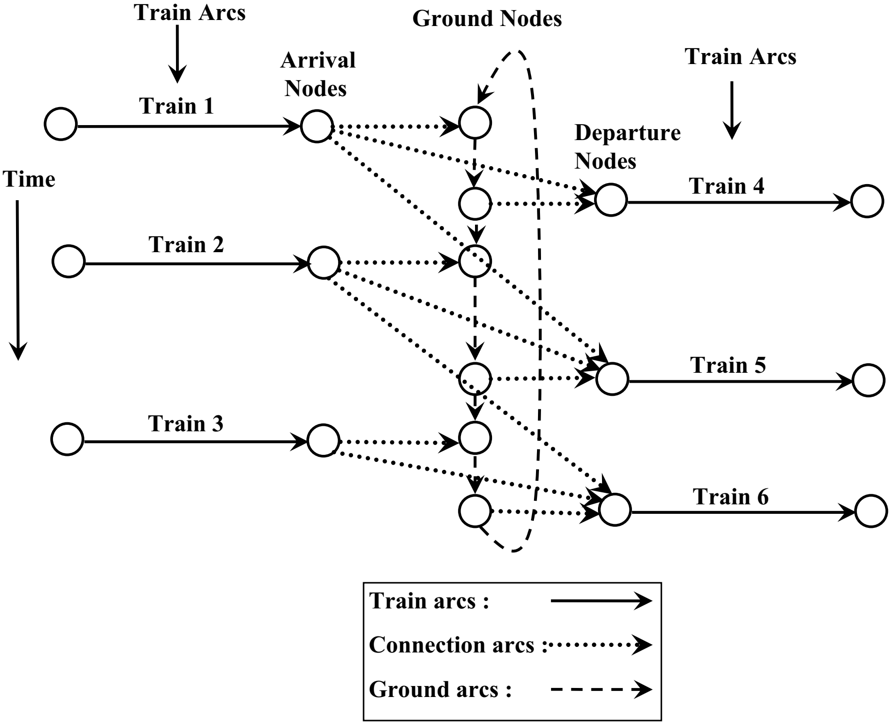


Fig. 2.1 A part of the weekly space–time network

the consistency constraints effectively. The main contribution made by Ahuja et al. (2005) was to develop a two-stage solution approach that captures the consistency constraints. In the first stage, the daily locomotive scheduling problem which is a simplified problem is solved, and in the second stage the daily locomotive schedule is modified to obtain the feasible weekly locomotive schedule.

This two-stage approach is motivated by the observation that in a typical problem more than 90 % of the train arcs in the space–time network correspond to the trains that run 5, 6, or 7 days. Based on this observation, the daily locomotive scheduling problem is created in the following manner: (1) all trains that run p days or more per week run *every* day of the week; and (2) all trains that run fewer than p days do not run at all.

To transform the solution of the daily locomotive scheduling solution into a feasible solution to the weekly scheduling problem, locomotives are taken from the trains that exist in the daily problem but do not exist in the weekly problem (Type 1 trains) and assigned to the trains that do not exist in the daily problem but exist in the weekly problem (Type 2 trains). This may lead to the model using additional locomotives to meet the constraints. The solution of the daily problem can be translated into the solution of the weekly problem more effectively if the number of Type 1 trains is less than the number of Type 2 trains but still as close as possible.

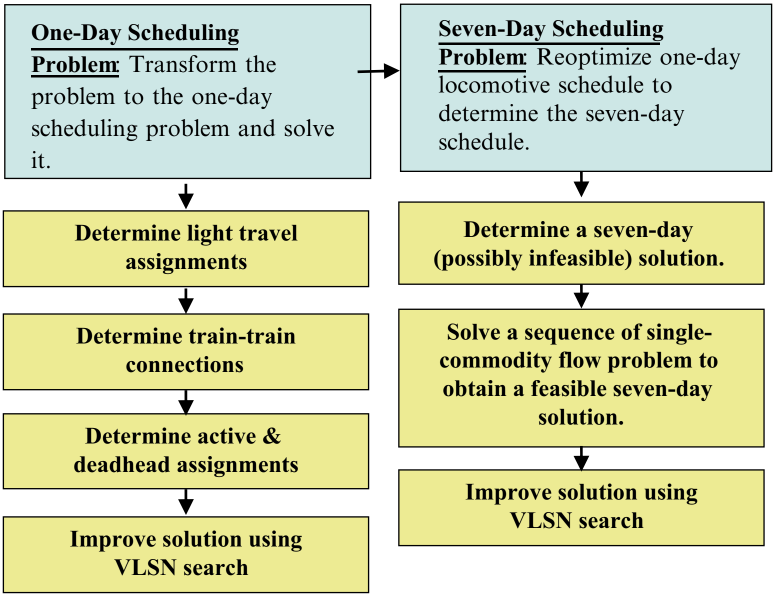

One-Day Scheduling Problem: Transform the problem to the one-day scheduling problem and solve it.
Determine light travel assignments
Determine train-train connections
Determine active & deadhead assignments
Improve solution using VLSN search
Seven-Day Scheduling Problem: Reoptimize one-day locomotive schedule to determine the seven-day schedule.
Determine a seven-day (possibly infeasible) solution.
Solve a sequence of single-commodity flow problem to obtain a feasible seven-day solution.
Improve solution using VLSN search

Fig. 2.2 Overview of the multistage locomotive scheduling algorithm

Another contribution made in Ahuja et al. (2005) involves determination of good train–train connections and good light arcs. Railroads often specify some candidate train–train connections and candidate light arcs out of which a certain number are fixed in the final solution. The candidate train connection or light arc has a fixed charge variable associated with it and these fixed charge variables make the mathematical formulation very hard to solve. Ahuja et al. (2005) describe a heuristic that can be used to determine a good set of train–train connection and light arcs. Figure 2.2 gives the various steps in their stage-wise approach.

2.3.3 Consist Flow Formulation for the LPP

Consist busting affects crew requirements, station fluidity, locomotive productivity, and mechanical maintenance processes. It consumes between 2 and 6 additional hours per locomotive within the station, asset time that could be productively used to pull trains on the mainline. Upon reassembly, each consist must undergo extensive operational testing as well. In addition, consist busting often results in outbound trains getting their locomotives from several inbound trains. If any of these inbound trains is delayed, the outbound train is also delayed, which potentially propagates to further delays down the line. Consequently, railroad managers seek to streamline and simplify processes in order to eliminate fragility in the operating plan. In reality, consists will be tactically busted as part of real-time operations to compensate for unplanned events.

In order to minimize consist busting, Vaidyanathan et al. (2008) extended the locomotive flow formulation described in Ahuja et al. (2005). While the solution approach was still a stage-wise one, consists are routed over the network instead of individual locomotives. In this formulation, referred to as *consist flow formulation*, each consist type (that is, a group of locomotives) is defined to be a commodity that flows on the network. Thus, the consist flow formulation differs from the locomotive flow formulation in the sense that locomotive types are replaced by the consist types. Every feasible solution of the consist flow formulation has a corresponding feasible solution to the locomotive flow formulation with the same cost, but the converse is not true. Thus, the optimal solution cost of the consist flow formulation cannot be better than that of the locomotive flow formulation. However, computational results revealed that if the number and types of consists are judiciously chosen, then both formulations produce solutions with comparable quality. This is indeed extremely beneficial since the consist formulation significantly reduces consist busting (whose cost is not reflected in the model).

The consist flow formulation is a multi-commodity integer programming problem formulation on the space–time network.

2.3.3.1 Notation

C: Denotes the set of consist types available for assignment and $c \in C$ represent a particular consist.

c_l^c: Cost of assigning an active consist of type $c \in C$ to train l.

d_l^c: Defined for every arc $l \in AllArcs$. For a train arc $l \in TrArcs$, d_l^c captures the cost of deadheading a consist of type $c \in C$ on arc l. For an arc $l \in CoArcs \bigcup GrArcs$, d_l^c captures the cost of idling for a consist type $c \in C$ on arc l.

α^{ck}: Number of locomotives of type $k \in K$ in consist type $c \in C$.

$I[i]$: Set of arcs entering node i.

$O[i]$: Set of arcs leaving node i.

S: Set of overnight arcs or arcs that cross the Sunday midnight timeline. (This time is chosen as the time for counting the number of locomotives used in the solution.)

2.3.3.2 Decision Variables

x_l^c: Binary variable representing the number of active consists of type $c \in C$ on arc $l \in TrArcs$.

y_l^c: Integer variable representing the number of non-active consists (deadheading, light-traveling or idling) of type $c \in C$ on arc $l \in AllArcs$.

zl: Binary variable which takes value 1 if at least one consists flows on arc $l \in LiArcs$ and 0 otherwise.

sk: Integer variable indicating the number of unused locomotives of type $k \in K$.

2.3.3.3 Objective Function

$$\min z = \sum_{l \in TrArcs} \sum_{c \in C} c_l^c x_l^c + \sum_{l \in AllArcs} \sum_{c \in C} d_l^c y_l^c + \sum_{l \in LiArcs} F_l z_l - \sum_{k \in K} G^k s^k \tag{2.1}$$

2.3.3.4 Constraints

$$\sum_{c \in C} x_l^c = 1, \quad \text{for all } l \in TrArcs, \tag{2.2}$$

$$\sum_{c \in C} \sum_{k \in K} \alpha^{ck} \left(x_l^c + y_l^c \right) \le 12, \quad \text{for all } l \in TrArcs, \tag{2.3}$$

$$\sum_{l \in I[i]} \left(x_l^c + y_l^c \right) = \sum_{l \in O[i]} \left(x_l^c + y_l^c \right), \quad \text{for all } i \in AllNodes, c \in C, \tag{2.4}$$

$$\sum_{k \in K} \sum_{c \in C} \alpha_{ck} y_l^c \le 12 z_l, \quad \text{for all } l \in LiArcs, \tag{2.5}$$

$$\sum_{l \in S} \sum_{c \in C} \alpha^{ck} \left(x_l^c + y_l^c \right) day(l) + s^k = B^k, \quad \text{for all } k \in K, \tag{2.6}$$

$$x_l^c \in \{0,1\}, \quad \text{for all } l \in TrArcs, c \in C, \tag{2.7}$$

$$y_l^c \ge 0 \text{ and integer}, \quad \text{for all } l \in AllArcs, \ c \in C, \tag{2.8}$$

$$z_l \in \{0,1\}, \quad \text{for all } \ l \in LiArcs. \tag{2.9}$$

$$s^k \ge 0, \quad \text{for all } \ k \in K \tag{2.10}$$

Constraint (Eq. 2.2) ensures that every train *l* is assigned exactly one active consist. Constraint (Eq. 2.3) ensures that the locomotive flow upper-bound on each train arc is satisfied. Constraint (Eq. 2.4) ensures that flow is balanced at every node for every consist type. Constraint (Eq. 2.5) ensures that the locomotive flow upper-bound on each light arc is satisfied. Constraint (Eq. 2.6) ensures that the number of locomotives used for each fleet type is no more than the fleet size.

Note that in this formulation, it is not required to explicitly specify the constraints that each train gets the required tonnage, horsepower and does not exceed the 24-active axle requirement. These constraints are implicitly handled in the formulation. The active axle constraints are handled by not creating consists which have more than 24 active axles. The tonnage and horsepower constraints are handled implicitly in the following way; if assigning consist $c \in C$ as an active consist to train $l \in TrArcs$ violates the tonnage or horse power constraints, then the corresponding variable is set to zero ($x_l^c = 0$); thus disallowing the assignment of consist *c* to train *l*. The consist flow formulation has significantly less side constraints compared to the locomotive flow formulation, resulting in faster solution time. Another speed-up in the consist flow formulation comes from the fact that each active consist assignment variable, x_l^c, is a binary variable, whereas in the locomotive flow formulation, it is a general integer variable; this makes the consist formulation a lot easier to solve. There are instances where the locomotive flow formulation could not give a feasible integral solution in 10 hours, but the consist flow formulation gave an optimal solution within a few minutes of computational time.

Railroads often impose complex rules on what locomotive types may be combined into ideal consists. Some locomotives do not work well together. Some railroads segregate AC powered locomotives and DC powered locomotives. These requirements are often very hard or impossible to honor in the locomotive flow formulation but are rather trivial in the consist flow formulation. Further, in the locomotive flow formulation, an outbound train often obtains its planned consist from locomotives coming from multiple trains and if any of these inbound trains is delayed, the outbound train is delayed as well. But in the consist flow formulation, all outbound trains derive their active consist only from one inbound train (but may derive their deadhead consists from other trains) thus reducing the impact of train delays. In summary, the benefits of using the consist formulation are (1) solution speed and robustness greatly improved, (2) consist busting is reduced to zero, and (3) constraints are more easily incorporated, resulting in more practical solutions.

Computational tests have shown that the consist flow formulation may have its optimal objective function value as much as 5 % higher than that of the locomotive flow formulation. However, the solution is far superior in terms of consistency, simplicity, and robustness. Thus, it may be easier to comply with and may need overall fewer locomotives in practice (considering train delays, for example).

2.4 Incorporating Practical Requirements

2.4.1 Cab-Signal Requirements

Each locomotive in the fleet is equipped with specialized equipment that may enable it to operate in certain restricted territories while at the same time disqualify it from operating in other territories. Some of this equipment is required to enable the locomotive to be the first in a consist or the *lead* locomotive. Other equipment may be required by union rules or regulatory rules unique to certain geography. Some of these constraints are handled during tactical (real-time) assignment of locomotives to trains. However, one particular constraint, which we call the *cab-signal constraint*, must be a part of the locomotive plan.

Railroad corridors are equipped with signaling systems to control the movement of trains. Most corridors are outfitted with wayside signals only; the crew in the locomotive observes the signal and slows, stops, or proceeds depending on it. To increase safety, some corridors do not have wayside signals but instead are equipped with cab-signal systems, where the signal is displayed inside the locomotive. Not only do cab-signals aid the crew in foggy weather or in cases where the wayside signal may be obscured but the cab-signal systems also interface with the locomotive throttle and brake and in cases where a crew does not honor the signal, the system will automatically slow down and then stop the train to avert a possible collision. Federal law requires that all consists operating in cab-signaled territory must have a lead locomotive that is outfitted to interface with the cab-signal system. For planning purposes, railroads like to have all locomotives in a consist equipped with cab-signals so that if the lead locomotive breaks down, the units can be swapped on the line of road and the train will continue; or, at the end of the line, the consist can reverse direction without being turned. Outfitting a locomotive with cab-signals costs in excess of $100,000 and there are increased maintenance and inspection requirements as well. Consequently, railroads do not equip every locomotive with cab-signals since that is too costly.

To incorporate the cab-signal logic, we partition each consist type into two consist types, one with the cab-signal capability and the other without the cab-signal capability. For example, consist type 2-[SD40] can be decomposed into the categories: 2-[SD40-Normal] and 2-[SD40-Cab]. Similarly, we decompose the fleet-size requirements for SD-40 locomotives into two parts: SD40-Normal and SD40-Cab (this decomposition is an input provided by the railroad). For each train l, which operates in a cab-signal corridor, we ensure that the all the consist flow variables on it, that is, x_l^c, y_l^c are zero when c is not a cab-signal compatible consist. This change guarantees that all the trains in cab-signal corridors are assigned cab-signal consists, but those in normal corridors can be assigned either kind of consist.

2.4.2 Foreign Power Requirements

Railroads often cooperate to run trains directly from the origin station on one railroad to the destination station on another railroad. These trains are called *run-through* trains. By allowing locomotives from the originating railroad to stay with the train through to destination, the cooperating companies eliminate queuing of trains and locomotives at the interchange stations that would otherwise occur as one railroad would have to move extra locomotives to the interchange point in anticipation of the train arrival, or the train would have to sit at the interchange waiting for locomotives. Over the week, several inbound run-through trains bring in foreign power into the network and several outbound run-through trains return foreign power to other railroad networks. However, the inbound and outbound run-through train schedules are not perfectly balanced and the foreign power often flows back to other interchange locations or on direct or indirect trains. But, at the end of the week, each railroad is obligated to return as many locomotives that it receives from each connecting carrier in order to maintain the overall balance of power across North America.

The foreign power requirement is incorporated by making appropriate changes to the space–time network. The network is augmented in such a way that solving the problem would automatically guarantee a plan which accommodates foreign power. We create the *augmented space–time network* in the following manner. We first create the space–time network as described in Sect. 2.3.1. Then, we create a pseudo *super station* for each of the foreign railroads. All the trains which bring in foreign power into the system originate at their respective super stations and terminate at their respective destinations. All the trains which send foreign power out of the system originate at their respective origins and terminate at their respective super stations. Due to the flow balance constraints, the number of locomotives (or consists) of a particular type entering a super node will be equal to the number of locomotives (or consists) of the same type leaving the same super node. However, the model may use the super station as a shortcut between two geographic stations. For example, a consist that needs to travel between Chicago and Memphis may be routed from Chicago to a super station and then from the super station to Memphis because it is cheaper to do so. To prevent this kind of shortcutting, we can set the costs on ground arcs at the super node to a large value. This ensures that the solution does not misuse the super node as a shortcut between two geographic stations and that only essential foreign power movement takes place between railroads.

2.5 Applications of the Model

The locomotive scheduling model has various applications, and we describe a few applications in this section.

2.5.1 Quantifying the Impact of Varying Minimum Connection Time

Freight trains do not run on time and often arrive later than their planned arrival time, which makes it difficult for locomotive dispatchers to adhere to the locomotive plan. One method commonly recommended to improve plan compliance is to increase the train–train minimum connection times. Although increasing the minimum connection time may improve plan adherence, it also increases locomotive costs as more locomotives will be held in inventory at terminals. The model could be used to quantify the impact of increasing the minimum connection time.

Figure 2.3 gives an example of one such study. Depending upon the lateness of trains and the willingness of railroad planners to improve locomotive plan compliance, appropriate connection times can be used.

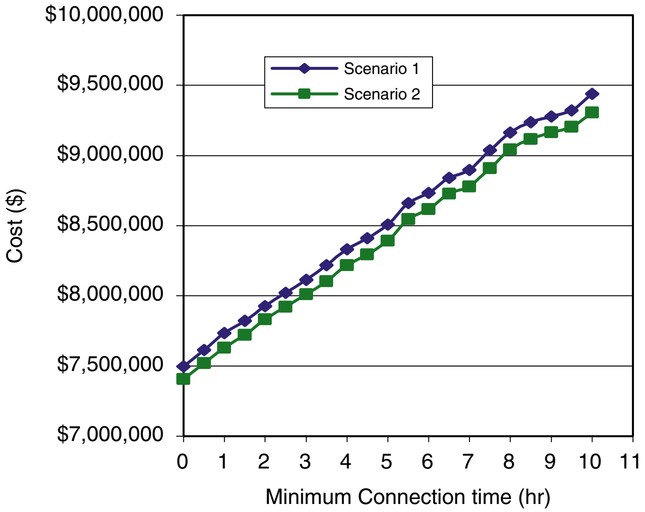


Fig. 2.3 Solution cost versus minimum connection time

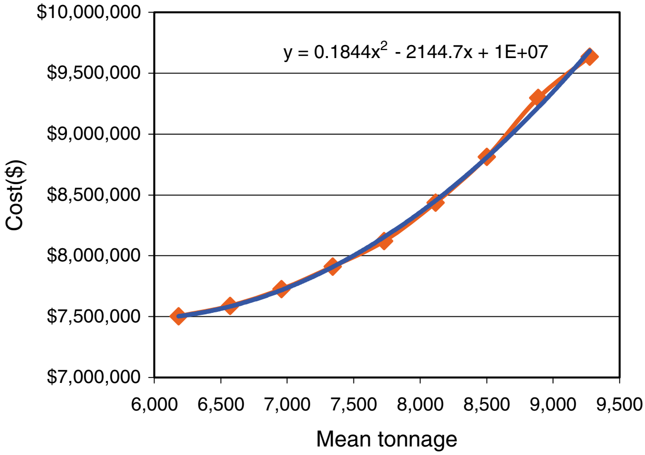


Fig. 2.4 Impact of transport volumes on solution cost

2.5.2 Quantifying the Effect of Changing Transport Volume on Key Performance Characteristics

The model can be used to measure the impact of varying transport volumes (or train tonnages) on the key transport characteristics such as number of locomotives used, solution cost, mean pulling power of a consist, and mean miles traveled per consist is measured. An example of one such analysis is shown in Fig. 2.4.

Thus, railroads can use the model to determine the relationship between rail freight transport volume and the optimal number of locomotives needed or the transportation cost.

References

Ahuja RK, Liu J, Orlin JB, Sharma D, Shughart LA (2005) Solving real-life locomotive scheduling problems. Transp Sci 39:503–517

Booler JMP (1980) The solution of a railway locomotive scheduling problem. J Oper Res Soc 31:943–948

Booler JMP (1995) A note on the use of Lagrangean relaxation in railway scheduling. J Oper Res Soc 46:123–127

Chih KC, Hornung MA, Rothenberg MS, Kornhauser AL (1990) Implementation of a real time locomotive distribution system. In: Murthy TKS, Rivier RE, List GF, Mikolaj J (eds) Computer applications in railway planning and management. Computational Mechanics Publications, Southampton, UK, pp 39–49

Cordeau JF, Soumis F, Desrosiers J (1998a) A Benders decomposition approach for the locomotive and car assignment problem. Technical report G-98-35, GERAD. Ecole des Hautes Etudes Commerciales de Montreal, Canada

Cordeau JF, Toth P, Vigo D (1998b) A survey of optimization models for train routing and scheduling. Transp Sci 32:380–404

Fischetti M, Toth P (1997) A package for locomotive scheduling. Technical Report DEIS-OR-97-16. University of Bologna, Italy

Florian M, Bushell G, Ferland J, Guerin G, Nastansky L (1976) The engine scheduling problem in a railway network. INFOR 14:121–138

Forbes MA, Holt JN, Watts AM (1991) Exact solution of locomotive scheduling problems. J Oper Res Soc 42:825–831

Nou A, Desrosiers J, Soumis F (1997) Weekly locomotive scheduling at Swedish State Railways. Technical Report TRITA/MAT-97-OS12. Royal Institute of Technology, Stockholm, Sweden

Ramani KV (1981) An information system for allocating coach stock on Indian Railways. Interfaces 11:44–51

Smith S, Sheffi Y (1988) Locomotive scheduling under uncertain demand. Transp Res Rec 1251:45–53

Vaidyanathan B, Ahuja RK, Liu J, Shughart LA (2008) Real-life locomotive planning: new formulations and computational results. Transp Res B 42:147–168

Wright MB (1989) Applying stochastic algorithms to a locomotive scheduling problem. J Oper Res Soc 40:187–192

Ziarati K, Soumis F, Desrosiers J, Gelinas S, Saintonge A (1997) Locomotive assignment with heterogeneous consists at CN North America. Eur J Oper Res 97:281–292

Ziarati K, Soumis F, Desrosiers J, Solomon MM (1999) A branch-first, cut-second approach for locomotive assignment. Manag Sci 45:1156–1168

Chapter 3
Simulation of Line of Road Operations

Roger W. Baugher

3.1 Introduction

In its simplest form, the operations of a railroad can be split into two disciplines: line of road operations and terminal operations. This chapter focuses on the management of line of road operations. Railroad management devotes the largest amount of analytic effort to this discipline, and tools exist for its analysis. Most railroads and many consulting firms have staffs dedicated to using tools to analyze line of road operation and justify capital improvements.

A critical element of line of road operation is the meet–pass planning process, which can be defined as the science of determining where a set of trains, either following or opposing one another (a pass and a meet, respectively), will be routed to resolve their conflict in a network of more than one track. While conceptually simple, the problem proves to be difficult to solve. Like many operations research problems, it involves a search for a feasible and optimal solution inside a large solution space. Any problem of practical size has many possible solutions—dispatching ten pairs of opposing trains which can meet at any of five sidings generates nearly ten million solutions. In general, the number of possible solutions (*N*) equals the number of meet locations (*S*) to the power of the number of trains (*T*) or $N = S^T$. In practice, many of the solutions can be rejected as illogical—it makes little sense to hold all westbound trains at their origin until all eastbound trains have arrived there. Additional factors a dispatcher must consider—many of which are described below—will further limit the number of feasible solutions that should be considered. However, finding a good plan—even one that is not optimal—remains a daunting challenge.

R.W. Baugher (✉)
Atlanta, GA, USA
e-mail: rwbaugher@aol.com

© Springer Science+Business Media New York 2015

B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3_3

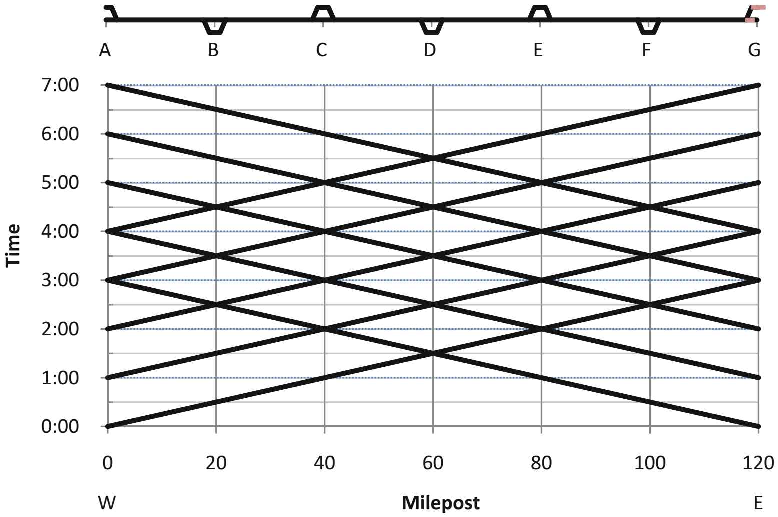


Fig. 3.1 Exhibit I

A convenient aid to visualizing the meet–pass problem is the use of what is known as a time–distance or stringline diagram. As shown in (Fig. 3.1), axes are drawn to depict time (shown here increasing upward along the *y*-axis) and distance (shown here increasing along the *x*-axis). Sloping lines in the diagram represent trains. Lines sloping up to the right represent eastbound trains moving toward Milepost 120 (time increases and milepost increases toward Milepost 120), while lines sloping up to the left represent westward trains moving toward Milepost 0 (time increases and milepost decreases toward Milepost 0). At the top is a schematic diagram of the track network—a single track line with sidings. The plot captures the meet–pass problem in its simplest form: five sets of opposing but identical trains arriving at even time intervals and running at constant speed on a line with equally spaced sidings. As simple as this problem is, the network of seven sidings and ten trains involves 19 meets.

In this diagram, trains perform "perfect meets" in that each pair of meeting trains arrives at a meet location in a manner that minimizes delay to either train. In practice, meets cannot be accomplished so precisely. At a minimum, one train must enter the siding, requiring it to slow for the diverging move through the turnout and the siding's typically slower track speed. There may also be delays to align turnouts and set signals. However, this scenario has a useful purpose in that it establishes the best possible capacity—the "theoretical capacity"—for this track network, a subject explored in greater detail later.

Figure 3.2 represents a nearly identical situation, but now the third eastbound train—EB3 (depicted by a dotted line)—is running 15 min late. Its arrival at Siding B will be delayed by 15 min, so its formerly perfect meet with the first westbound train—WB1—will be impacted. Train WB1 is now held in Siding B for 15 min, as indicated by the vertical line introduced into the diagram to indicate delay, as time

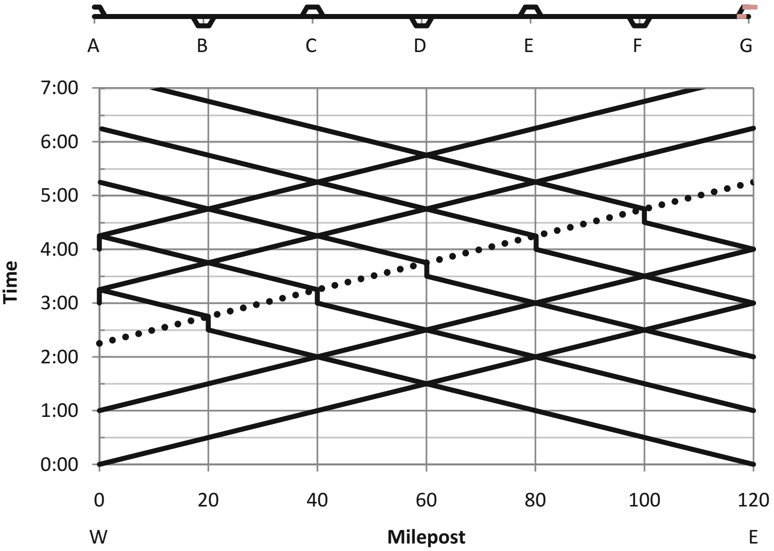


Fig. 3.2 Exhibit II

increases without the milepost changing. This 15 min delay to EB1, in turn, delays its arrival at Siding A to meet train EB4, which, in turn, requires it to be held at Siding A for 15 min. The departure of Train EB5 will be similarly delayed, as will the arrivals of Trains WB2, WB3, WB4 and WB5. So, a delay of 15 min on one train has translated into delays on seven of the other trains depicted in this diagram, producing a total delay of 105 min.

Figure 3.2 presents a reasonable resolution to train meet conflicts posed in this example—hold westbound trains in sidings and delay departure of eastbound trains. What if some trains in one direction were more important than trains in the other? This could occur if loaded trains ran one way and empties the other, or if a grade in one direction made it undesirable to stop a train at a siding. In this case, delays to trains in the preferred direction—let us call them superior trains—should be avoided when possible. If eastbound trains EB4 and EB5 in this example (depicted by dashed lines) are superior, westbound trains meeting them could be held in sidings to avoid any delay to these eastbound trains. Figure 3.3 depicts trains dispatched in this fashion. Overall delay has increased from 105 min to 600 min, but none of the delay is to the superior trains.

One can see how rapidly the resolution of train conflicts becomes a challenge. When the following real-life factors are introduced, conflict resolution becomes extremely problematic:

- Running time variability—The slope of the train lines in the time–distance plot represents train speed, and therefore the running time between sidings. Many factors affect running times:
  - Maximum authorized speed—this may be related to train type, higher for passenger, lower for freight trains

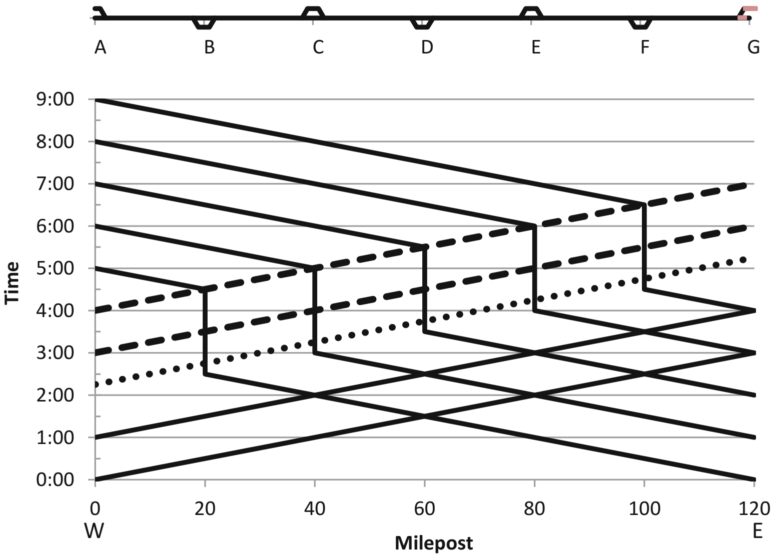


Fig. 3.3 Exhibit III

- Speed restrictions below maximum authorized speed—factors like curves, road crossings, and slow orders will limit speed
- Grades—steep grades may prevent trains from reaching the maximum authorized speed
- Train make-up (horsepower per trailing ton)—intermodal trains will have higher horsepower per trailing ton, unit trains will have less

- Relative train priorities—the example above illustrates how superiority by direction can impact conflict resolution. In practice, relative train priority can be quite complex:
  - Often, the traffic the train carries dictates its relative priority, with intermodal trains being superior to unit trains.
  - When trains of the same priority have conflicts, train direction, schedule adherence or other factor will determine their relative priority.
  - Priorities can change over time, a function of schedule adherence, hours-of-service considerations, and other factors. For example, an intermodal train running 4 hours early may be delayed in favor of a merchandise train running 4 hours late, and both may be held to get a unit train to the crew change point before the crew exceeds its hours of service.
  - Most train conflicts are between opposing trains, but conflicts can also arise between trains moving in the same direction. Advancing one train around another on line of road is similar in nature to a motorist passing a truck on a two lane road, something to be avoided when possible. Often, such “overtakes” or “passes” can be resolved at terminals during crew changes or other

terminal work, where the inferior train is simply held until the superior train has departed. If conducted on lines with single track and sidings, the pass will be performed with the inferior train in the siding, allowing the superior train to hold the main and move at track speed. On lines with multiple tracks, the inferior train may proceed on one track while the superior train passes it on a parallel track, a move that can consume an enormous amount of track capacity, impacting the ability to meet opposing trains.

- Track network—The track network—the arrangement of sidings, single and multiple tracks, terminals, control systems—is perhaps the most critical factor in line-of-road performance. It determines the theoretical capacity of a line—the number of trains that can be moved over a route with minimal delay under ideal conditions. The features of the track network that impact dispatching include:
  - Number of tracks—single main with sidings or multiple main track
  - Siding spacing—distance between sidings affects running time
  - Spacing of crossovers (sets of switches that enable a train to move from one track to another)—distance between crossovers on multiple tracks affects dispatching options
  - Length of sidings—a siding shorter than maximum train length makes meeting or passing trains there more problematic
    (a) If only one of the two trains exceeds siding length, the conflict can be resolved (typically with the shorter train taking the siding).
    (b) If both are longer than the siding, the conflict must be resolved at another location.
  - Siding location
    (a) Are sidings equally spaced (have equal running time)?
    (b) Is the siding free of highway crossings which cannot be blocked? If not, the first train to arrive will be held short of the crossing until the other train arrives, producing increased train delay.
    (c) Is the siding on a heavy grade? If so, the train moving up the grade should not be stopped, so the opposing train will typically be held in the siding.
  - Intermediate terminals—can be a source of congestion, especially if road trains must hold the main to set out or pick-up cars or change crews
  - Industry tracks—road train movement will be less impacted if industry tracks allow locals to work industries without fouling the main
  - Railroad crossings at grade—meets and passes are the most typical types of conflicts, but a third type arises at railroad crossings, where trains of other railroads cross the tracks of the home road, producing conflicts with train movements on the home road. Unlike the other types of conflicts, these are seldom scheduled, may not be known in advance and may occur erratically. Typically, one of the two roads involved physically controls dispatching over the crossing, and the other road's trains will proceed only after the controlling road's trains have been dispatched.

- Foreign road control—to avoid the high cost of track construction and maintenance, a railroad might use another railroad's track. Most often, this takes the form of trackage rights, where the track's owner allows another road to operate trains over the route. The owner will control dispatching, so the tenant's trains may be treated unfavorably.
- Signaling and other movement control systems—while railroads employ a variety of systems to control train movement, all systems have one element in common—movement authority is controlled by the train dispatcher. The systems differ in how this authority is communicated and implemented, which has a major impact on the efficiency of train movements.

  (a) Track Warrant Control—this system, deployed extensively on non-signaled lines, relies on radio communication between the dispatcher and the train crew to communicate and acknowledge movement authority. The limits of such authority may be a station sign (fixed block) or a fractional milepost (moving block). Once a train has reached the limits of its authority, the train crew advises the dispatcher to release the block, then receives, acknowledges and implements a new movement authority.

  (b) Automatic Block System (ABS)—in this system, a track is segmented into a series of blocks, where the occupancy of each block is indicated by signals at the ends of the block. The minimum length of a block is the distance required to bring a train to a stop from full speed (an approach (yellow) signal indicates that the next signal is set to stop (red)). Importantly, the presence of a clear (green) signal in such a system does not constitute the granting of movement authority—it simply indicates that the track in the next block is not occupied.

  Double track—in this situation, each of the two tracks has a specified direction of travel, much as highway lanes do. In North America, right-hand running predominates. Once a train has received authority, typically by radio, to enter the appropriate track of the pair, it can proceed on signal indication—continue at track speed if the signal is set to clear (green), prepare to stop at the next signal if this signal is set to approach (yellow), and stop if the signal is set to stop (red). Its movement authority does not permit movement on the other track, nor allow reverse movement out of its current block; such movements have to be authorized by the dispatcher.

  Double track ABS allows trains to follow one another efficiently, as long as the trains are running at the same speeds. However, many railroads operate a mix of slow, low priority trains (e.g., unit trains of bulk materials) and various faster, higher priority trains (e.g., passenger, intermodal, merchandise). In this case, priority trains may be delayed as they follow slow trains preceding them.

  Single track—in this system, trains must receive authority, typically by radio, to enter the single track segments between sidings, but once entered, the train may proceed on signal indication to the next siding. Tracks are no longer operated directionally. The train crew will be responsible for throwing and restoring the switches to enter and leave a siding, when necessary.

(c) Centralized traffic control (CTC)—this system changes both the communication and implementation of movement authorities. Dispatchers now directly control key signals (known as home signals), and the aspects that the signals display conveys movement authority to the train. A crew, viewing a clear (green) on a home signal is now authorized to occupy the track beyond the signal. As with ABS, further movement is governed by signal indications that reflect track occupancy ahead. CTC also permits the dispatcher to directly control switches and set signals to indicate the alignment of the switches. An approach (yellow) or other restricting signal will convey to the train crew that the switch ahead is lined for the siding or other diverging route. Crews are no longer required to handle the switches themselves, greatly speeding train movement.

- Signal systems have a positive impact on dispatching efficiency and track capacity, but they do limit how closely trains can follow one another. Following trains must be separated by at least one block, preferably two or more to provide a margin of safety. On high-density lines, these signal effects, known as signal wakes, must be considered when performing dispatch analysis.
- Defect detectors—another feature of the track network that will impact train movements is the location and number of train defect detectors. These devices are mechanical and electronic systems that monitor bearing temperature, look for dragging equipment (e.g., derailed cars, low-hanging hoses or other items that might pose a derailment hazard), cars whose wheels are out of round (e.g., have flat spots), etc. When these devices detect a defect, train crews are directed to stop and inspect their trains, possibly dropping off defective cars at a nearby siding. Clearly this may cause unexpected delays on heavily trafficked lines.

- Scheduling—besides the track network, scheduling has the greatest impact on line-of-road performance, playing a major role in determining how much of a line's theoretical capacity can be realized. Ideally, trains could be scheduled so that the track capacity demands they produce best match the track network's ability to supply the necessary capacity. This is unrealistic for a number of reasons:

  - Scheduled and unscheduled trains—some trains run on a regularly scheduled basis, since their traffic is tied to daily business cycles. Intermodal trains, for example, often depart their origin in the evening with traffic that arrived throughout the business day, and are targeted to arrive at their destination early morning, to allow shipment delivery when business first opens. On the other extreme are trains originating at mines, chemical plants, grain elevators or ports, where commodities are produced throughout the day and trains are dispatched when fully loaded. Now, no standard or default dispatching plan can be developed because each day's operation is different.
  - Peaking—on lines where train traffic is tied to daily business cycles, there will be a natural bunching of train schedules, with late-evening departures and early-morning arrivals. In this case, demands on the track network's capacity

peaks during one period, rather than being spread throughout 24 hours. The trains operated in this period will likely experience greater delay than if they had been dispatched evenly throughout the day.
- Train priorities—schedules determine how many trains of different priorities will operate over a line segment, impacting dispatching decisions as described previously.
- Interference by maintenance of way, local/industry switching activities and curfews—Track capacity will be seized to repair track or to switch an industry or interchange track, producing delays to road trains. On a line where traffic peaks during one period, these activities would best be scheduled during a non-peak period, typically during daylight hours. Curfews are sometimes imposed as well; freight trains running on commuter lines around Chicago cannot operate during morning and afternoon rush hours. In these cases, road trains will be held until the track is freed up.

3.2 Fundamental Elements for a Dispatching Algorithm

Before designing a dispatching algorithm, several fundamental tools must be available and a critical concept should be fully understood. The tools relate to moving trains over the track network, while the concept ensures that the solution will be feasible.

Think about the stringline diagrams above and the slopes of the lines depicted there. The slope reflects train speed and, therefore, running time between sidings, a critical factor in any solution. But where did these running times come from? If the trains are short passenger trains with nearly instantaneous acceleration and deceleration moving on level, straight track, the track speed limit and the siding distance could establish the running time between sidings with sufficient accuracy. In more realistic cases where heavy freight trains operate on undulating and curvy terrain, running time would not be so easily calculated. Analysis of historic running times can provide baseline data, but if the track network or train speeds are to change, such historic data are no longer valid. What is needed is a tool that models train operation, considers speed limits and track and train characteristics, and predicts train speed and running time. Such a tool is called a Train Performance Calculator (has also been called a Train Performance Simulator or Train Performance Model).

A Train Performance Calculator is a program that applies the physical laws of motion to compute train acceleration, deceleration, velocity and running times. Newton's second law, Force = Mass × Acceleration, is the cornerstone of the calculations. Starting from a stop, a train's acceleration is determined by the net accelerative force present at the given velocity—locomotive tractive effort (pulling power) minus train resistance (rolling, curve and air resistance to movement) and plus or minus grade and wind resistance. When stopping, a train's deceleration is similarly determined by the net declarative force—braking force (railcar brakes, engine air brakes and engine dynamic brakes) plus train resistance plus or minus grade and wind resistance. Because the forces change as a function of velocity and because the grades, curves and speed limits are a function of location, calculations must be made at small time increments, often 1 s or less. Many inputs will not be known precisely: the tractive effort of old locomotives and the braking characteristics of old cars may not perform according to their original specifications, train weights are often estimates and may vary from expected values, and wind forces may be unknown. A Train Performance Calculator can be constructed to predict a train's performance given a locomotive engineer's specific throttle and braking commands, but, more generally, the program is written to predict how an engineer would operate the train. The human engineer's look-ahead planning must be integrated into the program to ensure that the train obeys speed limits and stops at the desired locations, a requirement that greatly complicates the construction of a Train Performance Calculator and can cause different Train Performance Calculator programs to differ in their results and their replication of actual train performance.

The second critical tool related to moving trains over the track network is a routing engine. This is needed to determine if an alternative route exists if a specific track is occupied. The alternative may be running through a siding if the main adjacent to the siding is occupied, or it may be using an alternative set of tracks in a complex terminal network. Generally, the dispatching algorithm employs a shortest path algorithm sensitive to track occupancy and relative track speeds. Proper calibration of this tool favors North America's right-hand running preference and ensures that all trains hold the main whenever possible, not darting into and out of sidings unnecessarily. A number of commercial tools have been known to display such unrealistic behavior, which immediately undermines their credibility.

Now that movement through the network has been addressed with a TPC and a shortest path algorithm, it is critical to introduce a concept that ensures solution feasibility. The concept is known as "Safe Point" or "Safe Harbor" dispatching. It ensures that the dispatching algorithm does not behave like some commercial dispatching software that produces a "lock-up" situation where no feasible solution can be found.

The safe point concept basically requires that no train be advanced from its current location unless it can reach a safe point where it will be clear of conflicts with other traffic. Such a point might be a terminal at the end of the subdivision where there are multiple tracks to hold trains clear of the main, or could be a junction with another subdivision that has two or more tracks. A safe point could be an industrial location where a local can clear the main and hold there until further movement can be made. In general, a safe point is a location where a train can clear the main so as not to interfere with other movements. A siding may or may not be a safe point; consider the example in Fig. 3.4 where advancing the train from Siding A to Siding B would "lock-up" the network. Commercial software lacking safe-point logic can make such a foolish decision.

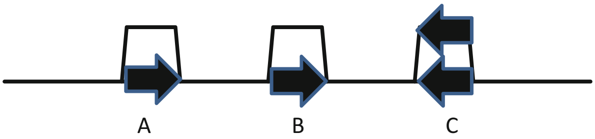


Fig. 3.4 Exhibit IV

3.3 Developing a Dispatching Algorithm

Dispatching algorithms are necessarily complex, and successful ones must address each of the many factors discussed above. Many papers have been written that address in detail specific elements of the problem, going far beyond the intended scope of this chapter. Described below is a basic framework to resolve meet–pass conflicts without regard to many of the complicating factors. A generic and greatly simplified step-by-step procedure is first described, followed by an example employing that procedure.

3.3.1 Overview

Resolution of meet–pass conflicts proceeds as follows:

1. Identify the trains expected to enter the track network during the study period.
2. Sort the trains in the order of their next "event." For originating trains, an event will be its entry into the dispatching problem; for existing trains, an event will be its arrival at a safe point.
3. Using running times derived from a Train Performance Calculator or other source and a shortest path algorithm, project each train's route from its origin to its destination on the track network without regard to meet–pass feasibility.
4. Move clock time ahead from the current time to the time of the next event.
5. Analyze the event or events that occur at the time referenced in step 4. The event may be train origination or train arrival at a safe point. In either case, examine the train's potential conflicts with other trains and assess its movement options, realizing that several trains may compete to occupy the same track segment.
   (a) If the train is originating, consider whether to allow the train to enter the track network and advance to the next safe point, or whether to hold it out of the network.
   (b) If the train is arriving at a safe point where more than one track is present, a decision has to be made as to which track it should occupy (e.g., main or siding).
   (c) Typically, more than one dispatching alternative is available at any decision point. Each alternative must be enumerated and its value quantified. Use of a hierarchical decision tree structure may prove useful. Be sure to respect any "Track Reserved" status (see step 6) when identifying dispatching alternatives.
6. Advance trains based on decisions reached in step 5, setting a "Track Reserved" flag for any track segment occupied by a train being advanced. This ensures that the train has its exclusive use until the train clears into the next track segment, at which time the "Track Reserved" flag is cleared. The shortest path algorithm must respect this reservation when considering route options.
7. Repeat steps 2 through 7 until all trains have arrived at their destinations.

3.3.2 Example

Before computers were widely available, dispatching analysis was performed by hand, using time–distance (stringline) diagrams. Computers enabled the processes to be automated, but the sequence of processes itself is largely unchanged. Consequently, stringline diagrams will be used below to visually document how the processes work.

As before, the problem starts with a track network and a set of train schedules. To keep the problem of manageable size, the track network will be limited to four meet locations (two sidings and terminals at both ends) and two pairs of opposing trains. The number of possible train meet solutions is limited to 4^2 or 16. For simplicity, all trains operate at a uniform 40 miles/h and are of equal priority. Train overtakes (passes) are not specifically considered. The train schedule and running time information is provided in the table below.

Train name	Origin station	Destination station	Origin time	Minutes between A and B	Minutes between B and C	Minutes between C and D
E1	A	D	00:00	60	60	60
E2	A	D	01:00	60	60	60
W1	D	A	00:30	60	60	60
W2	D	A	01:30	60	60	60

An examination of the stringline diagram associated with this problem and drawn without regard to conflict resolution is shown in Fig. 3.5. Note that this diagram does not depict a feasible solution, as the opposing trains do not meet at a siding or in a terminal.

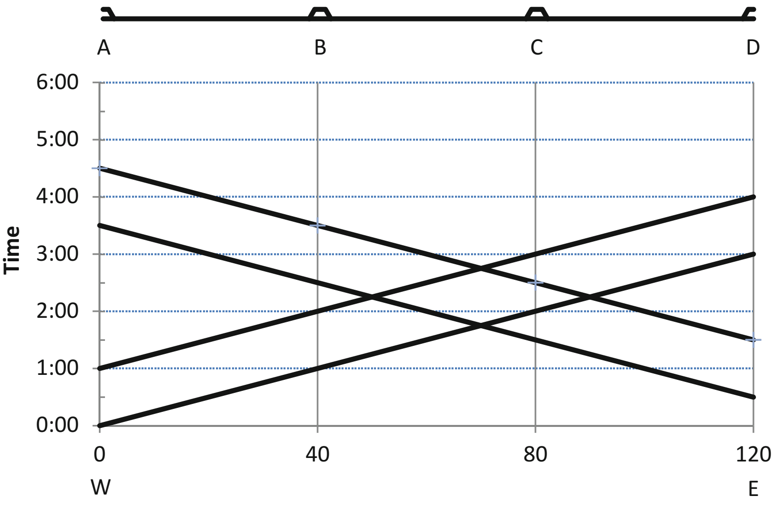


Fig. 3.5 Exhibit V

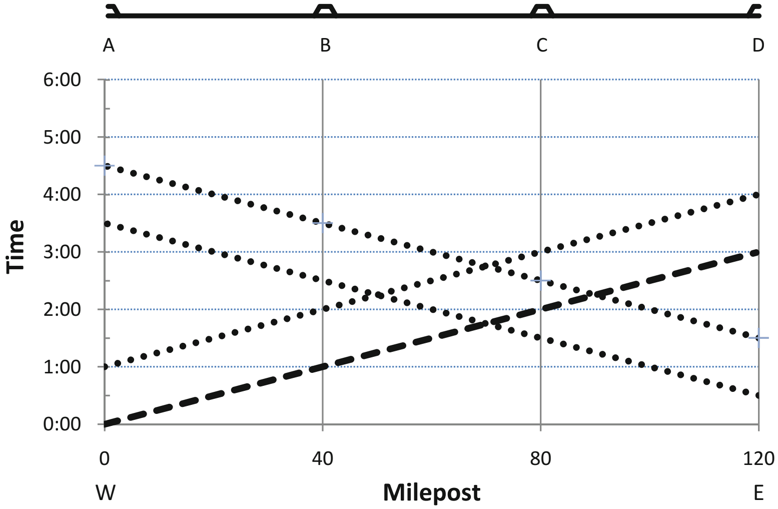
While it does not represent a feasible solution, the stringline diagram in Fig. 3.5 can be the starting point for the development of one. A major North American railroad has implemented a real-time system that depicts unresolved meet–pass problems in stringline format, asking the dispatchers to move train conflicts to meet points to make the plan feasible. The computer provides feedback to the dispatcher to guide his decisions by quantifying the value of the specified solution.

While a dispatching algorithm can probably be built that starts with an infeasible solution such as that in Fig. 3.5 and works to make it feasible, a better approach builds a solution incrementally, ensuring feasibility at every step. The algorithm would proceed as follows:

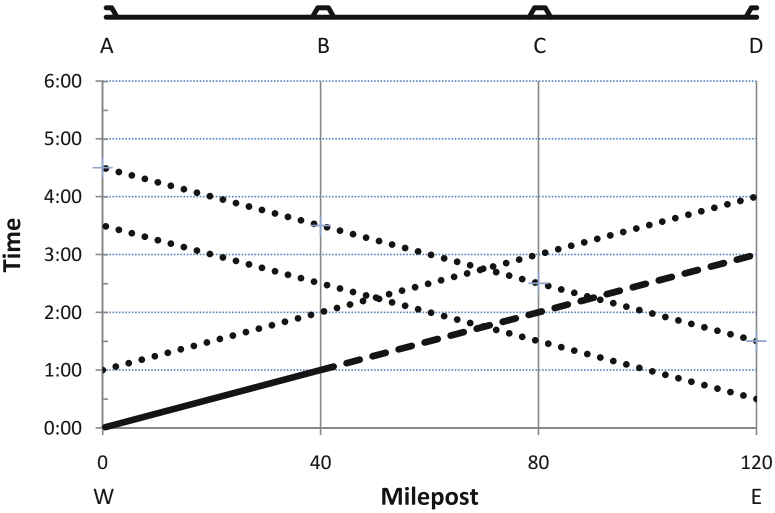
1. Sort the trains in the order they will enter the problem. Train E1 will arrive first, followed by train W1, E2 and W2, in that order.

Train name	Origin station	Destination station	Origin time
E1	A	D	00:00
W1	D	A	00:30
E2	A	D	01:00
W2	D	A	01:30

2. Using running times derived from a Train Performance Calculator or other source and a shortest path algorithm, project each train's route from its origin to its destination on the track network without regard to meet–pass feasibility.
3. Move clock time ahead to coincide with the first event at 00:00.
4. Bring the first train (E1) into the dispatching problem at its origin and highlight its route to its destination (dashed line).
5. Identify all trains that are likely to conflict with train E1 en-route to its destination. These would be trains W1 and W2 (and E2 if an overtake of E1 is contemplated), represented by dotted lines.



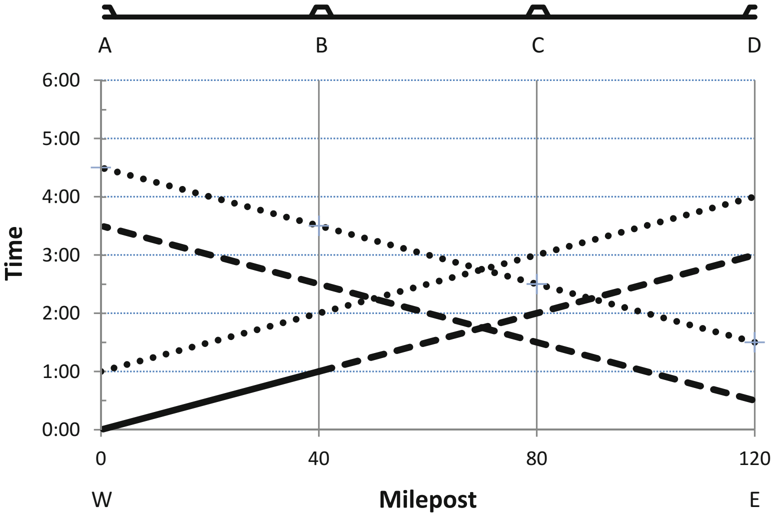
6. Taking into consideration the trains in step 5, determine the next safe point in the route of train E1. As no other trains have actually entered the problem at time 00:00, station B is clearly a safe point for train E1, as are stations C and D.
7. Advance train E1 from its current position at A to the next safe point in its route, which will be either track at B (represented by a solid line). Note that step 4 projected train E1 all the way to its destination at D, but that the dispatching logic has only authorized its movement as far as B, its next safe point, thereby preserving a broader set of movement options. Set or clear the appropriate "Track Reserved" flags.



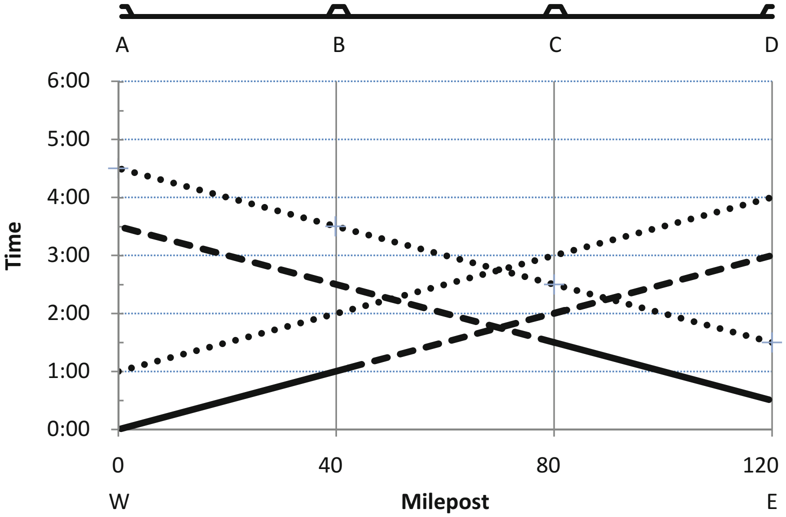
8. Update route and running time projections if necessary.
9. Determine the next event for each train. For train E1, this will be its 01:00 arrival at station B; for the others, it will be their originating time.
10. Sort the trains in the order of their next events. The first event will be train W1's origination at 00:30, followed, at 01:00, by train E1's arrival at station B and train E2's origination at station A, then the origination of train W2 at station D.

Train name	Next event	Next event station	Next event time
W1	Originate	D	00:30
E1	Arrive	B	01:00
E2	Originate	A	01:00
W2	Originate	D	01:30

11. Move clock time ahead to coincide with the next event at 00:30.
12. Bring train W1 into the dispatching problem at its origin and project its route to its destination (dashed line).
13. Identify all trains that are likely to conflict with train W1 en-route to its destination. These would be trains E1 and E2 (dotted lines).



14. Taking into consideration the trains in step 13, determine the next safe point in the route of train W1. Neither track at station C is currently occupied, making station C a safe point.
15. Advance train W1 from its current position at station D to the next safe point in its route, which will be either track at station C (represented by a solid line). Set or clear the appropriate "Track Reserved" flags.

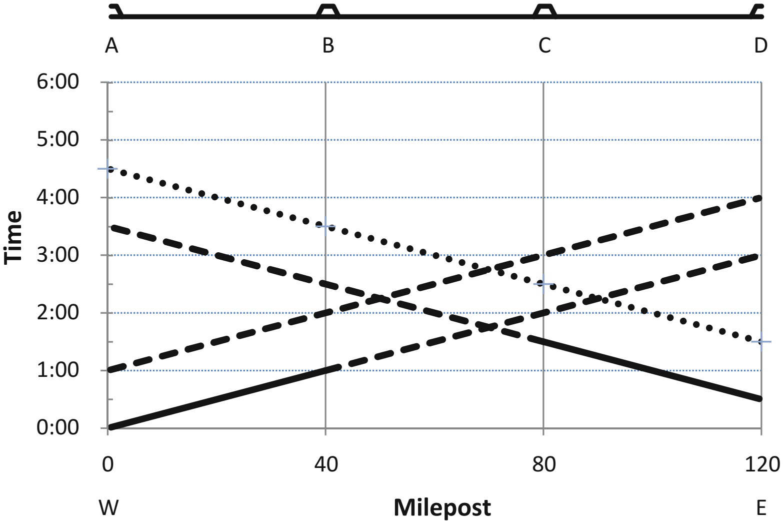


16. Update route and running time projections if necessary.
17. Determine the next event for each train. For train E1, this will be its 01:00 arrival at station B; for train W1, this will be its 01:30 arrival at station C; for the others, it will be their originating time.

18. Sort the trains in the order of their next events. Two events will occur at 01:00—train E1's arrival at station B and train E2's origination at station A; the next events, at 01:30, will be train W1's arrival at station C and W2's origination at station D.

Train name	Next event	Next event station	Next event time
E1	Arrive	B	01:00
E2	Originate	A	01:00
W1	Arrive	C	01:30
W2	Originate	D	01:30

19. Move clock time ahead to coincide with the next event at 01:00.
20. Bring train E2 into the dispatching problem at its origin and project its route to its destination (dashed line).



21. Identify all trains that are likely to conflict with train E2 en-route to its destination. As before, these include opposing trains W1 and W2, but now consideration must be given to preceding train E1 as well.
22. Taking into consideration the trains in step 21, determine the next safe point in the routes of trains E1 and E2.

    (a) Consider that station B does not represent a safe point for train E2 unless train E1 advances beyond station B, because advancing train E2 to station B while train E1 occupies the other track at station B could set up a lock-up situation. Consequently, train E2 has to be held at station A unless train E1 is advanced beyond station B.

(b) Determine whether station C constitutes a safe point for train E1.

i. Train W1 is only authorized for movement to station C, so it can be held there if necessary.
ii. Train W2 is not currently in the problem space, but can be held at station D when it originates at 01:30.
iii. Given i. and ii. above, station C is a safe point for train E1.

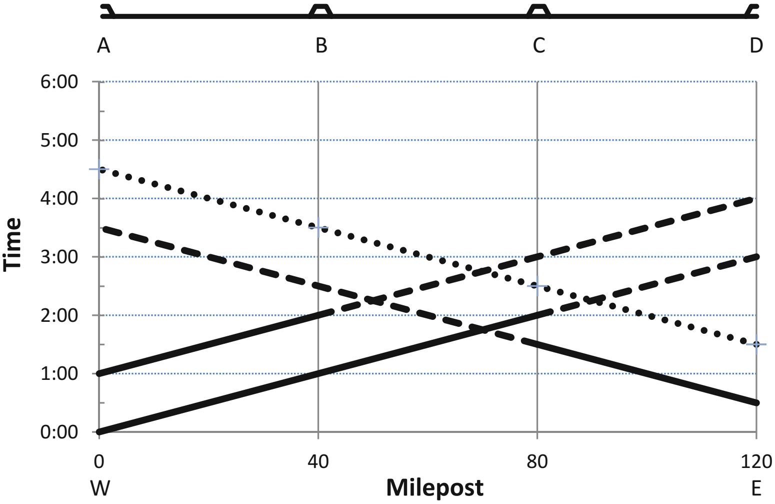
(c) Since train E1 can be advanced to its next safe point, station B constitutes a safe point for train E2.

Various dispatching alternatives exist at this point, so the decision can be made based on factors such as relative train priority, hours-of-service constraints and other considerations. For the purposes of this exercise, assume that it is decided to advance trains E1 and E2.

23. Advance train E1 from its current position at station B to the next safe point in its route, which will be either track at station C. Since train E1 will meet train W1 there, one train must take the siding while the other holds the main. An algorithm will determine which train occupies which track:

(a) If both trains are of equal priority and can fit in the siding, the first to arrive—in this case train W1—will generally take the siding and clear the main, enabling train E1—the second train to arrive—to hold the main and operate at track speed.
(b) If one train is longer than the siding, it is best practice to have it hold the main, accomplishing the meet faster than if the long train had to operate at siding speed while pulling through and exiting the siding.
(c) If one train has higher priority than the other, it will generally hold the main.
(d) Many other factors, including hours of service considerations, train lateness, etc., must also be considered.

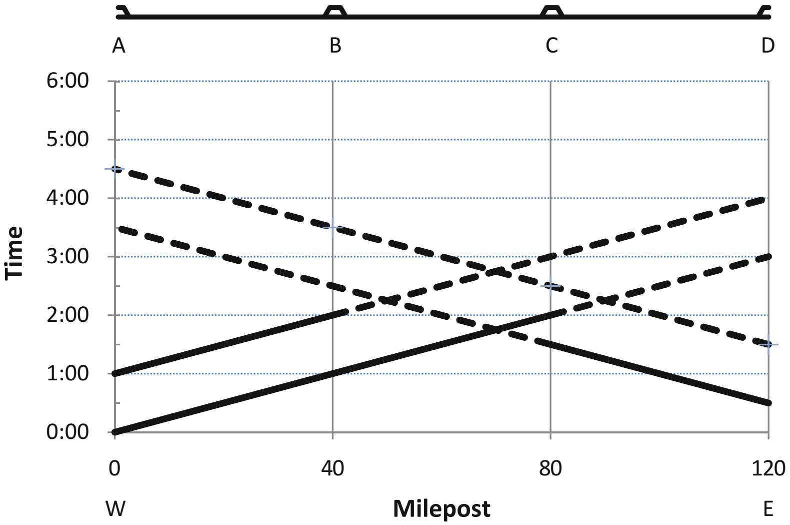
24. Advance train E2 from its current position at station A to the next safe point in its route, which will be either track at station B.



25. Set or clear the appropriate "Track Reserved" flags.
26. Update route and running time projections if necessary.
27. Determine the next event for each train.
28. Sort the trains in the order of their next events. Two events will occur at 01:30—train W1's arrival at station C and train W2's origination at station D. Two events will also occur at 02:00—train E1's arrival at station C and train E2's arrival at station B.

Train name	Next event	Next event station	Next event time
W1	Arrive	C	01:30
W2	Originate	D	01:30
E1	Arrive	C	02:00
E2	Arrive	B	02:00

29. Move clock time ahead to coincide with the next event at 01:30.
30. Bring train W2 into the dispatching problem at its origin and project its route to its destination.
31. Identify all trains that are likely to conflict with any train en-route to its destination. These will include both opposing and preceding trains.



32. Taking into consideration the trains in step 31, determine the next safe point in the routes of trains W1 and W2. (The other trains are en-route to their next safe point.)
    (a) Train W1 cannot advance beyond station C because train E1 is en-route to station C.
    (b) Since W1 must be held there, station C does not represent a safe point for train W2. Consequently, train W2 has to be held at station D.

    In this step, no trains are advanced (they are either en-route to their safe point or held at a safe point), and the logic continues at the next step.

33. Update route and running time projections if necessary.
34. Determine the next event for each train.
35. Sort the trains in the order of their next events. Two events will occur at 02:00—train E1's arrival at station C and train E2's arrival at station B. Trains W1 and W2 are currently held at stations C and D, respectively.

Train name	Next event	Next event station	Next event time
E1	Arrive	C	02:00
E2	Arrive	B	02:00
W1	Holding	C	–
W2	Holding	D	–

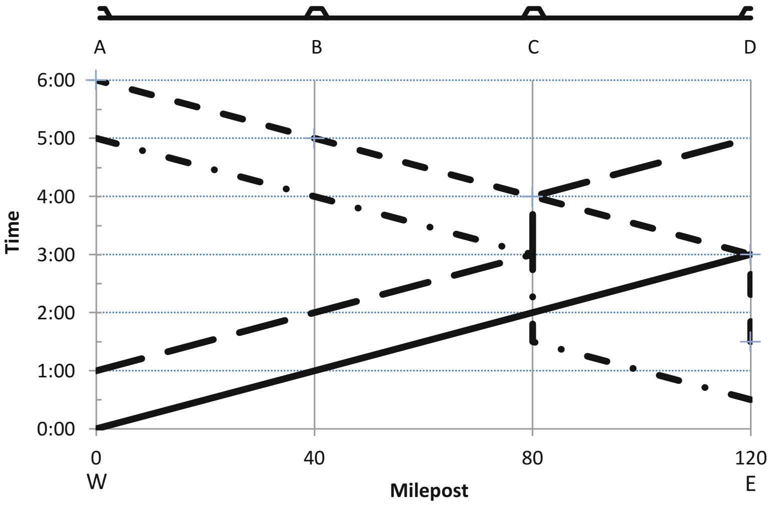
36. Move clock time ahead to coincide with the first event at 02:00.
37. As all trains are now in the dispatch problem, the next step is identification of safe points for each train:

    (a) Since tracks at station C are fully occupied by trains E1 and W1, station C cannot be considered a safe point for train E2 or train W2 unless train E1 or train W1, respectively, can be advanced beyond station C.
    (b) Station D is a safe point for train E1 only if train W2 is held at station D.
    (c) Station B is a safe point for train W1 only if train E2 is held at station B.

    Clearly, the resolution to this problem is advancing one or both trains out of station C. Here are the options:

    (a) Advance train E1 to station D, holding train W2 there. Train E2 can then be advanced to station C, or
    (b) Advance train W1 to station B, holding train E2 there. Train W2 can then be advanced to station C, or
    (c) Advance train E1 to station D and train W1 to station B, holding trains E2 and W2 in place.

    Any of these options provides a feasible solution, so the choice can be made based on factors such as relative train priority, hours-of-service constraints and other considerations. For the purposes of this exercise, assume that train E1 is advanced to station D and train E2 is advanced to station C.
38. In similar fashion, all remaining meets are resolved. Figure 3.6 provides a feasible and practical solution producing 240 min of delay.



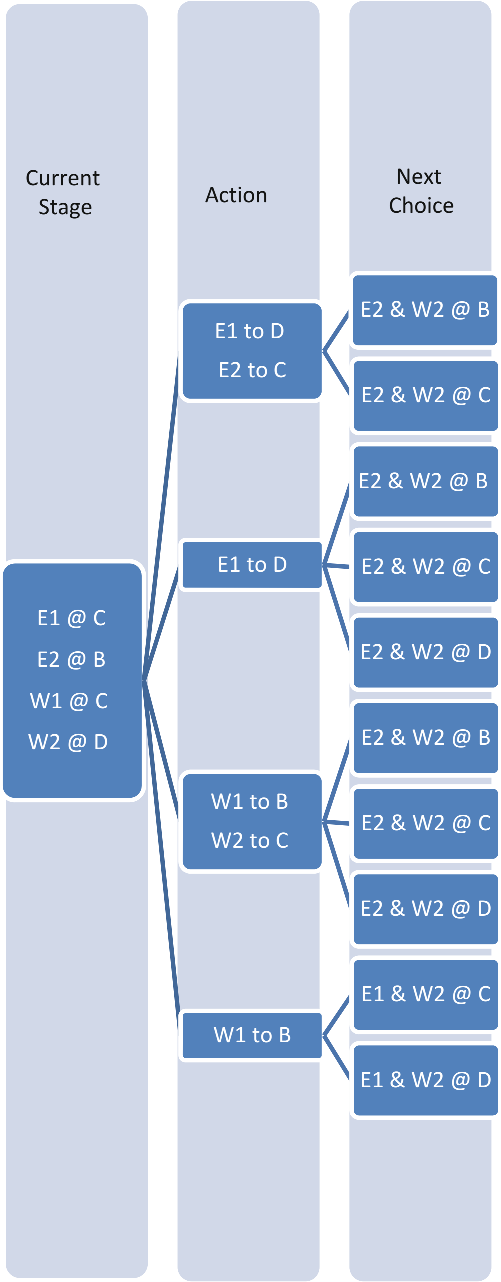
Train Name	Line Style in Exhibit	Met Opposing Train	At Station	Minutes Delay	Met Opposing Train	At Station	Minutes Delay	Total Delay
E1	Solid	W1	C	0	W2	D	0	0
E2	Long Dash	W1	C	0	W2	C	60	60
W1	Dot Dash	E1	C	30	E2	C	60	90
W2	Short Dash	E1	D	90	E2	C	0	90

Fig. 3.6 Exhibit VI

While this meet plan is feasible and achievable, is it a good plan? To evaluate its quality, the plan must be measured against some goal. The objective could be to minimize total delay, minimize delay to priority trains, minimize the number of trains exceeding their hours-of-service limits, maximize on-time performance and/or achieve some other target. Also, as demonstrated above, meet–pass planning is a sequential process where decisions made in a previous step affect subsequent steps, and, at each step, there may be a number of valid options available. Combining these two concepts, the idea of an objective function to quantify "goodness" and a dynamic programming/decision tree structure to organize and evaluate alternatives emerges.

Consider the alternatives being contemplated in step 37 above. Figure 3.7 depicts the decision tree at time 02:00. Where there were once 16 alternative meet plans to be considered, only seven remain, nine having been resolved at earlier stages. Each of the seven current options can be walked to its logical conclusion, branching where alternative choices exist in future stages. Note that some options require a coordinated decision to be executed at the same stage; often a preceding train must be advanced to free up a meet location. Delays to each train can be tallied for each option, and the quality of that sequence of decisions can be compared with the other alternatives to find the best.

Fig. 3.7 Exhibit VII



Based on that concept, Southern Railway (now Norfolk Southern Railway) in the late 1970s developed a real-time meet planning system relying on a branch-and-bound algorithm. At each stage, a feasible solution would be generated, its “goodness” established, alternatives identified, and the sequence of decisions would be walked until fully evaluated or until their value exceeded the previously discovered best. This effort, described in a paper by Sauder and Westerman (Interfaces 13: 6 December 1983, pp. 24–37), earned Southern Railway recognition as a finalist in the 1983 Edelman Award competition.

Since that time, many algorithms have been developed, some finding commercial success for planning purposes. One of the most successful is the Rail Traffic Controller (RTC) from Berkeley Simulation Software, LLC, which has been widely adopted by railroads and consulting firms for analysis of proposed capital and operating changes. There have been recent advances in development and deployment of real-time systems for line of road operations as well, with Norfolk Southern and GE Transportation Systems installing a Movement Planner component for NS' Unified Train Control System (UTCS). However, challenges remain in both planning and real-time environments when road trains must operate within complex terminal areas—a challenge which can only be resolved when yard operations can be fully considered in the line of road dispatching tools.

3.3.3 Simplified Assumptions

The dispatching example provided above does not fully capture the complex logic of dispatching simulations. As an example of a key simplification, consider that in step 22(a), train E2 cannot be advanced beyond station A unless train E1 advances beyond station B. The question arises, how can E1 be assured that it can advance beyond station B? The algorithm above alluded to the answer—use of a resource request or reservation system that allocates a resource—a main line track, a siding or a track at a terminal—to a specific train for a specific time period, denying that resource to other trains. In practice, this is key to practical dispatching solutions and warrants far greater consideration than can be provided here. The reservation may pass through several stages, from pending to confirmed, as dispatch logic ensures its feasibility and desirability. A pending reservation can be assigned to another train; a confirmed reservation is sacrosanct.

Also, to simplify the description of the algorithm, attention has been focused exclusively on train meets, which is generally the larger part of resolving train conflicts. However, planning for train passes (overtakes) is essential as well. Fortunately, logic developed for meet planning can be expanded to consider alternatives where superior trains must pass preceding trains. In many ways, a train pass is similar to passing another vehicle on a highway—it is accomplished over a large distance. On single track, one siding is effectively unavailable for train meets while the passing train operates over the track segments on both sides of that siding. In multiple main territory, the two tracks between two sets of crossovers are similarly unavailable for train meets (a crossover is an arrangement of switches that enables a train to move between tracks). When possible, passes should be accomplished at terminals while the inferior train is delayed for crew change or some terminal work.

Operation in multiple main track territory, overlooked here, obviously makes dispatching logic far more complicated, as does the introduction of railway signaling. At a minimum, signals impact how closely trains can follow one another on the same track, and may cause trains to run significantly below maximum track speed.

Since the lines typically analyzed by dispatching software are high-density routes, they are very likely to be signaled, so the signal system's impact should be captured in the software.

3.4 Future Directions

Train dispatching is an important function in railroading today. Years of line rationalization and corporate mergers have shrunk the rail network while traffic volume has continued to increase. Cost-saving initiatives create pressure to reduce dispatching staffs at a time when dispatchers enter their positions with far less railroad experience than dispatchers of a generation ago. Consequently, railroads increasingly look to operations research for help. To date, the efforts have primarily focused on two areas: real-time dispatching and integrated road and terminal planning.

Computers have long been able to enumerate and evaluate meet–pass alternatives, but, given the current state of the art, the best dispatchers can often find better solutions. Consequently, real-time computer-based dispatching systems were initially sought to simply free the dispatcher from the many tedious tasks—mostly paperwork and recordkeeping—that distract from his primary responsibility, which is the development and execution of high-quality meet–pass plans. Systems are now emerging that provide the dispatcher with insights and recommendations, which makes the human–computer interface a critical element in system design. At present, such systems can improve the performance of a less-skilled dispatcher, but may still not perform as well as the best dispatcher.

While railroads have developed computer programs for both line of road dispatching and terminal management, there has been inadequate coordination between the two systems. Until recently, an active real-time line dispatch program managed meet–pass planning on either side of a terminal—but not within the terminal area itself. Slowly, interfaces are being developed to enable line-of-road and terminal operations to be coordinated. In time, a yard will programmatically communicate what time and in which order it wishes to receive or depart trains, and the line of road dispatching system will integrate these considerations into its meet–pass planning.

Chapter 4
Car Scheduling/Trip Planning

Carl Van Dyke and Marc Meketon

4.1 Introduction and Background

Prior to the widespread adoption of unit trains and the rise of intermodal, most traffic moved in “loose car” or “manifest” service (also called “carload traffic”). In this type of service, sets of railcars are grouped together on a temporary basis into blocks. A block is a group of cars that may have disparate origins and destinations, but will be moved together as a group from one point to another before being broken apart and formed into other blocks. These blocks are moved by trains, where each train may carry a single block, or may carry multiple blocks. In this manner the cars are relayed from their origin to their destination by being placed in a series of blocks, which are moved by a series of trains. This series of blocks and trains represents the core of what we know as a “trip plan” or “car schedule.” See Chaps. 1 and 5 for a discussion of the train scheduling and blocking.

Based on the above, two questions become foremost in determining the car schedule or trip plan for a railcar:

1. What block should a shipment be placed in given its current location?
2. What train should be used to advance the block to its destination?

All existing Class I trip planning or car scheduling systems are built around these two questions. In part this is because it represents the commonly used logic to route railcars found throughout the industry, and in part this is because all of these systems share a common intellectual heritage.

C. Van Dyke (✉)
TransNetOpt, Princeton, NJ, USA
e-mail: carl@cvdzone.com

M. Meketon
Oliver Wyman, Princeton, NJ, USA
e-mail: marc.meketon@oliverwyman.com

© Springer Science+Business Media New York 2015

B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3_4

Trip plans are used for a variety of purposes. Assuming that a railroad achieves reasonably high adherence to the carload operating plan, then trip plans underlie the entire carload management process, providing a forward view of expected train sizes and yard workloads, instructions for the make-up of trains, and the basis for overall performance monitoring. They also provide a means of providing predictive arrival times to customers and for the management of empty railcars.

Southern Pacific Railroad (SP) in the late 1960s and early 1970s developed the TOP system, which was one of the earliest systems for tracking the location of railcars, and using computers to determine the car-to-block assignments. TOP and variants of it were adopted by many other railroads (IBM; Railway Age 2014), including Burlington Northern (BN), Canadian National (CN), and the Missouri Pacific (MP). In the late 1970s the Missouri Pacific, with funding assistance from the U.S. Federal Railroad Administration (FRA), extended its version of this system (called TCS) to include the generation of trip plans by adding logic to select the train a block should use on top of the blocking system already built into the TCS system (FRA 1980, 1981). Thus was born the "car scheduling system," the concept of a computer-generated trip plan or car schedule, and the basic logic that is used by all Class I railroads today.

The Missouri Pacific system survives today in the form of the TCS system currently in use at the Union Pacific Railroad. It is the author's understanding that individuals fully familiar with this system used it as a design guide in the development of a new car scheduling system for the Santa Fe railroad in the late 1980s or early 1990s. This system was subsequently purchased by the Canadian National (Murray 2006; McBain 2000), and is the parent of the systems used by both the CN and BNSF at the present time. Based on discussions with CSX, the authors have learned that other individuals familiar with the design of the Missouri Pacific car scheduling system went on to create the system in use at CSX. While the legacy of the Norfolk Southern (NS) trip planning system is somewhat different, the core building blocks and logic are the same, due in large part because this system was developed after the others cited above had already been created, and the creators of the NS system were very familiar with the concepts used in the car scheduling process.[1] Based on the authors work with Canadian Pacific, we know that as a purchaser of the NS system, Canadian Pacific also became an adoptee of the basic trip planning or car scheduling logic employed by the other Class I railroads.

In general, the nature of railroading has changed significantly since the 1970s when the foundation of the trip planning or car scheduling systems was laid. Railroads have gone from 75 % or more carload traffic to a complex mixture of service offerings including dedicated unit trains, separately operated intermodal trains, complex gathering systems for grain and other products, and a variety of other specialized services. Carload traffic now represents anywhere from a high of about 50 % to a low of about 20 % of the overall traffic handled by an individual railroad (see chart below). For the most part, the computerized car scheduling or trip planning systems remain focused on only the carload segments of the business, with

[1] One author of this chapter had a direct role in the design of the NS system, and bases this statement on this experience.

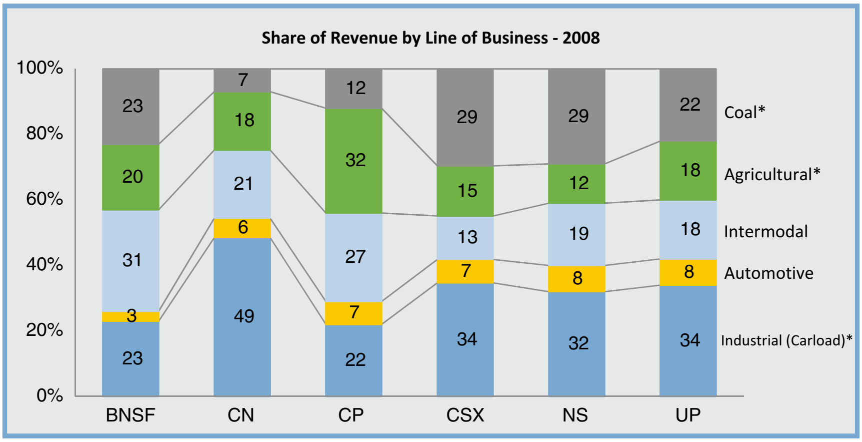


Fig. 4.1 Share of revenue by line of business—2008. Source: 2008 annual reports and investor fact books produced by each railroad. *"Coal" is largely unit trains and may include coke and iron ore in some cases; "Agricultural" often moves in unit trains and may include some finished or consumer products; "Industrial" may include some bulk products. Note: Revenue tends to overstate some lines of business in terms of cars handled, and understate others. This difference can be up to ± 30 %; that is the market share numbers for a product such as coal could be up to 30 % higher when expressed in carload, and for manifest up to 30 % lower

some railroads having made extensions to address intermodal and to a limited extent commodities such as grain. The coal segment moves almost exclusively in unit trains, as does a large portion of the agricultural segment, while the intermodal and much of the automotive traffic tends to move in dedicated, single purpose trains. By and large the unit train segments of the business are handled externally to the car scheduling process through either separate systems or manual processes (Fig. 4.1).

Given the primary focus on the carload business segment, our descriptions and discussions of the trip planning systems and processes will focus first on how these systems approach the "classic" carload trip planning problem. We will then use this carload framework to discuss variations on the carload logic to address other segments of the business.

4.2 Car Scheduling/Trip Planning Systems in Context

To generate trip plans, and to fully leverage the information they provide, a relatively large and complex set of systems must be developed, and a significant pool of accurate supporting data must be available. The overall relationships and components of a trip planning or car scheduling system are depicted in the diagram presented in Fig. 4.2:

Many of these components are described in the sections that follow, including their overall role in the trip planning process, the logic that they employ, and the user groups and systems that they interact with. The following is a brief guide to these components:

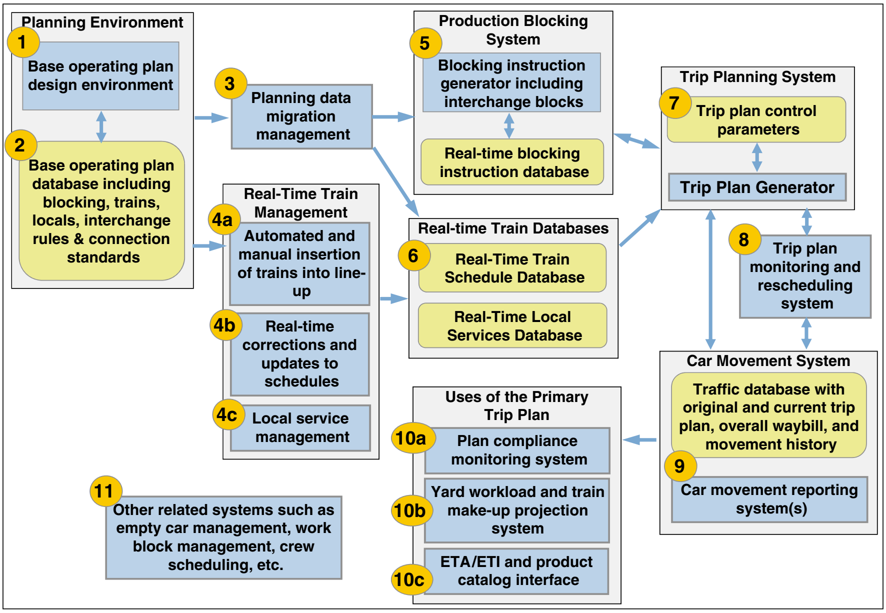

Planning Environment
1
Base operating plan design environment
2
Base operating plan database including blocking, trains, locals, interchange rules & connection standards
3
Planning data migration management
Production Blocking System
5
Blocking instruction generator including interchange blocks
Real-time blocking instruction database
Trip Planning System
7
Trip plan control parameters
Trip Plan Generator
Real-Time Train Management
4a
Automated and manual insertion of trains into line-up
4b
Real-time corrections and updates to schedules
4c
Local service management
Real-time Train Databases
6
Real-Time Train Schedule Database
Real-Time Local Services Database
8
Trip plan monitoring and rescheduling system
Car Movement System
Traffic database with original and current trip plan, overall waybill, and movement history
9
Car movement reporting system(s)
Uses of the Primary Trip Plan
10a
Plan compliance monitoring system
10b
Yard workload and train make-up projection system
10c
ETA/ETI and product catalog interface
11
Other related systems such as empty car management, work block management, crew scheduling, etc.

Fig. 4.2 Trip planning (or car scheduling system) system relationships

1/2. *Planning environment*: much of the trip planning process and train management process is based on a set of blocking, train schedule, local service, and other specifications that are maintained by one or more planning groups, working in a railroad-specific planning environment. See Chaps. 1 and 5 on train planning and blocking for more details.

3. *Planning data migration management*: this is the mechanism for the plans created by the planning groups to be moved or migrated to the real-time production systems. How this is done can influence the accuracy of the plans and the timeliness of any updates to these plans. This topic is explored briefly in this chapter.
4. *Real-time train management*: this information system provides the interface to manage and capture train activities such as the dispatching, annulling, and consolidating of trains, the addition of extras, and the changes to the schedules or timing of trains, and provides the inputs to the databases in item 6 below.
5. *Production blocking system*: as discussed earlier, the grouping of railcars into specific blocks is one of the two core elements of the trip planning process—this system provides those car-to-block assignments, and is discussed in Chap. 5 on blocking.
6. *Real-time train databases*: to generate practical trip plans, we need to know the actual trains to be operated by the railroad, not a set of theoretical trains—this is the source of those actual trains. This subject is discussed briefly in this chapter, and the development of the underlying train schedules is explored in Chap. 1 on train plan design.

7. *Trip planning (car scheduling) system*: this is the core of the trip planning or car scheduling process, using a variety of inputs, plus built-in business logic to generate specific trip plans, and is the primary focus of this chapter.
8. *Trip plan monitoring and rescheduling system*: it is an unfortunate fact of life that railcars do not always adhere to their original trip plan—as a result the status of each railcar against its plan must be monitored, and new trip plans generated when the trip plan becomes infeasible or invalid.
9. *Car movement system*: as railcars advance through the system, their progress is recorded in this database, which typically includes the overall waybill or movement instructions for a car, the history of events the car has experienced, and one or more trip plans related to the movement. The basics of this system are described in this chapter.
10. *Uses of the primary trip plan*: there are various systems and users of the trip plans. From this chapter's perspective, the core users are the yard management system, train management system, and performance monitoring system. All of which are explored briefly in this chapter.
11. *Other related systems*: there are many other systems that use the results of the trip planning process. These include customer service, empty railcar management, systems and users needing ETA and ETI (estimated time of arrival/ interchange) estimates, sales, costing, and financial analysis. These users are touched on briefly in this chapter.

We must remind the reader that the above processes are largely focused on only the carload business. While fairly straight forward variations of this logic can be used for business segments such as intermodal or automotive, it is much more difficult to address the needs of products such as grain or coal using the "classic" car scheduling system. The solution to this fundamental incompatibility may be the creation of specialized systems to address segments such as grain or coal, or it may be to create specialized variations of the existing trip planning processes. We explore these issues further later in Sect. 4.7.

Each of the internal components of the "classic" car scheduling or trip planning system are described below, along with a discussion of the other lines of business handled by the railways, and potential alternative approaches. Planning techniques, and the potential for applying algorithms and optimization to the process are also explored.

4.3 Plan Compliance and the Value of Trip Plans

It is important to note that while the trip plan generator or "car scheduling engine" is at the core of the process of generating trip plans, the quality of what it produces is only as good as the data it is supplied with. Furthermore, the utility of producing trip plans is highly dependent on both the availability of systems to take advantage of the trip plans, and existence of an operating philosophy that values both adherence to the plan and the use of the information the trip plans and related systems contain.

Trip plans can be used to forecast the workloads at yards, understand the expected size of specific trains, and help to direct the manner in which cars are classified and trains are built. However, if actual operations bear little resemblance to assumptions built into the blocking plan and base train schedules, then the trip plans, and the workload forecasts based on them, will also be equally meaningless. Thus, trip plans only become valuable if the underlying operating plan is achievable, and largely adhered to by the field. The greater the level of compliance to the operating plan, the greater the value of the trip plans.

The value of compliance to a well thought out operating plan is well recognized (FCUP 1981; Harrison 2005; Norfolk Southern 2002, 2003, 2004). As a result, it is the author's observation that most Class Is strive to be above 75 % compliant with the initial trip plan, and above 95 % compliant with the specific blocking and train make-up instructions at an individual yard. Keep in mind that if the typical railcar undergoes a local switch at origin and destination, plus three intermediate classifications, that represents a total of five potential connection failures. If each operates at a 95 % level of reliability, this yields an overall plan compliance rate of only about 77 %!

4.4 Current Industry Practices: Basic Car Scheduling/ Trip Planning Concepts

The generation of railcar schedules based on the operating plan is a well established process that is used by major railroads in North America, Europe, and elsewhere. The basic idea is to start with the current location, and determine the block the car should be placed in for movement of the car to the next location, and then to determine the exact train to be used to move the block (see Chaps. 1 and 5 on train and blocking plan design). This process is then repeated until the destination of the shipment is reached. Figure 4.3 depicts the core of this process:

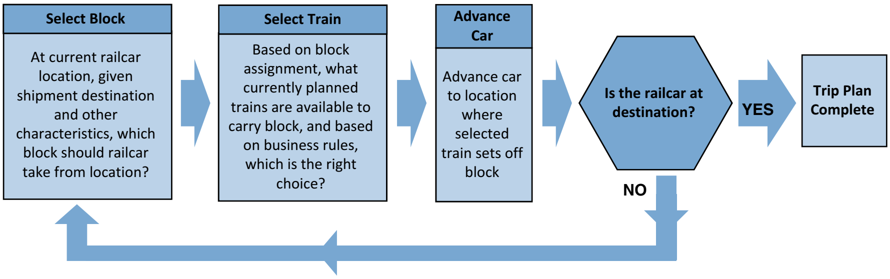

Select Block
At current railcar location, given shipment destination and other characteristics, which block should railcar take from location?
Select Train
Based on block assignment, what currently planned trains are available to carry block, and based on business rules, which is the right choice?
Advance Car
Advance car to location where selected train sets off block
Is the railcar at destination?
YES
Trip Plan Complete
NO

Fig. 4.3 Trip planning process

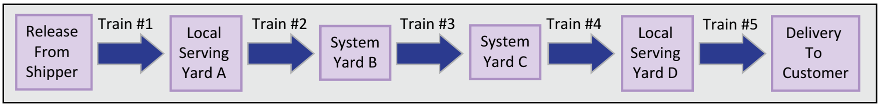
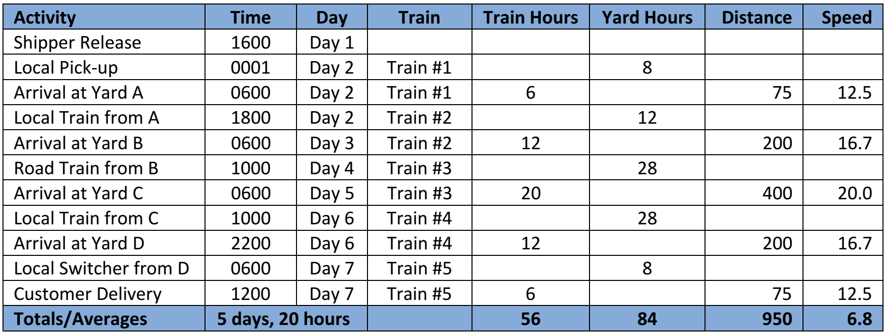


Fig. 4.4 Overview of a trip plan



Activity	Time	Day	Train	Train Hours	Yard Hours	Distance	Speed
Shipper Release	1600	Day 1					
Local Pick-up	0001	Day 2	Train #1		8		
Arrival at Yard A	0600	Day 2	Train #1	6		75	12.5
Local Train from A	1800	Day 2	Train #2		12		
Arrival at Yard B	0600	Day 3	Train #2	12		200	16.7
Road Train from B	1000	Day 4	Train #3		28		
Arrival at Yard C	0600	Day 5	Train #3	20		400	20.0
Local Train from C	1000	Day 6	Train #4		28		
Arrival at Yard D	2200	Day 6	Train #4	12		200	16.7
Local Switcher from D	0600	Day 7	Train #5		8		
Customer Delivery	1200	Day 7	Train #5	6		75	12.5
Totals/Averages	**5 days, 20 hours**			**56**	**84**	**950**	**6.8**

Fig. 4.5 Detailed trip plan with summary statistics

Putting the above into context, an overall trip plan itinerary might look something like the following (Fig. 4.4):

In table form, the trip plan would appear as follows (Fig. 4.5):

While the above represents the core of the trip planning process, there are a number of issues that can complicate the overall process. These include:

- *Yard-blocks/train-blocks*: perhaps for historic reasons, most blocking systems do not provide a definition of a car to yard-block assignment in terms of a block origin, destination, and block name. Instead, they provide a "yard-block code," which is variously referred to as a "tag" or "class code," or in the case of CSX Transportation an "IYSC" or "inter-yard switching code." The exception to this is the NS/CP system, which produces a full block definition including a destination (Norfolk Southern). In most systems, trains specify a separate concept called a "train-block" that provides the pick-up location for the train-block, the set-off location, and a train-block name. Yard-blocks (class codes/tags) are then associated with the train-block. More than one yard-block can be assigned to the same train-block. This is done to provide visibility to subsets of the traffic in a train-block (both codes are displayed by most systems), and to allow sets of traffic to be easily shifted from one train or destination yard to another for capacity management purposes. Since the yard-blocks (class codes) do not have a destination, the destination becomes the location where the train-block is set-off. On the one hand, this makes it very hard to validate that appropriate class codes have been assigned to a particular train-block; on the other hand, it also provides flexibility to send the same class code/yard-block to different locations by day-of-week or based on other factors related to the available train service. See Chaps. 1 and 5 on train and blocking plan design for further discussion of this topic.

- *Block swaps*: a block swap is defined as the movement of a group of cars (a block) from one train to another on an intact basis without intermediate classification. For example, if a block is made at A, destined to C, but the train sets off this block at B instead, for pick-up by a second train, the activity at B is called a block swap. The benefit of a block swap is that it can help create larger trains by adding more blocks to them, and move blocks that do not easily support direct train service, while reducing intermediate switching work at the yards and the associated delays. However, it can also create:

  – More complex train operations
  – A potential loss of network capacity when the block is set-off at a siding
  – Additional delays and costs at the block swap location

  In environments such as those used at NS and CP, a block swap can be defined as a case where the set-off location does not equal the block destination. In systems that use yard-blocks or class codes where the destination of the block is unknown, block swaps become essentially impossible for the computer system to identify. As a result, in many of these systems two things occur. First, the existence of block swaps are identified in the design notes for a train, and become a manual management issue to carry out. Second, because the block swap set-off looks like any other set-off, the cars being set-off are passed through the classification system, and a new yard-block or class code is obtained. The result is that significant effort must go into ensuring that the same class code is produced at this intermediate location as was generated at the block origin in order to maintain a correct specification for the car's trip—this can become a significant maintenance headache for the planning group. Various extensions have then been applied to the train specifications and car scheduling systems to identify and properly protect these types of connections.

  Local services: the movement of railcars to/from industry often poses special challenges that require an alternate set of specifications for trains and blocks. This is caused by several factors, including the nature of how local switching services are provided, the "addresses" for customers, and the large number of unique customers that must be served.

  – One factor to consider is the number of customers that need to be served. The trip planning system must have a means of generating a solution to every station and customer that might generate a railcar movement, not just the ones that consistently generate such movements. A local train that serves all of the customers along a line might have 50, 100, or even more potential customers within its service area. If we were to generate a block, and a block-to-train assignment, to and from each and every one of these customers, this could result in a set of local trains that had dozens, or even hundreds of blocks on them. To avoid this, many railroads have systems that allow local trains to be defined as serving a range of stations or customers, without specifying specific blocks. While effective from a data management perspective, this results in the need for special logic in the blocking system and trip planning system to handle these alternate train definitions and "implicit" blocks. This also makes the local service component difficult to model and optimize during the planning processes, or in a simulation.

  - A second factor is that some local blocks and trains do not leave the area covered by a single station or node. For example, customers and interchanges to other railways that take place within a yard area are all specified with a single station number. Most train scheduling systems require that a train go to more than one station. To specify yard switching operations that serve customers and interchanges at the yard, special trains must be designated that do not match the pattern of other trains and require special logic for blocking and trip planning purposes. Furthermore, one has the need to assign a class code or yard-block to the car movement at the destination location. Normally, once one has reached the destination for a trip, there is no further action required. However, in the case of customers or interchanges located at the yard, it is likely the yard will need to switch these cars into specific blocks for delivery to these customers even though the destination has been reached. The result is the need to support the designation of a "final class code" for each shipment, and special logic to determine when these class codes are required. Special timing rules, or "single station trains" are then used to represent the time factors and capture the workloads in the car scheduling system associated with these intra-station activities.
  - A final factor to consider is that a single station may contain multiple customers. Most of the train schedule, block, and trip planning processes are built around the concept of the station. However, when providing local services one must operate at the level of the specific customer, and in some cases for large customers a specific siding at the customer's site. This results in a second addressing system below the level of the station. Often called the zone-track-spot (ZTS) system, it goes by many names across the industry. Most blocking systems need to have overrides of some form to assign block codes by customer and/or ZTS type information, and train services must have ways of specifying the timing of services to be provided at the ZTS level. Generally this is handled within the process of generating final class codes, and the specification of local train services. Upstream from the destination, all railroads support the ability to limit blocks to specific customers, and some railroads support the ability to apply ZTS related rules to ordinary road blocks. Other complications may arise when road trains provide local switching services en-route, raising the need to specify the specific customers to be served by these trains.

- *Interchange blocks/run-through trains*: railroads often enter into agreements with other railroads to build blocks for each other (called "pre-blocks"), and in some cases to operate "run-through" trains with the other railroad. A variety of issues arise with respect to these pre-blocks and run-through trains.

  - For the pre-blocks, the railroad making the blocks generally agrees to make a certain number of blocks for the other railroad for delivery at a specific interchange, and the foreign railroad specifies which cars (destinations) should go into each block. In some cases the railroad simply codes the definition of each pre-block into its blocking system, and in others, a special process is used to obtain the block designation from the foreign road. The most common situation is to do both. When a car is assigned to a block to be interchanged, the car is first assigned a block based on the railroad controlled blocking system. At the same time, a message is sent off to the foreign railroad (called a 419 message in North America) requesting a block designation. Eventually an answer is received (called a 420 message), and the block assignment for the car is updated based on this message. In general, the railroad is free to make the pre-blocks anywhere it wants, as long as it is delivered to the correct interchange with the correct content. Many railroads build the same interchange block in more than one location if that makes for a more efficient operation. Maintaining the local version of the pre-blocking rules requires constant vigilance to stay consistent with the foreign road's requirements, and requires good data on the final destination of the cars on the foreign railway.

– The next level up from pre-blocking is the creation of run-through trains, where entire single or multi-block trains are created and passed to the other railroad on an intact basis. In some cases special logic is required to specify these trains since their routes extend off of the railroad's home network.
– Both pre-blocks and run-through trains pose particular problems for the trip planning process, particularly for cars coming onto the railroad at an interchange. Many interchanges receive both pre-blocks and local interchange blocks. The local blocks are generally switched at or near the interchange location by the receiving railroad, while the pre-blocks are often moved straight through to a much more distant location on the receiving railroad before they are touched. In general, cars for specific online destinations at an interchange will be found in both a specific pre-block, and in the local interchange block. This occurs for a variety of operational reasons, and does not necessarily represent a failure of the foreign road to follow instructions. In many computer systems it is often difficult or impossible to determine if a car received at interchange is in a specific block, and as a result the trip planning system must "guess" if the car will be locally switched or moved straight through to the pre-block destination. In some cases the foreign road may provide the blocking for each car through electronic messages, and this can then be used to determine the classification of the car. Unfortunately this is more the exception than the rule, and most railroads must make assumptions about which block the car is in when it is received, resulting in some number of incorrect trip plans.

Clearly the two most basic keys to this process are selecting the right block, and selecting the right train. All of the Class I railroads have separated this process into two sub-systems, one to address blocking, and one that uses the blocking information in combination with the trains to generate the actual trip plans. As mentioned earlier, there are separate chapters on blocking and train plan design that explore these subjects in more detail (Chaps. 1 and 5). Train selection and to some extent block selection are discussed in this chapter.

4.4.1 Current Industry Practices: Block Selection Logic

The first step in the generation of a trip plan or car schedule is to select the “block” that the shipment should be assigned to at the current yard. This is done based on the current location of the railcar, and a variety of attributes related to the shipment and the railcar itself. A combination of these user supplied block specifications and a set of business rules are then used to select a specific block for the shipment. Included in this process are a variety of special cases that address some of the issues cited above including block swaps, interchange blocks, and local blocking. These systems fall into two types: table driven and algorithmic. NS and CP use algorithmic blocking, while the other major railroads in North America use table driven solutions. Both approaches are discussed in Chap. 5 on blocking.

In general, most of the railroad blocking systems produce only a “yard-block” or “class code” for the railcar, as discussed above. The exception to this is NS/CP system, which produces a full block definition including a destination, which in principle makes the identification of block swaps much easier. Some other railroads are working to better define the destinations of blocks within their classification systems in an attempt to address some of the short comings of the class code-based approach.

The bottom line is that the block selection process is able to provide a block assignment based on the current location of the rail car in terms of either a full block definition or a class code.

4.4.2 Current Industry Practices: Train Selection Logic

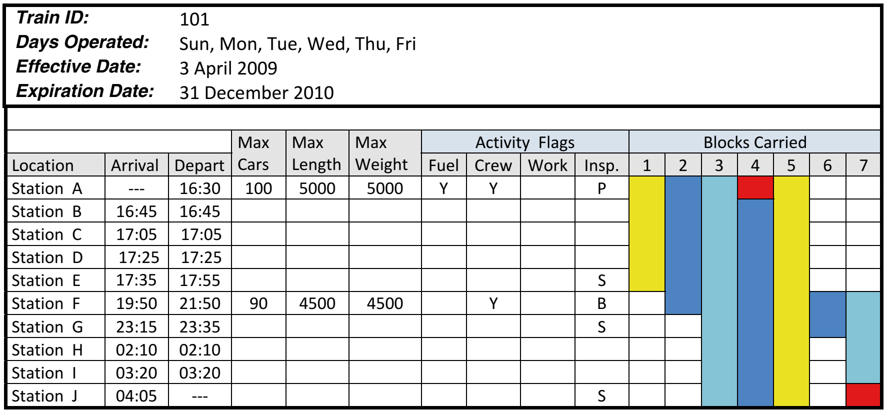
The train selection process can be viewed as a two-step process. In the first step, the eligible trains that can carry the block are identified; and in the second, a specific train is selected from among the eligible trains.

Train eligibility is determined through the use of block-to-train assignments. Every train consists of a number of data components. The four-core components are an overall description of the train (symbol, effective dates, days operated, train type, etc.), a train route (locations that will be visited, train timings by location, and special actions such as crew changes, inspections, etc.), connection standards (rules on how long it will take to process shipments connecting to a train), and a set of block-to-train assignments. It is the block to train assignments that determine which trains are eligible to carry the block.

A typical train schedule might appear as follows (Fig. 4.6):

Each train may carry a number of blocks. For each block the pick-up location, set-off location, and block attributes are specified. Based on the block the shipment has been assigned to at the current location, all trains that pick-up that block at the shipment’s current location are identified.

Even if only one train symbol can pick-up the block, different instances of that train likely operate on various days of the week. These date-specific trains are viewed as independent from each other for trip planning purposes, and each is considered a separate, eligible train to carry the block, and thus the railcar.



Train ID:	101
Days Operated:	Sun, Mon, Tue, Wed, Thu, Fri
Effective Date:	3 April 2009
Expiration Date:	31 December 2010

			Max	Max	Max	Activity Flags				Blocks Carried						
Location	Arrival	Depart	Cars	Length	Weight	Fuel	Crew	Work	Insp.	1	2	3	4	5	6	7
Station A	---	16:30	100	5000	5000	Y	Y		P	🟨	🟦	▨	🟥	🟨		
Station B	16:45	16:45								🟨	🟦	▨	🟦	🟨		
Station C	17:05	17:05								🟨	🟦	▨	🟦	🟨		
Station D	17:25	17:25								🟨	🟦	▨	🟦	🟨		
Station E	17:35	17:55							S	🟨	🟦	▨	🟦	🟨		
Station F	19:50	21:50	90	4500	4500		Y		B		🟦	▨	🟦	🟨	🟦	▨
Station G	23:15	23:35							S			▨	🟦	🟨	🟦	▨
Station H	02:10	02:10										▨	🟦	🟨		▨
Station I	03:20	03:20										▨	🟦	🟨		▨
Station J	04:05	---							S			▨	🟦	🟨		🟥

Fig. 4.6 Representative train schedule with block display (*yellow* and *blue* colors represent different block categories, *red* represents a block swap)

It is important to note that during actual operation of the railroad, the near-term trains are likely known with greater precision than the trains to be operated further into the future. Thus, most trip planning systems use "dated" or actual trains in the near term, and planned or "template" trains further out in time. As train schedules change, trains are added, annulled, etc., the near-term dated train schedules are updated so that these changes are reflected in the trip plans. Thus, the train schedules can still be somewhat dynamically created, as long as an up-to-date, complete, forward view of the plan is maintained in the computer system with a 7- to 14-day planning horizon.

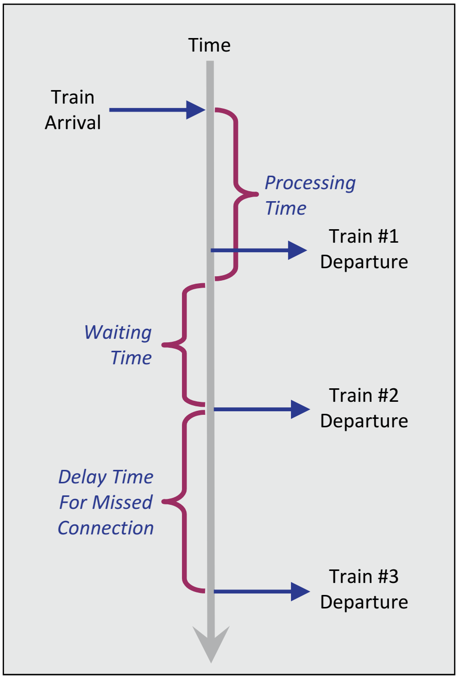
One can visualize a typical intermediate connection for a railcar as follows (Fig. 4.7):

In the above scenario, a railcar arrives on a train. It must then be processed and prepared for movement on an outbound train. This processing includes in-bound inspection, the switching of the car into a specific classification track containing the designated outbound block, and the assembly of that block with others into a train.

There is generally a minimum processing time that a yard is willing to commit to for the completion of the processing of an in-bound car and placement of it into a specific train. In the trip planning process, this is generally called either the "connection standard" or "cutoff time." In some systems it is expressed in terms of an elapsed time, for example, "the connection standard is 12 hours." In others, it is expressed as a clock time relative to a specific out-bound train (there are also examples of setting the cut-off based on using the in-bound train, or a combination of the in-bound and out-bound train). For example, if a train is leaving at 18:00, one might set the cutoff to be 06:00, meaning only railcars arriving in the yard prior to 06:00 can make the 18:00 departure.

Cutoffs can be very simple, or very complex. Some trip planning systems support a variety of both global and train specific cut-offs or connection standards. The global standards typically apply to standardized situations, such as movements

Fig. 4.7 Illustration of connection times. Trains #1, #2, #3 carry the necessary outbound yard-block. Train #1 leaves too soon, Train #2 would be the intended train, and Train #3 would be used if for some reason there was a missed connection for Train #2



to/from industry, movements to/from interchange, block swaps, and conventional intermediate connections. In most cases the train-specific cut-offs or connection standards are either yard-based or train-based rules that specify a specific standard for some combination of the following four core factors:

- In-bound train
- In-bound block (train-block and/or yard-block)
- Out-bound train
- Out-bound block (train-block and/or yard-block)

At some railroads, additional factors can be applied as well, such as the connection type (industry, interchange, etc.) or highly specialized knowledge of a customer's operating hours or shipment preferences. These connection standards can also be used to specify shorter connection times for block swaps compared to regular classification events. While these connection-specific standards can be useful for fine tuning the trip plans they also come at a price. This price exists both in terms of pushing individual yards to provide customized processing on a train-by-train basis, and the potential for creation of many thousands of rules that must be maintained. Thus, railroads often try to use these types of standards as sparingly as they can.

Going back to the train connection diagram shown above, the typical trip planning system develops a timetable or "line up" of all of the eligible outbound trains for a particular block. The minimum connection time for each outbound train is determined, and the earliest departing outbound train that is later than or equal to the minimum connection time is identified and selected as the train to be used to advance the car to the next yard. In the above example, this would be Train #2.

Based on this train selection process the railcar is then advanced to the next yard, and the process is repeated until the destination is reached. This process of course does not account for issues related to capacity constraints on trains, and the related trade-offs against other shipments also trying to get on the same train. It also does not consider the potential for benefits from taking a later train that might have a more favorable schedule for the block. Thus, most current car scheduling systems are "uncapacitated" and "myopic" (greedy). While there is a general interest in the industry to move toward capacitated solutions, the fact that many shipments do not reliably follow their initial trip plans due to a variety of factors tends to make fully capacitated solutions impractical. The issues related to capacitation, reservations, and the simulation of prospective operating plans are discussed further later in this chapter and in Chap. 8 on simulation.

4.4.3 Current Industry Practices: Other Special Considerations

Several considerations that modern car scheduling systems include are:

- Data clean-up issues
- Plan accuracy, falling off the plan, and the self-correction process
- Capacities
- Fill blocks, extras, and annulments

Each of these topics will be touched on below.

1. *Data clean-up issues*: a large number of data sources are used as inputs to the trip planning process, and many of them can have data issues associated with them. As a result, most trip planning processes have a variety of mechanisms for correcting these data issues in order to improve performance. In most cases, the core document or data record that drives the trip planning process is the waybill. As a result, almost all railroads have some form of waybill correction process. One of the most common elements of these waybill correction processes is to change the origin or destination of traffic. For example, waybills will often designate the destination using a generic code or station number for a place like Chicago or Toronto. This designation is not enough to determine exactly where in Chicago or Toronto the car is to go, but is sufficient for billing purposes. As a result, the correction process will look at things like whether the car is being locally terminated or interchanged, who the customer is, what type of car or commodity is involved, and many other factors. Based on this, a more specific destination will be applied. Other types of corrections include applying the results of reroute agreements with other railroads to use specific interchanges, making customer spellings more consistent, enriching the data with line-of-business or service type designators, or applying preliminary or final pre-block designations.

2. *Plan accuracy, falling off the plan, and the self-correction process*: the accuracy of trip plans depends on many factors including the general adherence to the plan by the railroad, proper understanding of pre-blocking assignments at interchanges, quality of the business logic, quality of the underlying data, etc. In general, railroads have two trip plans for each car—a benchmark plan and a current or active plan. Railroads create an initial plan for each shipment when the car is released by the customer. This initial plan is used as a benchmark for measuring service quality and plan compliance. In some cases this benchmark plan may be replaced with a revised plan for a variety of reasons, such as not knowing the interchange block that a car was placed in, or gaps in the data on the origin local service to be used. Nonetheless, at some point early in the movement of a railcar, a benchmark plan is established. At the start, the benchmark plan and the current plan are the same. Over time, they may diverge. A monitoring process exists that is constantly checking each current trip plan for accuracy, and updating or replacing the current plan when it is found to be no longer valid. A variety of strategies exist to determine when the plan should be checked. These strategies range from regenerating the trip plans every time an event occurs that may be related to the car, to regenerating all trip plans on a fixed time basis, or watching for specific conditions that may warrant a regeneration of the trip plans. Each has its pros and cons, but the end goal is to correct the trip plans so that the current trip always represents the best estimate of what will occur in the future. The most common triggers for reviewing and updating the trip plans are the arrival and departure of trains at handling locations, and cases where trains are expected to be off schedule by more than some prescribed level of tolerance.
3. *Capacities*: the short summary is that no North American railroad is taking capacities into account in its trip planning processes, and we are only aware of one railroad worldwide that currently does so on a broad basis (Green Cargo in Sweden). In essence, most uncapacitated systems assume that all the eligible trains being considered in the trip planning process have sufficient room to carry the shipment being trip planned. If a railroad operates with a reasonably high adherence to the plan, then the forward projections of train sizes based on the trip plans will reflect the likely situations where trains will be under or over capacity. Based on this information, the line managers can make decisions on when to run extra trains, and when to fill out light volume trains. The specification of capacities and their potential use in both planning and actual operations, along with related reservation concepts, are explored further later in this chapter.
4. *Fill blocks, extras, and annulments*: most railroads support the designation of block-to-train assignments as either primary blocks or fill blocks. The concept behind a fill block is that it will only be used if the train is below capacity. In general, most trip planning systems will not automatically use a fill block, and thus will never reflect this type of assignment in a computer-generated trip plan. However, when the self-correcting monitoring system discovers that the car has been placed in a fill block on a train, it will then auto-correct the current plan to reflect this assignment and update the plan appropriately. Field operations may also add extra trains, or annul trains. When this happens, the dated train database used

by the trip planning system is updated to reflect these actions. While annulments will always be reflected in the updated trip plans, use of extras will depend on how they are designated, and if their block-to-train assignments are designated as primary or fill.

4.5 OR Challenge: Typical Reasons of Trip Plan Failures

There are three primary dimensions of trip plan "failures":

1. Failure to generate an end-to-end trip plan of any form
2. Failure to generate an accurate trip plan
3. Failure of the shipment to comply with the trip plan

Each is discussed briefly below.

- *Trip plan generation failures*: in most systems, trip plans are generated incrementally from the origin point to the destination, as depicted in Fig. 4.3. An information failure at any point along this process can result in a trip plan failure. The most common failure modes are (a) an inability to identify a block for a shipment out of a specific yard, and (b) no identifiable assignment for the block to a train. In some cases these failures are in effect planned, because the trip planning system is not designed to address a particular class of shipments such as unit train movements. In our experience, one of the most common failure modes is the identification of local blocks and trains to move the shipment to the first serving yard. Most blocking systems are "forward looking" and focused on the serving yards and above. This results in good blocking information from the serving yards and other larger yards to the next locations, including the final destination, but tends to mean that gaps exist in the blocking information at customer origin sites that fall outside of the normal definition of a yard. Data integrity checks on the operating plan can go a long way toward identifying these problems and addressing them.
- *Trip plan accuracy failures*: these are cases where a trip plan is generated, but the result is a plan that does not "make sense" or adhere to the actual operations of the railroad. For the most part these are simply plan definition errors. For example, a block with a complex definition for what traffic should be included is created or modified, and the specification is not quite correct and causes some traffic to be either inappropriately included or omitted. There are analytic techniques that can be used to identify some of these situations such as testing for the "ideal" block for each shipment versus the currently assigned block. Many other "accuracy" failures are in reality operating plan compliance failures that might be caused by execution issues or by an impractical plan design.
- *Shipment compliance failures*: the causes of such failures are myriad. Examples include:
  - Failure of a shipment to be switched into the correct block
  - A late inbound train causing the expected outbound train to be missed

- Slow processing of a shipment such that it misses its expected outbound train
- The annulment of a train
- The advancing of a shipment onto an earlier train
- The operation of an extra train to carry the shipment
- A lack of capacity on a train to take a shipment
- Rerouting of a train due to local, short-term operational issues
- Rerouting of a block due to local, short-term operational issues
- Delay of a shipment due to mechanical problems
- Loss of paperwork or routing instructions for a shipment
- Rerouting of a shipment based on customer requirements.

Internal studies with which the authors are familiar of major classification yards in the 1980s and 1990s found that the failure rate to make the expected outbound train could run 20 % or more. The industry has made great efforts to improve their performance and reduce failures of the types listed above, but they remain an issue. Addressing these types of failure issues remains the largest barrier to using trip planning systems to directly manage capacity and the industry moving to more of a "reserved" model for customer orders.

4.6 Trip Plan Output Usages

Trip plans are used for a variety of purposes. Assuming that a railroad achieves reasonably high adherence to the carload operating plan, then trip plans underlie the entire carload management process, providing a forward view of expected train sizes and yard workloads, instructions for the make-up of trains, and the basis for overall performance monitoring.

To understand how trip plans can do all of this, it is important to understand that trip plans can be viewed in two different manners. First, they can be viewed as a series of train movements (Fig. 4.8):

Second, they can be viewed as a series of yard connections or connection events (Fig. 4.9):

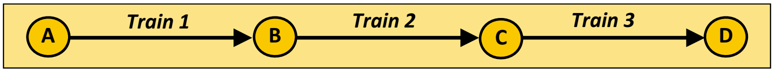


Fig. 4.8 View of a trip plan as a sequence of train movements

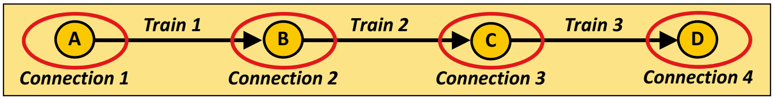


Fig. 4.9 View of a trip plan as a sequence of connection events

If one organizes the trip plans by date-specific train, and sums the individual car movements up, one gets a forward view of the expected size of each train and the specific traffic it will carry. If one organizes the trip plans by connection, and sums the individual car movements up, one gets a forward view of the expected workload at each yard, and the train make-up instructions for each train including the required in-bound to out-bound connections.

From a customer service and overall network view, monitoring performance against the trip plans provides a focus on customer service and asset velocity. It also provides a means for both identifying operational failures to adhere to the plan, and helps to identify where the plan may need refinement.

There are many other uses for trip plans beyond those cited above. Some include the provision of ETAs to customers, estimated times of interchange (ETIs) to other railroads, supporting the means to do proactive monitoring of service-sensitive traffic, providing transit time estimates for empty railcar distribution, providing service and cost factor data to sales, finance, planning and other departments, providing insight into what services can be offered to customers, and supporting customer enquiry tools showing the service the customer would receive when shipping on the railroad.

A number of these uses are explored in more detail in later sections of this chapter.

4.7 OR Challenges: Alternate Approaches to Car Scheduling and Special Cases

The above discussion has focused on the "traditional" approach to trip planning or car scheduling. The traditional approach is best characterized as a "fixed schedule, location centric, uncapacitated approach." Here is what is meant by that:

- *Fixed schedule*: for the most part the scheduling systems expect a set of trains to be predefined that carry a set of well defined blocks. This approach works well for the general merchandise or carload business segment, but may not work as well for other segments such as unit train operations or the movement of grain. This shortcoming can be addressed in part by more actively revising the train schedules over the near-term time horizon, and possibly through the addition of specialized logic for the business segments that do not fit well with the traditional car scheduling paradigm.
- *Location centric*: when you travel by air, you generally look at a number of possible itineraries, and then select or build one that meets your needs. This means that your routing plan is not "owned" by the airports you use, but by you as an individual. When a railcar is moved, it does not own its own routing plan. Instead, at each location the shipment visits, tables and other systems are examined, and based on the content of these tables, the next location for the shipment is determined. Thus, the routing plan is "location centric" and not "shipment centric."

This had significant advantages in an environment with limited communications, and no fully defined, centralized, computerized operating plan. Each location could have a "blocking book" or "routing guide" and know what to do with each shipment without having to consult with a central authority. Even today, this approach has advantages when shipments are misrouted, or fail to connect to their expected train, because it supports a straight forward way to determine what to do with the shipments.

- *Uncapacitated*: most car scheduling systems assume that the next train will always have sufficient capacity to take the shipment, making these systems "uncapacitated." This permits the generating of a car schedule independent of all other shipments, which greatly simplifies the process. However, it does mean that trip plan adherence failures will occur due capacity constraints. It is often argued that this is acceptable for two reasons:

  - First, due to the unpredictable nature of when specific railcars will arrive at a particular yard, and the need to have operational flexibility at each yard to manage it in the most efficient manner, one cannot predetermine which shipments should use a particular train when it is at capacity.
  - Second, by showing the expected volumes on a train, even when it is above capacity, the yard managers will have a clear understanding where they may have an issue, and thus can (a) make informed decisions on which shipments should be delayed or rerouted, (b) elect to take actions to increase the capacity of the train (e.g., add a locomotive), or (c) operate an "extra" train. Further, capacity is often viewed as a "soft" constraint that can be changed in many cases.

Given the above discussion, a number of alternatives are possible for how a car scheduling system could function including:

- *Shipment centric routing*: instead of scheduling based on a myopic, current location basis, take a broader network view of the scheduling process, and then tie the resulting routing to the shipment. When a shipment is processed at a yard, it is assigned to a train not based on local routing instructions, but based on the routing instructions owned by the shipment. A fallback solution will still be required when a shipment falls off its planned routing. This has a number of advantages, including the ability to support reservation type systems, customize routings for individual shipments/customers, and provide a foundation for supporting a capacitated routing process.
- *Capacitated routing*: the core concept is to track the current volumes of shipments assigned to each train, and take these volumes into account when shipments are routed. Under the simplest scheme, if a shipment cannot fit on a train it is "rolled" or delayed to the next train with available capacity. Under this scenario the overall routing remains fixed. Other options involve changing the routing of selected shipments to avoid capacity bottlenecks based on business logic that trades-off the costs of delays, the costs of extra handlings or distance, and the overall transit times of the various options. In some of these solutions each

shipment receives its own unique routing that may deviate significantly from other similar shipments. Concerns have been raised about this approach related to both introducing more variability in transit time for the customers and making the operations less predictable (which may cause field personnel to make more errors). Important in any process of this type is how priorities are set between shipments, which will determine which shipments are delayed. Other factors include whether a previously scheduled shipment can be "bumped" off a train in favor of a higher priority shipment, and if multi-railcar shipments can be split apart.

- *Specialized trip planning logic*: there are many specialized types of services that do not fit the conventional carload model, and need special handling in the scheduling system. A few are described below:
  - *Unit trains*: unit trains are generally groups of railcars (shipments) that move intact from a single shipper to a single receiver on a single train. There can be regular unit train services that operate like a "conveyor" between two points, and more ad hoc unit trains that have variable origins and destinations. By their nature, car scheduling systems are often poorly aligned with such shipments, particularly because the trains do not operate on a fixed schedule. Special logic is required to both define the blocks for these shipments, and ensure that dynamically scheduled trains exist to match each movement. At present, most car scheduling systems do not address the scheduling of unit trains.
  - *Grain*: there are three primary types of grain shipments: small lot, medium lot, and large lot. For the most part, large lot shipments move in unit trains, generally scheduled on an ad hoc basis. Many of these unit trains are operated as "shuttle trains" where the customer buys the train for some period of time and is responsible for its scheduling. Small lot shipments generally move in the carload network, and are subject to conventional scheduling rules. Medium lot shipments pose the largest challenge. In some cases these shipments move in the carload network, and in other cases two or more medium sized lots are combined into a "solid train" to a single destination. It can be very challenging to know when a grain shipment falls into one of these three categories, and as a result, such shipments can cause inappropriate volumes to be reported on the carload train network by the car scheduling system. Best practice appears to be the use of "hold" blocks, with manual determination of how the shipments will be handled from the hold location. The idea is for the carload network to drive these shipments to a staging point, where they are put into a hold status. A person then reviews the available shipment lots, and either dispatches the shipments in a solid train, or "releases" them into the carload network at which point they are scheduled using conventional trip planning logic. These shipments also pose significant problems in capacity management because the large lot sizes represent "lumpy" volumes that disrupt the statistical stability of expected train sizes within the carload environment.

  - *Intermodal*: for the most part it is the belief of the authors that intermodal traffic (trailers and containers) and operating plans are becoming increasingly similar to a specialized carload network, and that conventional car scheduling strategies can be applied. The main complication is that the scheduling of the railcar can be different from the scheduling of the intermodal unit. For example, intermodal units can be "switched" between railcars en-route by lifting the intermodal unit from one car to another, or by the use of highways to make some connections, particularly between railroads. Thus, intermodal units must be scheduled separately from the railcars. Intermodal scheduling remains an evolving problem, and no single, standardized approach is used within the industry.
  - *Hold blocks*: one important concept is the hold block. As cited in the section on grain above, hold blocks are used to classify or sort a set of shipments into a group that does not have an outbound train. Essentially they can be viewed as a forced trip plan failure. These hold blocks are used for a variety of purposes, but the most common is to collect shipments that will require manual intervention prior to onward movement. Grain is one example as described above. Two other common North American examples are empty autorack movements, and empty coal and grain cars. Both cases are similar. For North American autoracks, a central group determines where these railcars are needed next. Each railroad gathers them at staging yards into hold blocks, and then awaits instructions as to where to send them. For coal cars returning to mines for loading, and grain cars going to elevators, the specific mine or elevator may not be known at the time the cars are emptied. As a result, the cars are sent to a staging yard, and put into a hold status awaiting further instructions. Even if these cars are assigned to a specific final destination, the fungible nature of this car fleet may dictate that hold blocks be used allowing local operational convenience to determine the exact cars sent to a specific mine or elevator. Hold blocks are sometimes also used as a backup safety mechanism to prevent the movement of oversized cars or cars containing hazardous materials on restricted routes (such as through tunnels).

- *Time-based routing logic*: as discussed above, as one starts to consider capacity issues, the need to support routing variations among similar shipments arises. This includes potentially "rolling" of traffic to later trains, and changes in routings to take advantage of routes with available capacity. In some cases, the operating plan itself may also change by day-of-week or time-of-day, and these changes will need to be accounted for. A number of strategies exist to support these concepts. Three examples include:

  - *Last train out*: one North American Class I railroad seeks to identify the latest possible train one can take and still meet customer service commitments. This is then used in the capacity management process, where local terminal managers know that they can put shipments on this last train, or any train with the right routing that leaves earlier than the identified train. A simple time-space network of the available routing options is used, in combination with a shortest path type algorithm to identify the last possible train (the ending time is fixed by the customer delivery commitment). See Sect. 4.10 for an explanation of the time-space network.

  - *Fastest block routing*: given a fixed plan of blocks, algorithms can be used to determine the lowest cost routing for each shipment over the available blocks (see Chap. 5 for a deeper discussion on this topic). An alternative is to also introduce a time-based factor into this algorithm, reflecting the available trains to move the blocks. In some cases this may reveal that a different routing for some shipments may produce better results in terms of total transit time, with no appreciable change in the other cost factors. Routings for shipments would still be fixed, but time will thus be added to the decision variables. This approach is being used by one North American Class I railroad, and leverages a time-space variant of the existing shortest path algorithms described in Chap. 5. The solution remains uncapacitated.
  - *Full time-space network*: the concept behind this approach is described in more detail later in this chapter, and is based in part on a thesis by Edwin Kraft (1998). The basic idea is to create a time-space network of train-blocks that have attributes inherited from the underlying blocking plan, and that applies algorithmic type routing rules to the selection of train-block sequences to move shipments to their destinations. Such an approach has the potential to optimize for both capacity and the trade-off between time and other cost factors. No existing production solutions are known to exist that use this approach.

- *Tonnage-based scheduling*: the idea behind “tonnage-based scheduling” is to only operate trains when they are full. Under this approach, the blocking plan is fixed, and the skeleton train plan is typically fixed, but the timing of trains and their frequency is a variable based on volume. In this case, each yard makes a fixed number of blocks. When a block collects enough cars for an entire train, a train is formed with either a single block, or group of blocks, and is sent to the train’s destination. These trains are unscheduled—they operate only when there are enough cars. While it is possible to predict the block sequence, it is impossible to generate a trip plan in advance since the train departure time is not fixed in advance. While this approach appears to minimize the number of trains operated and associated costs, it also sacrifices shipment reliability due to the inability to provide complete trip plans, and operational costs due to increased handlings and the inability to plan sufficiently far in advance to properly plan locomotives, crews and other assets. We should note that the North American railways were closer to a tonnage-based scheduling system in the 1970s through the early 1990s, and many surveys revealed that shipment reliability was the primary factor in potential customers using trucks instead of the railways. The growth of the railways in the past decade and a half was in part due to increased reliability of shipments due to better planning and keeping with a fixed schedule. Railroads also report significant improvements in asset utilization from operating on a fixed schedule basis, instead of a tonnage basis, due to both better planning and higher asset velocity.

4.8 Capacitation and Reservations

Capacity management and the taking of customer reservations are currently done on only a very limited basis within the freight railroad industry. To some extent, one can argue that the management of unit trains represents a form of capacity management and reservations. For individual or small lot shipments, the most widespread case of reservations or guaranteed train assignments are for intermodal shipments. For the case of more general carload traffic, the authors are aware of only one fully functioning reservation and capacity management system, and that is the one used by Green Cargo in Sweden (Green Cargo; Jeppsson 2010). Much of the discussion in this section will be based on the authors' direct experience with the Green Cargo system, and will be focused on the general movement of carload traffic.

The management of capacities and the prediction of future train sizes and yard workloads have many advantages including:

- Ensuring that trains do not become overloaded, and that shipments do not become unexpectedly delayed.
- Identifying where actions should be taken to avoid congestion issues through the delaying or rerouting of selected shipments or the addition of extra capacity.
- Allowing the dynamic right-sizing of capacity to reflect the anticipated levels of demand, thus supporting tighter cost control.
- Supporting dynamic pricing (revenue management) to level peak demand and maximize total revenue.
- Allowing specific shipments to be protected in their movements by ensuring that the necessary capacity is available to handle them.
- Improving the management of locomotives, cars, and human resources to ensure that the right capacity is available at the right times in the right places.

However, relative to current practice, implementing a capacity management system has a number of challenges:

- At present most railroads have significant deviations from plan in their day-to-day operations, which means that capacity management at the level of prescribing specific itineraries for shipments will tend to experience a high failure rate.
- The deviations from plan would likely cause the volume projections to be unreliable.
- Projections of capacity utilization for trains and yards must be based on the joint processing of all shipments active on the network, which means that the current practice of rescheduling individual shipments independently from each other would not be viable.
- Proper capacity management must take into account traffic that is not yet moving on the railroad's network. However, under current practice, shippers are not required to provide advance notice on when shipments will be released for movement or the destinations to which those shipments will be directed. Further, a large percentage of traffic is received from other railroads, which may not provide accurate or complete information on this traffic in advance of its interchange to the receiving railroad.
- There are many operational challenges to following preplanned itineraries for each shipment that take capacity restrictions into account, particularly if this

means (a) having to hold some shipments at the origin or intermediate yards that may not be well suited to storing selected cars, and (b) individual shipments receiving different routings due to capacity limits, meaning that there is no longer a standard way to handle all shipments going to a particular destination.

Thus, in the authors' view, full capacity management is only possible when two conditions are achieved: (1) most traffic is booked in advance by customers, and (2) the railroad operates with a high degree of adherence to the plan. In the case of Green Cargo, customers must book all of their shipments several days in advance, including specifying the release date and time, and the overall railway achieves 95 % or better in its adherence to the trip plans promised to the customers. In part, customer adherence to the booking process is achieved because only by booking a shipment will an empty be provided for loading as the empty supply is integrated with the booking process.

4.8.1 Specifying Capacities

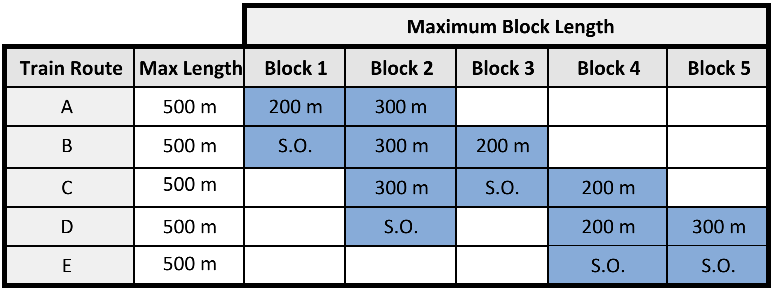
The potential capacities that could be managed include limitations on the number of railcars that can be handled at a yard, and limits on the sizes of trains. In looking at the Green Cargo example, the system only focuses on train capacities, and not yard capacities. Yard capacities are taken into account during the design process as the blocking plan and train plan are developed. Because shipments are only accepted if there is sufficient train capacity to move them to destination, and because the plan is largely fixed, yard workloads generally stay within acceptable limits as a direct result of using the reservation system and ensuring a high degree of plan adherence. Train size limits are typically both on the length and weight of the train; however some examples below only specify length capacity constraints to simplify the exposition.

From a train perspective, there are two types of capacities: overall train size limits and the allocation of capacities to individual blocks on the train. In general, the overall train size limits are treated as absolute limits that cannot be violated, and are specified in terms of a maximum weight and length. These limits typically have the option of varying along the train route due to changes in siding lengths, locomotive capacity, and other factors.

The more complex issue is the allocation of capacities to specific blocks on a train. For discussion purposes consider the following train and its block assignments (Fig. 4.10):

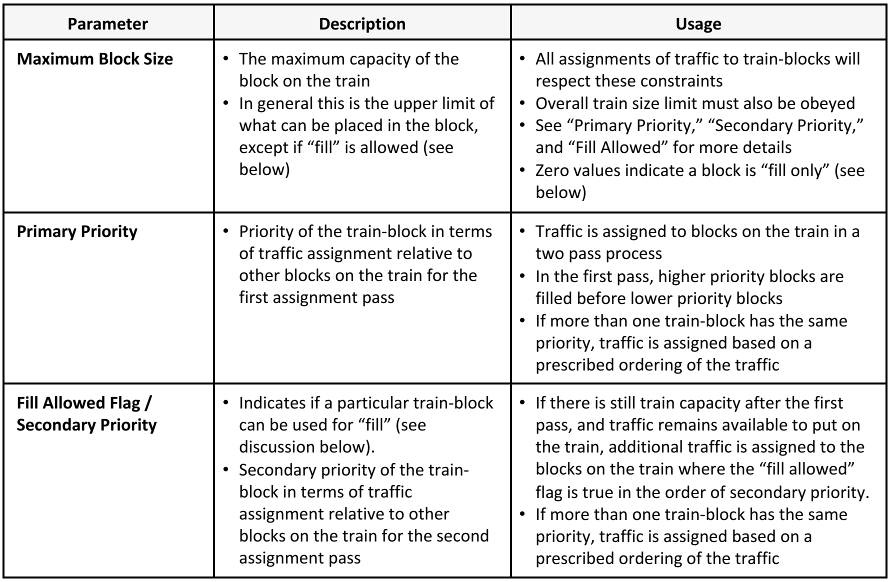
In the above table, the train goes from A to E, via intermediate points B, C, and D. Overall, the train has a maximum capacity of 500 m, which as discussed above is an absolute limit that cannot be exceeded. The train carries five blocks, which are picked-up at the first shaded location, and set-off at the location marked "S.O."

In this sample case, the capacity of each block is shown as a subset of the total train capacity, and the sum of the capacities allocated to each block never exceeds the total capacity of the train. However, what happens for example if there is more



		Maximum Block Length				
Train Route	**Max Length**	**Block 1**	**Block 2**	**Block 3**	**Block 4**	**Block 5**
A	500 m	200 m	300 m			
B	500 m	S.O.	300 m	200 m		
C	500 m		300 m	S.O.	200 m	
D	500 m		S.O.		200 m	300 m
E	500 m				S.O.	S.O.

Fig. 4.10 Setting train-block capacities as a subset of the overall train capacity



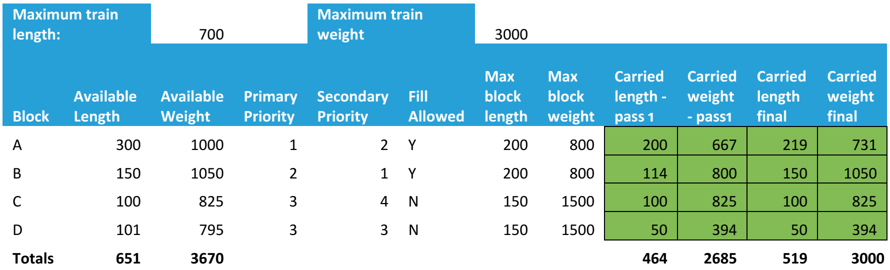
Parameter	Description	Usage
Maximum Block Size	• The maximum capacity of the block on the train • In general this is the upper limit of what can be placed in the block, except if "fill" is allowed (see below)	• All assignments of traffic to train-blocks will respect these constraints • Overall train size limit must also be obeyed • See "Primary Priority," "Secondary Priority," and "Fill Allowed" for more details • Zero values indicate a block is "fill only" (see below)
Primary Priority	• Priority of the train-block in terms of traffic assignment relative to other blocks on the train for the first assignment pass	• Traffic is assigned to blocks on the train in a two pass process • In the first pass, higher priority blocks are filled before lower priority blocks • If more than one train-block has the same priority, traffic is assigned based on a prescribed ordering of the traffic
Fill Allowed Flag / Secondary Priority	• Indicates if a particular train-block can be used for "fill" (see discussion below). • Secondary priority of the train-block in terms of traffic assignment relative to other blocks on the train for the second assignment pass	• If there is still train capacity after the first pass, and traffic remains available to put on the train, additional traffic is assigned to the blocks on the train where the "fill allowed" flag is true in the order of secondary priority. • If more than one train-block has the same priority, traffic is assigned based on a prescribed ordering of the traffic

Fig. 4.11 Parameters for describing train-block capacity with priorities and fill train-blocks

traffic for Block 2 than 300 m, and Block 1 has less than 200 m of traffic on a particular day? Should we not be able to shift capacity between these two blocks? However, if we let Block 2 exceed 300 m, we might then have a problem if Block 3 needs its full allocation of space.

To address these types of issues, the following capacity specification concepts have been implemented in the planning software used by Green Cargo (Fig. 4.11):

The idea of a fill block is to designate that traffic can be placed in a train-block only if there is space available after other higher priority traffic has been processed and placed in the train. This can be used to trade-off the use of space between two different blocks, or to ensure that a train is filled only if the "regular" blocks do not have enough traffic on a particular day. Section 4.4 also discusses fill blocks.



Maximum train length:	700	Maximum train weight	3000

Block	Available Length	Available Weight	Primary Priority	Secondary Priority	Fill Allowed	Max block length	Max block weight	Carried length - pass 1	Carried weight - pass1	Carried length final	Carried weight final
A	300	1000	1	2	Y	200	800	200	667	219	731
B	150	1050	2	1	Y	200	800	114	800	150	1050
C	100	825	3	4	N	150	1500	100	825	100	825
D	101	795	3	3	N	150	1500	50	394	50	394
Totals	**651**	**3670**						**464**	**2685**	**519**	**3000**

Fig. 4.12 Illustration of the use of train-block capacities with prioritization and fill train-blocks

Refer to the next table for an example that applies capacitation logic to a train that picks-up four blocks at the train origin and sets all of them out at the train destination. There are absolute limits on the train length of 700 m and train weight of 3,000 t. All four blocks combined have 651 m and 3,670 t of cars. The table below gives the primary and secondary priority of the train-blocks and whether fill is allowed. The last four columns are discussed below (Fig. 4.12).

A two-pass approach for calculating how much of the available traffic will be accepted for each train-block is used. The first pass begins with block A since it has the highest priority. The max block length of 200 m is the limiting constraint, so 200/300 = 67 % of the available cars are accepted, leading to 200 m length and 667 t allowed to be carried. The next block in priority order is B, and block weight is the limiting factor, resulting in 800/1,050 = 76 % of the available cars allowed to be carried, or 114 m and 800 t accepted. The carried length and weight so far is 314 m and 1,467 t, less than the absolute limits of train length and weight.

Blocks C and D have the same primary priority. In this case, it means that the combined limits on C and D are 150 m and 1,500 t (the planning process requires that if two blocks have the same primary priority they must have the same maximum length and weight). There are several ways to calculate how much traffic could be carried in this first pass between these two blocks, depending on how traffic is prioritized among these two blocks. In our example, we will arbitrarily prioritize the block with the smallest block name, in this case block C. We see that all of the available traffic of block C can fit within the combined limits. This leaves available for block D a capacity of 50 m and 675 t. Block D has more than 50 m and more than 675 t, with the most limiting factor being the length, so 50 m and (50/101) × 795 t = 394 t are allowed to be carried in block D.

So far cars totaling 464 m and 2,685 t have been allocated to the train, which means there is still room for 236 m or 315 t to be added to the train, whichever comes first. Since both the length and weight limits of the train have not been reached, the blocks designated as fill blocks are examined to see if more cars can be allocated to the train in the second and final pass of the process.

Block B has the highest secondary priority that is allowed to have fill. So far only 76 % of the available cars for block B have been allocated, and we find that we can add the remaining cars and still fit within the train capacity. Doing so drops the

		Maximum Block Length				
Train Route	**Max Length**	**Block 1**	**Block 2**	**Block 3**	**Block 4**	**Block 5**
A	500 m	**500 m**	300 m			
B	500 m	S.O.	300 m	**500 m**		
C	500 m		300 m	S.O.	200 m	
D	500 m		S.O.		200 m	300 m
E	500 m				S.O.	S.O.

Fig. 4.13 Example allowing train-block capacities to exceed overall train capacity

available room on the train to 200 m and 65 t. We then examine the next secondary priority fill block, A. It has 100 m and 333 t still unallocated. The tonnage is the limiting factor, and we can add in 65 more tons and (65/333) × 100 m = 20 m (it appears as 64 t and 19 m in the table above due to round-off). At this point the maximum weight of the train is reached, and the process stops.

There are many variations of the above approach, including multi-pass strategies for allocating traffic, and the potential for the sum of the train-block capacities to exceed the maximum train size. For example, consider the following revision to the train-block size limits (Fig. 4.13):

In the above example, Block 2 cannot exceed 300 m because we must protect space on the train for Block 3 when it is picked-up at location B. However, Block 1 can be larger, provided that the train size limit is not exceeded. Likewise, once the size of Block 2 is known, Block 3 can be larger, again subject to the overall train size limit. In this example, we would likely make Block 2 be of a higher priority than Blocks 1 or 3, but not allow fill for Block 2.

One of the key questions for the car scheduling or reservation system is whether the assignment process takes a greedy/myopic approach, or if a more holistic approach is taken. By myopic, we mean that the traffic assignment process only looks at the rules and traffic available to be assigned to a train at the current location, and does not consider the volumes that may be assigned at downstream route locations. For example, if we knew that there would only be 100 m of traffic for both train-blocks 3 and 4, we could increase the size of train-block 2 to 400 m. Without this knowledge, we would need to keep train-block 2 at 300 m to protect these downstream train-blocks. Current practice appears to favor a myopic approach, both to keep the processes simpler, and because the variability in operations and traffic demand is sufficient to make the prediction of the volumes for a downstream route location difficult to achieve on a reliable basis.

In the context of Green Cargo, the most common situation is to have one primary train-block on a train, and one or more secondary train-blocks. The primary block is considered an "anchor block" and is often customer-specific. Based on commercial negotiations, a specific amount of space is allocated to this anchor block, and it may form the primary commercial justification for the train. Reservations are then taken for the train, protecting the anchor block. As one approaches the departure date and time for the train, depending on the utilization of the anchor block, the anchor block's capacity may be released so that the remaining blocks on the train can be "filled" with additional traffic, or the anchor block may be allowed to take on "fill" traffic itself if customer demand is high.

The last major decision variable in the management of the capacities is to determine which traffic should be placed on a specific train at the time the train departs. In the absence of a reservation system, a prioritization of the traffic can be used. For example, sorting the traffic by arrival time in the yard, or the time that it is classified into an outbound block, may fit well with actual operations. In actual practice, and in simulation modeling, other criteria can be used such as a booking date/time recorded on the shipment records, or the release date/time for each shipment. The relative priority of the individual train-blocks must also be taken into account.

In a reserved system, the traffic that is pre-booked to take a specific train generally takes priority over other available traffic. At Green Cargo this process is tightly managed, and the plans for each train departure are reviewed and adjusted based on the booking information and the actual status of individual shipments against their schedules. In addition, as traffic is received from interchanges with other railroads, this traffic must also be entered into the booking system to ensure that it has space allocated to it on the trains.

4.8.2 Managing Reservations

To support a reservation-based approach, the car scheduling system must be modified to maintain a contindatabase of the expected volumes assigned to each date-specific train and train-block that will be operated. Furthermore, the trip plans must be generated in a manner that is consistent with the management of the available capacities. To support this there must also be a process of defining the trains to be operated and the associated blocking rules starting at the point in time at which reservations can be accepted. While most existing trip planning systems only maintain a 7- to 14-day forward view of date-specific trains to be operated, the use of a reservation system requires this to be expanded to include the trains that are expected to be operated for several weeks in advance of actual operations. For example, if customers can book shipments up to 2 or 3 weeks in advance, then perhaps 21–28 days of date-specific trains must be maintained at all times.

The reservation system must respond promptly as customer shipment requests are received. Thus, by its nature the process will be greedy, and will not consider potential future (unknown) business that might be received except through the management of the capacities on individual train-blocks or the potential use of dummy reservations to protect identified business opportunities.

The reservation process must be able to find a viable solution in terms of a trip plan or itinerary for each proposed shipment. In the case of Green Cargo, the routing of traffic is fixed, and shipments are simply delayed at the origin or intermediate points to later trains when earlier trains are full. This approach allows the blocking plan to remain static, but does present the challenge of potentially having to hold traffic at intermediate yards.

The Green Cargo system books both the inbound empty wagon and the outbound loaded movement at the same time. Only if both are feasible against the customer's proposed release date/time will the reservation be accepted and the empty delivered to the customer.

An alternative approach is to use dynamic routing as part of the reservation process, allowing for the blocking of traffic to change based on the capacities available. Such an approach is explored in Sect. 4.10.1 below.

4.9 Planning and Optimization

Car scheduling systems are meant for real-time use. They have little ability to analyze whether a plan is well designed. At best, existing car scheduling systems seek to drive plan adherence, and check for data gaps (such as unassigned blocks or invalid station codes). They generally do not support the evaluation of real-time plan changes, or seek out plan deviations that would improve performance. They instead rely on the underlying plan to be of good quality, and thus do not provide many features found in good planning systems.

The authors have been involved in creating MultiRail®—a planning system used by many railways in North America, as well railways in Europe, Asia and Africa. Almost all of the features and capabilities discussed below are in MultiRail, but the authors have also worked with, or have significant knowledge of, other planning systems, and the below discussion is somewhat of a union of the various planning systems they have encountered.

Major features of planning systems include:

- Ability to test if the car scheduling rules can provide trip plans for all possible traffic, or at least for all likely traffic
- Complex validations of the car scheduling rules, such as testing to see if block swaps are complete
- Estimation of block and train sizes
- Alerts if projected yard workloads exceed thresholds, or projected train sizes or train-block sizes exceed capacity limits
- Various diagnostic reports such as excess circuity, long trip plan transit times, shipments with many handlings, etc.
- Various key performance indicators such as estimated gross-ton-miles

More advanced planning systems may also offer:

- An enhanced ability to edit the various components of the car scheduling rules, such as blocking rules and train routes. A typical enhancement found in some planning systems, but never in the car scheduling systems, is the ability to see estimated block and train sizes change as edits are made on the car scheduling rules. Another enhancement is to see the set of traffic that uses a particular block. The planning system usually has a richer user interface compared to the (usually) mainframe "green screen" interface used for most car scheduling systems, making it more desirable to use the planning system for editing the mainframe rules.

- Traffic forecast processing—the ability to take a high-level, usually "loads only" forecast and transform it to a level of detail that yields valid train-size and block-size estimates. A big part of this is it includes estimating empty traffic movements.
- Blocking plan improvement diagnostics, such as bypass opportunities as discussed in Chap. 5.
- Train plan improvement diagnostics, such as suggesting which train-blocks should be carried on a particular train, as discussed in Chap. 1 on train scheduling.
- Blocking optimization, as discussed in Chap. 5 on blocking.
- Locomotive, crew and other asset estimation; as discussed in Chaps. 2, 6, 8 on crews, locomotives, and network simulation.

Planning systems typically have five major components:

- *Traffic file management*: this is the specification of a historic or forecasted set of traffic movements for use in the planning process to estimate volumes, generate trip plans, and test for plan completeness.
- *Blocking and train plan rule editor*: an editor for managing blocks (including the blocking rules) and trains (including the routes, block-to-train assignment and other rules). This editor is usually enhanced as described above.
- *Network viewing*: the ability to see the rail network on a map, as well as to see train routes and various other information about the plan graphically.
- *Trip planning*: the planning system must be able to mimic the car scheduling system business logic for producing trip plans. Due to the legacy of some of the existing car scheduling systems with specialized business rules built up over a long time, and often in a computer language that is not well supported, the ability to duplicate the car scheduling logic is often difficult.
- *Network simulation*: this is the ability to obtain trip plans for all the traffic records and from it produce analyses on block and train sizes and yard workloads, as well as shipment durations, dwell times and so on. This is discussed in Chap. 8 on simulation.

Central to the planning system is the concept of a traffic file. This file represents shipments that are used for testing the plan and estimating block and train size and other key statistics. The traffic file is most often a compressed version of history. For example, if the railroad is planning the "winter" schedule, they may use the previous year's history from November 30 to December 20 (3 weeks) plus January 4 to February 28 (8 weeks). For a large railway, this might involve two million records. Usually these records are compressed by grouping on most of the traffic/blocking attributes, and “sampling” selected other blocking attributes. For the compression, important attributes such as origin, destination, car type, commodity, shipper and client are grouped, but some little used blocking attributes that tend to be very car-specific, such as car initial/number, are “sampled” instead of grouped to preserve a reasonable compression effect.

Using a high-quality traffic file is critical in a planning system. Creation and maintenance of the traffic file, since it will most likely change every month, is often one of the most time-consuming activities in the maintenance of the planning system.

4.10 Time-Space Network Solutions

Car scheduling is typically built around the concept of “first legal train-block out.” Often, this strategy will get the car to final destination as quickly and as efficiently as possible. But there is no guarantee that this is always the most efficient routing strategy. For example, the “next train-block out” may result in a routing with extra block swaps. Or the train route may be more circuitous. It is possible that a train that departs at a later time is a better choice. This is illustrated in the below time–distance diagram, where a car is released at 10:00. The dashed line represents a trip plan based on taking the next-train-block out that arrives at 42:00. The solid line represents a different trip plan whose first train pick-up occurs later, but has fewer block swaps and ultimately arrives earlier at 37:30 (Fig. 4.14).

One North American railway has implemented a *k*-shortest path algorithm to find a variety of trip plans—all using the same block sequence—and then evaluating each trip plan according to business goals that are specific to the type of shipment. For example, for intermodal all of the options that will satisfy the commercial delivery time commitment for a shipment can be identified, and the option that best

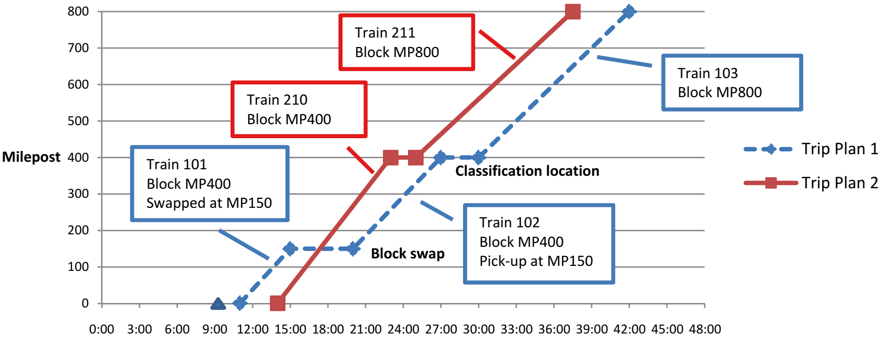


Fig. 4.14 Time-space illustration of two potential trip plans where the first-outbound train delivers the railcar later than the second-outbound train

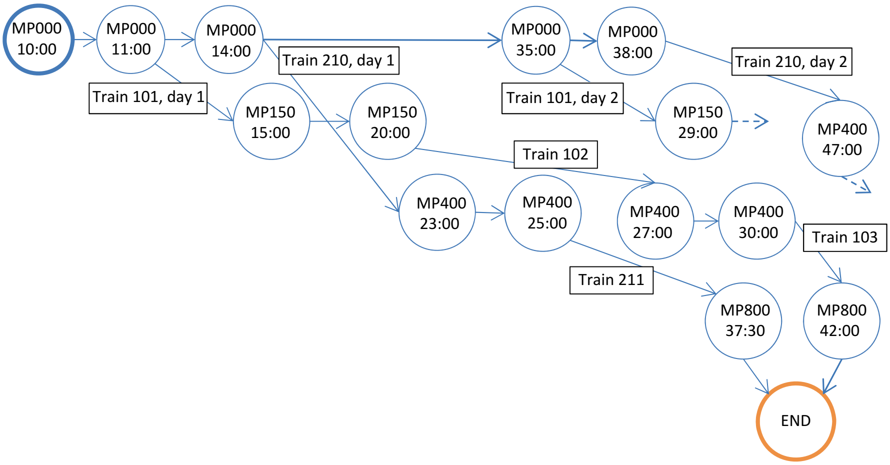


Fig. 4.15 Time-space network that is used in shortest-path calculation for example from Fig. 4.14

fits with the local operations selected. For intermodal, the goal might be earliest arrival at destination, while merchandise the goal might be to minimize block swaps or circuity. By looking at a variety of trip plans, and not just finding a single feasible path, the railway is able to use complex business rules to evaluate the possibilities.

The basic idea is to find all train-blocks over a multi-day period that carry the yard-blocks from the traffic's block sequence. Each train-block represents an arc in a time-space network: the tail of the arc is the pick-up yard and time-of-departure, and the head is the set-out yard and time-of-arrival. There are also dwell arcs within a yard. For the example above, the time-space network would appear as follows (Fig. 4.15):

Note that several instances of train 101 and train 210 are shown in the time-space network to depict that the solution may find it better to wait for a train on a subsequent day, and that each arc represents a single instance of a specific train symbol operating on a specific day. Improved service by taking a later instance of the same train could happen if downstream trains run less than 7 days a week, and it is decided that it is better to keep the car at origin than to have it dwell at some intermediate point while waiting for a train to depart.

There are generally two variants of using the train-block/time-space network. The first variant only allows use of a single block sequence for a shipment. This is the case in the example just discussed. The second variant allows the block sequence to change depending on a variety of factors such as changes in the service offered by day-of-week or time or day, or due to capacitation issues for selected trains, and is explored below along with a more formal description of the train-block/time-space network formulation.

4.10.1 Dynamic Car Scheduling

Kraft (1998) proposed a method for shipment-centric, capacitated car scheduling based on two main ideas:

- A reservation/revenue management system to determine if a shipment should be accepted by examining (a) the potential train-block routes of the shipment, (b) the revenue associated with the shipment, and (c) a forecast of all the demand for those train-blocks segregated into revenue buckets. Kraft specifically proposed developing "bid prices" on the train-blocks which provide an easy mechanism for quickly determining whether to accept or reject a shipment. See also Armstrong and Meissner 2010 for a discussion of Kraft's work.
- A real-time, dynamic car scheduling system that understands train-block capacities, the currently accepted orders, and the cars online. The scheduling is based on finding train-block sequences directly over a time-space representation of the operating plan for each shipment. Individual block sequences for shipments are not predetermined, and can be changed based on the available capacities and relative priorities of each shipment.

In a highly scheduled, precision run operation, there is less need for dynamic car scheduling. This is because the reservation system pre-routes all the shipments and only allows enough shipments onto the network to keep within capacity limits, and the composition of cars on the trains can be exactly forecasted. However, it is difficult to achieve that level of precision in operations. The authors are aware of only one fully capacity reserved, fully scheduled freight railway operating anywhere in the Americas, Europe or Africa, and that is Green Cargo in Sweden. Green Cargo does not use a dynamic scheduling approach, and instead used fixed block sequences, and delays shipments to later trains when accepting reservations that would exceed the capacity of the "default" trip plan.

While Green Cargo can achieve a reserved operation through extraordinary operating plan adherence, this is an exception that has not been replicated elsewhere. Recognizing the difficulty of achieving complete plan adherence, Kraft proposes a dynamic car scheduling system that allows cars to be rerouted or delayed to keep the system as close to schedule as possible in an economic fashion while respecting the capacities of the available trains and train-blocks.

Kraft also argues that dynamic car scheduling should be used to increase revenue by providing a mechanism to accept last minute high-revenue, high-priority traffic while "bumping" lower priority traffic that might already be moving or about to move.

The dynamic car scheduling system uses algorithmic routing that is extended to the time-space network of train-blocks, as described earlier in this section. Two important notes:

1. The car scheduling algorithm simultaneously examines all the cars that need to be routed and knows of the capacities of the trains and train-blocks. Some cars will take the most economical, quickest route. But by examining all the cars simultaneously, the algorithm may decide to have some cars take longer routes by either time or distance (or both) to take better advantage of capacities on other trains.

2. In addition to the time-space network of train-blocks, there is a special node called a "super sink." Arcs are connected from each train-block set-out node to the super sink when the location of the set-out is the shipment destination. The costs on these links could represent penalties for lateness of the shipment in cases where the set-out time is after the targeted arrival time.

More specifically, Kraft proposes a "multi-commodity-network-flow" arc-node formulation with the commodities being traffic records. It is important to note that this approach has the potential to change the solution for any individual shipment each time the network is reoptimized due to introduction of new shipments that may cause the relative importance and best route choices for existing shipments to change.

The following description of the network is for a single traffic record $t \in T$. One can view the problem as if each traffic record has its own set of nodes and arcs. In reality, there is a superset of nodes and arcs that represent all train-block movements, of which a subset are candidates that can be potentially used by a specific traffic record. In effect, there are parallel networks that will be linked later when considering the capacities.

For each traffic record t, we must generate a set of time-space nodes that represent the pick-up and set-off locations and times of the train-blocks, and arcs representing the train-blocks, dwell arcs, and arcs to a "sink." The set of time-space nodes for each traffic record t, is:

$$N = \left\{\left(p, dt\left(\tau, p\right)\right) \mid \left(\tau, b, p, s\right) \in \Gamma_t\right\} \cup \left\{\left(s, at\left(\tau, s\right)\right) \mid \left(\tau, b, p, s\right) \in \Gamma_t\right\} \cup \left\{\left(\sigma, \infty\right)\right\}$$

where Γ_t is the set of legal train-blocks for the traffic t, and each train-block has four components (τ, b, p, s) with τ representing the train, b representing the block, p is the pick-up location, $dt(\tau, p)$ is the departure time of train τ at the pick-up yard p, and s is the set-out location that occurs at time $at(\tau, s)$ The symbol σ is the sink, and to keep in the same time-space notation it is written as having a time at infinity. The sink node $\left(\sigma, \infty\right)$ will be common for all traffic records.

The time-space node that represents the earliest possible time that traffic t could be first carried from its origin $o(t)$ is the time-space node $\left(o\left(t\right), \hat{r}\right)$ with $\hat{r} = \min\left\{r \mid r \geq R_t \text{ and} \left(o\left(t\right), r\right) \in N\right\}$, where R_t is the release time of t.

There are three sets of arcs in this model:

- Train-block arcs that represent cars moving on train-blocks. Each legal train-block (τ, b, p, s) is an arc from $(p, dt(\tau, p))$ to $(s, at(\tau, s))$. By legal, we mean that the traffic t is permitted to be on the train-block. The cost of using this arc would generally be a function of the switching cost and the distance, and should be non-negative. While the cost could also be a function of the time spent on the train-block, it generally is not considered here but rather considered in the arcs that go into the super sink. Note that it is possible to differentiate between block swaps and classifications through the switching cost formulation.
  - The amount of cars on a train-block arc will be represented as the variable $A^t_{(y_1, r_1)(y_2, r_2)}$.

- Yard-dwell arcs that represent cars dwelling in a yard either in the process of being switched, or waiting for the next train-block to depart. To generate these arcs, for each yard find the set of times for train-blocks setting out or being picked up, order the times, and then build arcs between consecutive times. More formally, for yard *y*, let the set of event times at yard *y* be $\Lambda = \{dt(\tau, y) \mid (\tau, b, y, s) \in \Gamma\} \cup \{at(\tau, y) \mid (\tau, b, p, y) \in \Gamma\}$ and we order the elements of $\Lambda = \{l_1, l_2, \ldots, l_{n_y}\}$ so that $l_1 \leq l_2 \leq \ldots \leq l_{n_y}$. The yard-dwell arcs are (y, l_{i-1}) to (y, l_i) for all yards *y* and $i = 2, 3, \ldots, n_y$. There is generally no cost for the yard-dwell arcs, but in practice a small positive cost is applied that depends on each yard's type so that cars prefer to dwell in certain yards and not dwell so long in other yards.
  - Yard-dwell arcs are sometimes called inventory arcs—they represent the inventory of cars at the yard at a moment in time. They will be represented as $I^t_{(y,r_1)(y,r_2)}$.
- Arcs from the time-space network to the sink. These arcs are from the destination of the traffic, $d(t)$, to the sink, and symbolize the completion of the trip. More formally, the set of nodes at the destination yard $d(t)$ used for these arcs are $\{(d(t), at(\tau, d(t))) \mid (\tau, b, p, d(t)) \in \Gamma_t \text{ and } at(\tau, d(t)) > R_t\}$.
  - We will represent these arcs as $S^t_{(y,r),(\sigma,\infty)}$, and their value is the number of cars that will go onto that arc.

  The cost associated with these arcs is based on the arrival time at the destination yard, and represents a non-negative penalty to being too early or too late at destination.

The above is a somewhat loose formulation. It is possible when solving the system to avoid directly introducing the sink arcs. Also, additional arcs and/or other features are needed to properly model minimum connection and block swap times (for example, advancing the set-off time for a train-block to the node representing the earliest possible departure from the set-off location after the minimum processing time has elapsed). These details will not be elaborated here.

The variables are the amount of cars that travel on an arc. Let $X^t_{(y_1,r_1),(y_2,r_2)}$ = the volume of cars on an arc from time-space node (y_1, r_1) to either the time-space node (y_2, r_2) or to (σ, ∞) for traffic record *t*. This applies for all three arc types. That is, $\{X^t_{(y_1,r_1),(y_2,r_2)}\} = \{A^t_{(y_1,r_1),(y_2,r_2)}\} \cup \{I^t_{(y,r_1),(y,r_2)}\} \cup \{S^t_{(y_1,r_1),(\sigma,\infty)}\}$.

Most of the constraints are straightforward and come from the network definition:

1. Let the number of cars associated with traffic *t* be $w(t)$. To force that all cars are carried, we ensure that all arcs emanating from the first legal node at the origin of the traffic, $o(t)$, have total volume equaling to $w(t)$. Recalling that $(o(t), \hat{r})$ is the first legal node, we have:

$$\forall t \in T,\ w(t) = \sum\nolimits_{\{(y,r) \| (o(t), \hat{r}),(y,r) \text{ is a valid arc for } t\}} X^t_{(o(t),\hat{r}),(y,r)}$$

2. Flow balance must be achieved at all other nodes

$$\forall t \in T, \forall (y,r) \in N \setminus \{(\sigma,\infty),(o(t),\hat{r})\}, \sum\nolimits_{(\tilde{y},\tilde{r})} X^t_{(\tilde{y},\tilde{r}),(y,r)} = \sum\nolimits_{(\tilde{y},\tilde{r})} X^t_{(y,r),(\tilde{y},\tilde{r})}$$

3. Capacity is constrained—for each train, and for each route point on the train, the volume on the train must be less than its capacity denoted by Cτ. In this explanation, we only have one type of capacity which is the number of cars. However, most railways also use both length and weight as capacity.

$$\begin{aligned}
&\forall \tau,\\
&\forall y \text{ on the train route of } \tau,\\
&\forall \text{train block } (\tau, b, p, s) \text{ so that } y \text{ is on the train route between } p \text{ and } s,\\
&\sum_{t \in T}\sum A^t_{(p,dt(p)),(s,at(s))} \le \mathrm{C}\tau
\end{aligned}$$

If the capacity constraints were eliminated, then finding the optimal solution to the model is simple—starting at the node $(o(t),\hat{r})$ find the shortest path to (σ,∞). Since every arc $X^t_{(y_1,r_1),(y_2,r_2)}$ has the property that it moves forward in time $(r_1 < r_2)$ the underlying network is acyclic, and with all the cost coefficients being non-negative there are very fast acyclic shortest path algorithms that could be employed.

Kraft suggested using a technique called Lagrangian decomposition on the capacity constraints to decouple these constraints from the other two constraints. The general principle is to move the difference between the capacity of the train and the volume of the train to the objective with a penalty cost. The penalty cost is called the Lagrangian multiplier and represented as $\lambda_{\tau,y}$. Using $c^t_{(y_1,r_1),(y_2,r_2)}$ to represent the cost of the arcs, the key to the theory of Lagrangian decomposition is that the solution to:

$$\min_{\text{subject to equations } 1,2,3} \left\{ \sum c^t_{(y_1,r_1),(y_2,r_2)} X^t_{(y_1,r_1),(y_2,r_2)} \right\}$$

is the same as the solution to:

$$\max_{\lambda_{\tau,y}} \min_{\text{subject to equations } 1,2} \left\{ \sum c^t_{(y_1,r_1),(y_2,r_2)} X^t_{(y_1,r_1),(y_2,r_2)} - \lambda_{\tau,y} \left(\mathbb{C}_\tau - \sum_{t \in T}\sum A^t_{(p,dt(p)),(s,at(s))} \right) \right\}$$

The above equation essentially adds a cost—the Lagrangian multipliers—for going over capacity.

If the optimal set of Lagrangian multipliers is $\{\lambda^*_{\tau,y}\}$ and known, then the optimal car flows are the solution to a problem that does not use the capacity constraints:

$$\min_{\text{subject to equations } 1,2} \left\{ \sum c^t_{(y_1,r_1),(y_2,r_2)} X^t_{(y_1,r_1),(y_2,r_2)} - \lambda^*_{\tau,y} \left(\mathbb{C}_\tau - \sum_{t \in T}\sum A^t_{(p,dt(p)),(s,at(s))} \right) \right\}$$

Since the capacity constraints are the only place that different traffic records appear in the same equation, eliminating this constraint allows each traffic record to be optimized separately. Hence this solution can be calculated by performing the shortest path algorithm for each traffic record t but using adjusted costs. Kraft developed a specialized dual adjustment heuristic (Fisher 1981; Keaton 1992) employing a Tabu search (Glover and Laguna 1997; Gorman 1998) for estimating $\{\lambda^*_{\tau,y}\}$. He shows how his technique overcomes the usual problems of slow convergence. At a high level, for each estimate of $\{\lambda^*_{\tau,y}\}$, the algorithm solves the set of shortest paths for each traffic record, then examines which train capacity constraints are violated and re-adjusts the estimates of $\{\lambda^*_{\tau,y}\}$ in such a manner that it leads to provably good solutions. He envisions this car scheduling algorithm to be run in real-time, in which case for each new run the previous estimates of $\{\lambda^*_{\tau,y}\}$ are already close to the true optimal, and hence the algorithm runs very fast.

Several railroads that have looked at this approach have proposed a modified strategy that instead assigns the traffic to the network in a manner that tends to be more FIFO in nature in order to protect the commitments made for existing shipments as new shipments are added to the network. These railways view that when they accept a shipment from a customer, then they should also be making a firm commitment to that customer with respect to the service that will be provided. Under that philosophy, a simple, greedy approach may be the most effective, where a shortest path is used to find the best available routing over the time-space network for each shipment as it is accepted. Once routed, this routing is frozen, and later shipments cannot "bump" the already planned shipments. Such an approach readily lends itself to the use of a successive shortest path type strategy, with arcs being removed from the network as they reach capacity.

All of these capacitated trip planning strategies using dynamic routing are at present strictly theoretical, as no railway is known to the authors to have actually implemented such an approach.

4.11 Opportunities

A variety of opportunities exist to further refine and improve the car scheduling process. A number of these were discussed in this chapter including:

- More predictive, real-time analysis of volumes. The volume of a train expected to depart over the next week is dependent on the cars already in the system and their current trip plans, cars coming from interchange, ship/vessel arrivals (especially for intermodal trains), known empty movements, new cars originating over the next week, and any plans to move additional empties. The current predictive abilities in most railways are limited, and significant research and modeling could be done to give a better predictions of near term car movements.

- Adjustment of train plans: through the use of predictive volumes, there may be optimization opportunities to make real-time adjustments in the plan. While we generally discourage stepping away from a schedule, especially in a short time period, there may be opportunities to do so. We could envision tools to:
  - Evaluate planned train frequencies based on projected train volumes and suggest when trains might be annulled, extras operated, or trains consolidated, or possibly even suggesting when selected traffic might be rerouted to keep the expected volumes within capacity limits.
  - Leverage projected yard volumes, both to plan resources and identify strategies to mitigate situations when volumes will exceed the yard capacity. Such strategies could include rerouting or delaying of selected trains, or rerouting of selected traffic.
- Short-term asset planning: there are almost no systems available that take full advantage of the available trip plan information, as well as the predictive volumes, to tune decisions around locomotive, crew, and yard assets. For example, crew deadhead policies could be more dynamic to reflect the most likely set of trains that will be used, which in turn could be predictively estimated by using train volumes based on current trip plans and business rules.
- Grain train car scheduling: the issues with grain trains have been discussed in this chapter, but developing an optimized grain car scheduling strategy is largely untouched. One effort (Huntley et al. 1995) is not in current use.
- Time-based solutions: Most railways use a fixed blocking plan and FIFO trains for developing trip plans. In Sect. 4.10 we present several approaches that used time-space networks to develop alternative trip plans that retain the block sequence. The current implementations of such solutions are very limited due to complexity.
- Dynamic scheduling that does not fix the block sequence and uses a more shipment-centric approach, such as explained in Sect. 4.10 has not been implemented due to various barriers. It needs further study and far more experimentation before the railways would be convinced of its benefits. Not stated in the section is also the concept of using revenue management techniques for pricing and capacity control; these should be studied and explored further as well.
- Intermodal management: mentioned earlier is that intermodal scheduling has three complications not found in the carload scheduling: (a) the need to trip plan both intermodal units (containers and trailers) and trip plan cars, (b) the ability to switch intermodal units between trains by lifting off the unit and placing it on a different car within the yard, or between yards by using the road network, and (c) the potential need to incorporate intermodal terminal gate constraints and shipping vessels schedules into the trip plans. There is a lack of a uniform, industry accepted trip planning method.
- Unit train management systems support, including the scheduling of train sets with respect to both the loaded and empty movements against the orders that need to be filled.

- Research on the best ways to specify local services and manage them within the context of the car scheduling system.
- Development of methods to improve the visibility of the block assignments for interchange received traffic in order to improve the accuracy of the trip plans.
- Refining the logic used for monitoring plan adherence, and leveraging this to both improve trip plan accuracy and support the more effective management of potential scheduling failures before they occur.
- Design of support mechanisms for other specialized situations such as the management of empty cars for automotive and some types of unit traffic.

Car scheduling systems are critical to the functioning of the railroad. They are very complex systems, and typically are embedded in a larger systems environment that touches on every aspect of the railroad's operations. As a result, railroads tend to approach changing the car scheduling systems with great caution. Nonetheless, these systems also represent a great opportunity to leverage the huge amounts of data collected by each railroad, and seek out ways to improve their efficiency and the level of customer service they deliver.

References

Armstrong A, Meissner J (2010) Railway revenue management: Overview and models. Working Paper (available at http://www.meiss.com). Lancaster University Management School

Federal Railroad Administration (FRA) (1980) Missouri Pacific's computerized freight car scheduling system, advanced system study, June 1980. U.S. Department of Transportation, Federal Railroad Administration, report number FRA-OPPD-80-4

Federal Railroad Administration (FRA) (1981) Evaluation of the MOPAC's freight car scheduling system, June 1981. U.S. Department of Transportation, Federal Railroad Administration, report number FRA-ORRP-81-6

Fisher ML (1981) The Lagrangian relaxation method for solving integer programming problems. Manag Sci 27(1):1–18

Freight Car Utilization Program (FCUP) (1981) The impact on freight car utilization of operating and service plans, task force 2, FCUP, Association of American Railroads, final report, March 1981, AAR report no. R-464

Glover F, Laguna M (1997) Tabu search. Kluwer Academic Publishers, Norwell, MA

Gorman MF (1998) The freight railroad operating plan problem. Ann Oper Res 78:51–69

Green Cargo. Event management: self study materials for customers. http://www.greencargo.com/Global/Kundservice%20jvg/Tracktrace/Eng/EM%20selfstudy%20material%20for%20customer%20eng_final.pdf

Harrison EH (2005) How we work and why: running a precision railroad. Canadian National Railway Company

Huntley CL, Brown DE, Sappington DE, Markowicz BP (1995) Freight routing and scheduling at CSX Transportation. Interfaces 25(3):58–71

IBM. The optimization of global railways. http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/smarterrail/

Jeppsson E (2010) Green Cargo capacity booking process, 2010 Oliver Wyman rail planning conference, Princeton, NJ. http://blogs.oliverwyman.com/rail/2010/10/21/2010-rail-planning-conference-presentations/

Keaton MH (1992) Designing optimal railroad operating plans: a dual adjustment method for implementing Lagrangian relaxation. Transport Sci 26:262–279

Kraft ER (1998) A reservations-based railway network operations management system. Ph.D. dissertation, Department of Systems Engineering, University of Pennsylvania, Philadelphia, PA. UMI order # 9829930

McBain JT (2000) Leading a transportation revolution, Canadian National Railway Company, U.S. Securities and Exchange Commission Filing, Commission file no. 333-94399, p 14. http://secfilings.nyse.com/filing.php?doc=1&attach=ON&ipage=1031421&rid=23

Murray T (2006) Rails across Canada: the history of Canadian Pacific and Canadian National Railways. Voyageur Press, MBI Publishing Company, Minneapolis, MN, p 301

Norfolk Southern (2002) Norfolk Southern Corporation annual report for 2001, Norfolk VA

Norfolk Southern (2003) Norfolk Southern Corporation 2002 annual report and form 10-K, Norfolk VA

Norfolk Southern (2004) Driving performance: Norfolk Southern Corporation 2003 annual report and form 10-K, Norfolk VA

Norfolk Southern Corporation. Next generation car routing system at Norfolk Southern. http://www.informs.org/content/download/239255/2274025/file/SC1.pdf

Railway Age (2014) Obituary for Guerdon Sterling Sines, 1928-2014, railroad computer systems pioneer. Railway age, Friday, September 05, 2014. http://www.railwayage.com/index.php/news/guerdon-sterling-sines-1928-2014-railroad-computer-systems-pioneer.html

Chapter 5
Railway Blocking Process

Carl Van Dyke and Marc Meketon

5.1 Introduction and Background

Prior to the widespread adoption of unit trains and the rise of intermodal, most traffic moved in "loose car" or "manifest" service (also called "car load traffic"). In this type of service, sets of railcars are grouped together on a temporary basis into "blocks."

> A *block* is a group of cars that may have disparate origins and destinations, but will be moved together as a group from a common assembly point to a common disassembly point. At the disassembly point the block will be broken apart and the railcars will be formed into new blocks along with other railcars arriving from other locations. Thus, for an individual railcar, the origin and destination of a block may be either the same as the ultimate origin or destination of the railcar, or may be intermediate points in the railcar's route where the car is to be marshaled.

These blocks are moved by trains, where each train may carry a single block, or may carry multiple blocks. In this manner the railcars are relayed from their origin to their destination by being placed in a series of blocks, which are moved by a series of trains.

> In this context, a *marshaling or blocking plan* is the set of rules governing which blocks will be made at each location, and which cars will be put in each block.

Thus, the two main decisions in the design of a blocking plan are:

- The overall blocks to be created at each location.
- The specific traffic that should be placed into each block.

C. Van Dyke (✉)
TransNetOpt, Princeton, NJ, USA
e-mail: carl@cvdzone.com

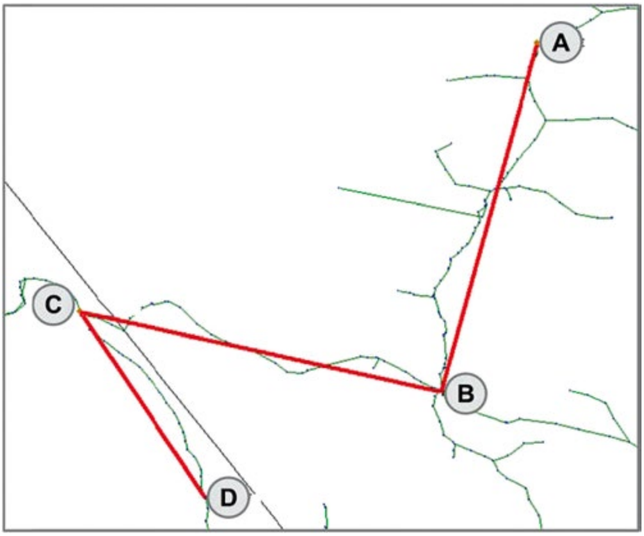
M. Meketon
Oliver Wyman, Princeton, NJ, USA
e-mail: marc.meketon@oliverwyman.com

© Springer Science+Business Media New York 2015

B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3_5

5.1.1 Impact of Blocking on System Efficiency and Service

The efficiency of a railroad's production system for carload traffic is underpinned by the quality of the blocking plan. To understand this, let us first consider the routing of an individual shipment using the blocking plan. For this discussion, please examine the "block route" map shown below:



In the above figure, a shipment moves on a series (or "sequence") of three blocks: A–B, then B–C, and finally C–D. If we think about the process of moving the shipment, it would be something like the following:

Activity	Location	Time impact driver
Shipper release at A	A	N/A
Pick-up car from shipper	A	Train schedule
Process car into block to B	A	Yard processing
Wait for train to B to depart	A	Train frequency
Arrive at B	B	Train schedule
Process car into block to C	B	Yard processing
Wait for train to C to depart	B	Train frequency
Arrive at C	C	Train schedule
Process car into block to D	C	Yard processing
Wait for train to D to depart	C	Train frequency
Arrive at D	D	Train schedule
Process car for delivery	D	Yard processing
Wait for delivery time to customer	D	Train frequency
Deliver to customer	D	Train schedule

While the train schedules influence the transit times, and can have some impact on the routing of the blocks, the blocking plan determines where the shipments will be handled, and the aggregate or overall routing of the shipment. Various analyses of carload shipments show that shipments can often spend more time in yards being processed and waiting for trains to depart, than in actual transit on trains (Little et al. 1992). Thus, the blocking plan strongly influences the efficiency and service level that shipments experience by determining where and how often shipments will be handled, and how direct the overall routing will be.

The core influences of the blocking plan can be summarized as follows:

Service levels: each handling (classification) of a shipment represents a delay in the forward progress of the shipment. If you consider a typical North American example of having one departure per day for each block based on the train schedule, an 8 h processing time for cars, and a perfectly random arrival pattern for the cars being placed in each block, then the average time in the yard would be [Processing Time] + [Headway]/2, where headway is the time interval between departing trains carrying the block. Using our example this would be 8 + 24/2, or 20 h of delay every time a car is handled.

Reliability: each handling (classification) of a shipment represents an "opportunity" for a failure, where failure is defined as the shipment not departing on the expected outbound train at the expected time. There can be many causes of such failures, for example they could be due to late in-bound train arrival, a lack of timely processing of the shipment into the out-bound block, miss-classification of the shipment, a problem with the out-bound train, detection of a mechanical defect in the railcar, or a lack of capacity on the outbound train (Kwon et al. 1995). Based on the author's experience, it has been found that such failure rates for connecting to a specific outbound train can exceed 20 % at major yards. These failures introduce variability into the overall transit time and thus adversely impact the product quality experienced by the shipper.

Circuity: shipments do not always take the most direct path from their origin to their destination. The difference between the most direct path and the actual path represents the excess distance the shipment travels, which is called circuity. Circuity can be introduced by both the blocking plan and by the train plan. Arguably, the largest source of such circuity is the blocking plan. Because the processing of railcars into blocks benefits from economies of scale both in the overall processing of the cars and in the ability to form larger blocks going longer distances, the author's direct experience indicates that it is often the case that shipments are taken out of route to reach larger yards. In other cases, the initial or final movement of the shipment may require an out-of-route local move to reach the origin or destination "serving yard" for the shipment. For these and other reasons, circuity may be introduced, and this circuity is often determined by the design of the blocking plan.

Yard workloads: the blocking plan determines where shipments will be handled. This means that the blocking plan determines the workloads at yards, in terms of both the number of blocks being made, and the total number of railcars being processed. The selection of which yards should perform which actions based on the blocking plan will thus determine the cost drivers for the yards, and can also influence the capital investments needed. The blocking plan modeling or design process can also be used as a tool for determining if specific yards can or should be closed or downgraded, and whether benefits would accrue from the upgrading of existing yards or the opening of new yards.

5.1.2 Specifying the Blocking Plan

Most blocking systems are location-based, and strive to be consistent—that is to provide the same instructions to all railcars or shipments with similar or identical attributes.

It is the author's understanding that before computers, blocking instructions were maintained in written form at each yard. Based on a railcar's destination, and perhaps a small number of other attributes or special conditions, a clerk could look up the block assignment in a paper blocking guide, and determine how to route the railcar.

Blocking plans were computerized well before the ability to create a computer-generated trip plan or car schedule was developed (see Chap. 4), with the Southern Pacific TOP system, and Missouri Pacific TCS systems being among the best examples (IBM; Railway 2014). This computerization process focused on converting the idea of the location-based blocking book into a similar set of location-based computer rules. These systems were enhanced by allowing a large set of shipment attributes to be considered when selecting a block for a shipment. This created a double-edged sword, simultaneously providing a great deal of control over how shipments were routed and greatly increasing the potential complexity of the rule sets.

The vast majority of railroads worldwide use some type of location-based, rules-driven, blocking look-up tables. The one major exception is the concept of using an algorithm for the generation of the railcar (shipment) to block assignments. Such an algorithm was developed by Norfolk Southern, and is now also used by Canadian Pacific Railway through adoption of the NS system (Norfolk Southern Corporation). It is the author's understanding that the development of an algorithmic capability is also under consideration at several other railroads as well. The algorithmic approach still uses rules, but also relies on business logic that takes a network perspective to determine the best or lowest cost sequence of blocks to use for each shipment. On the one hand, algorithms can increase the ease of plan maintenance, and allow for faster changes to the plan. On the other hand, algorithmic blocking can be more challenging to manage, and the user may have less control over the routing of specific shipments.

While concepts of "dynamic blocking" exist (Kraft 1998; Norfolk Southern Corporation), at most railroads the blocking plan is fairly static, and rarely changed on a "real-time" basis. This is true of both the algorithmic and table-based systems.

The concepts of table-based and algorithmic-based blocking are explored in more detail below, and the dynamic blocking concept is explored further in Chap. 4 on car scheduling. The authors have had direct experience with the design and data contained in the blocking systems at about a dozen major railroads in North America, Europe, Asia, and Africa (plus numerous smaller railways), as well as several planning systems and modeling tools for the design of blocking systems. The discussion that follows is based largely on this first-hand knowledge.

5.1.3 Plan Complexity

In the simplest approach to blocking, the final destination of each traffic record would be the sole determinant of which traffic should be placed in each block. Thus, the blocking plan specification would be based on a single attribute—final destination. However, for a variety of reasons, railroads use many other attributes to determine which shipments go in each block, greatly increasing the complexity of the blocking plans. A typical railroad will use between 20 and 30 attributes on a regular basis in assigning shipments to blocks.

Examples of reasons why these additional attributes are used, and specialized blocks are created include:

- Service differentiation—alternate blocks and trains are often provided for specific types of traffic such as intermodal, automotive or grain traffic, or to separate out unit train traffic.
- Restricted routings—some traffic must take specific routes due to safety considerations or to avoid damage to selected commodities. For example, some railcars may be speed-restricted, have clearance restrictions (cannot use some routes due to the height of bridges or tunnels), contain hazardous materials that must be taken over specific routes, or contain commodities that should not pass through a hump yard due to potential for damage.
- Interchange blocks—traffic bound to an interchange point with another railroad may need to be separated out by destination on the receiving railroad in order to improve service. In some cases there may be more than one receiving railroad at an interchange, and separate blocks may need to be made for each.
- Local blocking—traffic destined to the same station may need to be broken out by customer at the station, or by different parts of a customer's plant based on commodity.
- Empty blocking—in some cases empty railcars are routed differently, or the railroad wants to group the empty cars together (by car type) in order to expedite the movement of these empties to customers for loading.
- Specialized services—for a variety of reasons railroads enter into commercial agreements with customers to provide various specialized services that require shipments to be handled in a specific manner, and this results in the creation of specialized blocking instructions.

The end result of the above considerations, and a variety of others, is that the rules for specifying the blocking plan can become quite complex. While the majority of traffic may use fairly simple, destination-based rules, the level of effort for specifying and maintaining the plan becomes driven by these specialized blocking rules.

5.2 Current Industry Practices: The Blocking Rules Concept

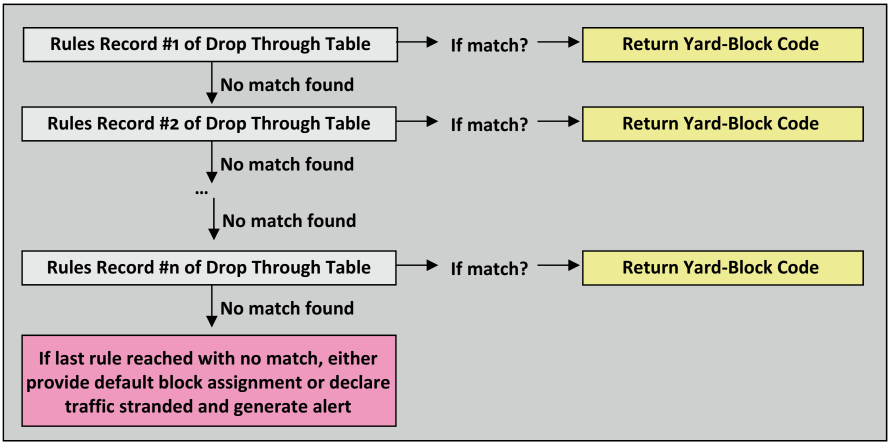
As noted above, most railway production blocking systems use a set of rules for determining which shipments should go in which blocks. These assignment processes are based on the current location of the railcar, and a variety of attributes related to the shipment and the railcar itself. Included in this process are a variety of special cases, which are discussed later in this section including block swaps, interchange blocks, and local blocking.

These rule or table-based blocking systems work as follows:

1. A set of rules are maintained for each location in the railroad network where blocking instructions must be generated. These rules are used for determining which block a particular shipment will be assigned to.
2. A request is made of the system to identify the block for a specific shipment. The blocking system is passed the current location of the shipment, the online destination for the shipment, and a variety of data on the overall shipment, the physical railcar being used, the content of the railcar, and the current status of the shipment.
3. The system processes the request by obtaining the blocking rules for the shipment's current location, and looking through those rules for the best match among the available blocks based on the information the system is given about the shipment.
4. The system returns a blocking code that it obtained based on its analysis of the rules for the current location.

While the details vary, all of the table-based systems work in a similar manner.

Central to the table-based blocking systems in widespread use is the concept of a drop-through rules table. In this type of table, the system starts with the first record in the drop-through table and tries to match the current shipment record to the criteria on the record in the drop-through table. If it matches, then the corresponding yard-block recorded on the record is returned. If it does not match, then each subsequent record in the drop-through table is checked until a match is found.


Rules Record #1 of Drop Through Table
If match?
Return Yard-Block Code
No match found
Rules Record #2 of Drop Through Table
If match?
Return Yard-Block Code
No match found
...
No match found
Rules Record #n of Drop Through Table
If match?
Return Yard-Block Code
No match found
If last rule reached with no match, either provide default block assignment or declare traffic stranded and generate alert

Each rule is typically composed of a series of attributes that the shipment must match in order to be assigned to the yard-block. As noted above, the rules are maintained by the planner or through business logic in a specific order, and the first rule that matches is used, thus ending the search. The rules are generally organized by location or a small group of locations, so only those rules that apply to the shipment's current location are considered.

Some railroads do not support the manual ordering of the rules, but instead use business logic to order the rules. This approach has been observed by the authors at two major international railways. In these cases, the rules are typically ordered by complexity, where the rules with more attributes specified come before the rules with fewer attributes. Priorities are then assigned to the attributes to further order records with the same number of attributes specified. In some cases, the user may also be able to specify rule priorities to change the relative order of the rules.

Based on the author's experience, most railroads use in the range of 20–30 separate attributes in their rules systems. Each rule typically uses only a small number of attributes. The assumption is that the values of all of the non-referenced attributes do not matter, and the shipment can have any value for those attributes during the matching process for that rule. By putting the rules in a specific order, lower rules can take advantage of the filtering effects of the prior rules. For example, consider the case where we want intermodal cars destined to location X to go in one block, and all other car types destined to X to go in a different block. We would specify the first rule as requiring all cars of car type P, Q, or S (intermodal car types) with a destination of X as going in block one. The second rule would simply say that all cars with a destination of X should go in block two, taking advantage of the fact that we already siphoned off the intermodal cars into a different block.

At a large yard, there can be several hundred blocking rules, and in some cases over a 1,000 rules. A smaller yard that only makes one or two blocks may have very few rules. The interactions and ordering of the rules are critical, and as a result the rules are generally maintained by a small number of highly trained individuals.

The rule attributes can generally be broken into several groups that include:

- *Primary traffic destination attribute*: There is generally a single destination code that is treated as the primary traffic destination attribute. The set of possible primary traffic record destinations that can be carried by a block is present in almost all rules. Several railroads organize their rules so that the rules at the end of the drop-through list are made up of only primary traffic destination attributes and each traffic destination appears exactly once among all of the destination-only rules emanating from a given yard. This ensures that all traffic will be routed. Because other attributes (as listed below) are also used in routing the traffic, the same primary destinations may appear multiple times across the more complex rules containing a mix of attributes. The primary destination codes are usually coded as stations within the railway, and are not coded macroscopically as city/state or microscopically as zone-track-spot. Destinations outside of the railway network typically have a predetermined interchange location that is used as the primary traffic destination code.
- *Shipment attributes*: these typically include the origin of the shipment, the destination, and the customer. Each of these pieces of information can be broken into a variety of separate pieces of information. For example, the offline origin can include the origin city/state, the origin SPLC code, the originating railroad, the railroad delivering the car to the railroad currently marshalling the shipment, the interchange received location, the online zone-track-spot data, etc. Similar details will exist for the destination, and the customer may be described in terms of both a code and a name, with distinctions made between the shipper, the consignee, the entity paying the freight bill, and the legal owner of the freight.
- *Railcar attributes*: the most commonly used attributes include the car type, the car's plate size, height, length, tare weight, the car's initial, the car's owner, whether the car is system, foreign or private, and the pool the car may be assigned to. In some cases, blocking systems can also use the car number.
- *Content attributes*: the most commonly used attributes are the net weight or gross weight, the load/empty status, and the commodity in the car. The commodity is typically expressed in terms of the STCC code or an internal, railroad commodity designation. For some commodities such as hazardous materials a special version of the STCC code may be used, or specific routing instruction codes may be applied. In some cases there may also be codes related to customs clearance, oversized dimensions, or other special considerations. For empties, the previously loaded commodity is often identified. Even the load/empty status can come in multiple flavors on some railroads.
- *Current status attributes*: this information relates to specific information on the status of the car at the moment that the classification request is made. Typically this consists of some combination of the current location of the car, the train it arrived on, and the yard-block or train-block it was on when it arrived at the current location.
- *Other attributes*: a variety of other attributes are used at various railroads. Examples include special codes specifying the run-through block the car is to be placed in (when going to a different railroad), routing instructions such as a

requirement that the car be weighed or cleaned somewhere in its routing, or a requirement to make intermediate stop offs. Another example is cars that have a mechanical problem and must be sent for repairs.

A variety of special cases can be handled. Two common examples are the blocking for local services and cars that do not have a destination specified. In some cases, cars that do not have a destination must be placed in "hold blocks" for manual handling, and in others "flow rules" are used to advance the cars in what is considered to be generally the right direction.

From a modeling perspective, the above approach represents three significant challenges:

1. Most production systems use a rules-based approach. If an algorithm is used in the planning process, how does one translate the results into a table-based form that can be used to direct the flow of the shipments?
2. The wide range of attributes and decision criteria reflect the overall complexity of the problem and the need to view the blocking problem from a multi-commodity perspective. This ultimately represents a huge challenge in the design and use of optimization or algorithmic strategies to design and improve blocking plans.
3. Given the nature of a rules-based system, significant challenges can arise in trying to improve the plan or even make modest changes. For example, determining what traffic should use a new yard can be very difficult because under a table-based approach, traffic will not naturally "flow" to a new yard, no matter what its cost.

Solutions to these problems and others will be explored further in this chapter.

5.2.1 Yard-Blocks, Train-Blocks, Class Codes, and Block Swaps

To provide some additional context for railroad blocking systems, there are a number of additional concepts that need to be understood, including the ideas of a *class code*, *yard-block*, *train-block*, and *block swap*.

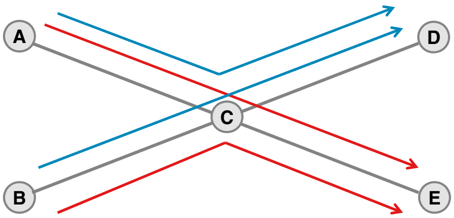
Perhaps for historic reasons, most blocking systems do not provide a definition of a block assignment in terms of a block origin, destination, and block name. Instead, they provide a "yard-block code," which is variously referred to as a "tag" or "class code," or in the case of CSX Transportation an "IYSC" or "inter-yard switching code." This "class code" is simply a name for the block, and does not specify where the block is going (except to the extent that the name matches a physical location on the railroad). The exception to this is the Norfolk Southern/ Canadian Pacific system, which produces a full block definition including a destination. In most systems, trains specify a separate concept called a "train-block" that provides the pick-up location for the block, the set-off location, and a block name.

Class codes are then associated with the train-block. Since the class codes do not have a destination, the destination becomes the location where the train-block is set off. On the one hand, this makes it very hard to validate that appropriate class codes have been assigned to a particular train-block; on the other hand, it provides flexibility to send the same class code/yard-block to different locations by day-of-week or based on other factors related to the available train service.

In many cases, railroads assign multiple yard-block codes to a single train-block. This is often done to give visibility to subsets of cars within a block. There are at least three common reasons for doing this:

- One reason is to provide information on special cars in a block. For example, if a block contains some "hot traffic" that must be protected in the classification process at the next yard, this traffic can be identified by giving it a unique class code.
- A second reason is to provide visibility to the classification work that must be done on a block when it arrives at a yard. For example, for a local block arriving at a serving yard, one might show the out-bound class codes from the serving yard for each car on the in-bound block.
- A third is to allow easy "swinging" of traffic for capacity management purposes. For example, in a block going from yard A to yard B, one might segment the traffic into several groups. One group would be traffic that is absolutely best served by going to B, and there might be two other groups, traffic that would do all right going to yard C, and traffic that would be handled satisfactorily if it went to yard D. If yard B becomes congested, one could then easily divert some traffic to yard C or D by "swinging" the appropriate class code(s) to a different train-block that was going to C or D.

One consequence of the class code-based approach is that many railroad production systems do not know the intended destination of a class code or yard-block, making validation and support for block swaps difficult. A block swap is a situation where a train-block is passed from one train to another on an intact basis with no switching of the individual railcars within the train-block. Block swaps are done primarily for operational reasons in situations where the scheduling of a through train is not practical. The classic example of a block swap is shown in the figure below:



In this example, yard A makes a block for yards D and E, and yard B makes a block for yards D and E. In the simple case that each of these four blocks is only large enough to fill one-half of a train, a block swap provides a potentially desirable operational option. Yard A would create a train A–E for yard E, and yard B would create a train B–D for yard D. From their origins, each train would carry both a D and an E block, meaning each train would be full. At yard C, the A–E train would set off its D block, and pick-up an E block set off by the B–D train. Likewise, at yard C, the B–D train would set off its E block, and pick-up a D block set off by the A–E train. In this manner, a full switching of the railcars at C is avoided (minimizing work and potentially dwell time), and each train can operate to its capacity from origin to destination.

For the class code-based blocking systems, where the destination of the block is implied by the set off of the trains, block swaps represent a significant challenge. These systems often pass all shipments through the classification or blocking system every time a set off occurs. This could cause the shipments in the block swapped blocks to have their class codes or block assignments changed by the system (in our example above, by processing the cars on the D and E blocks through the classification system at location C). To counter this, many railroads end up putting in extensive instructions into their blocking systems to identify block swaps at the block swap locations, and ensure that the block assignments are protected as part of block swaps. Typically the systems use the in-bound train number, in-bound class code or yard-block, and other factors to create rules to ensure that the out-bound class code or yard-block remains the same.

These block swap rules are very difficult to maintain, and can be a significant nuisance during the creation of planning systems. In general, we will not address this issue further in this chapter. Most optimization systems and algorithmic blocking systems have full visibility to the block destination, which greatly simplifies the block swap issue—a block swap is assumed anytime a block is set off short of its destination.

5.2.2 Local Service

The usual notion of a block is that it is a group of cars that are assembled at one yard, which is then transported and delivered intact to a disassembly point. Specification of local blocks represents a challenge in that many distinct blocks or grouping rules may be required to identify all of the required, customer-specific groupings. One approach is to fully enumerate each customer block in full detail, resulting in numerous individual blocks with very specific rules. An alternate approach used by some railroads is to organize the blocking or routing information around the local services that will directly pick-up and deliver cars to customers. In this second approach, rather than specify individual blocks to each customer, a range of stations or customers may be specified instead, tied to the specific local service that will serve the stations or customers. In effect, the individual local blocks become implicit within the specification of the local service.

One example of local blocking or routing rules tied to a specific local service is known as a *gatherer/distributor* block. Typically, such blocks use a description that on the surface appears to be a single block to actually represent multiple blocks. To give an example, suppose a block has origin A, destination F, and traffic destinations B, C, D, E, and F, but also has a special flag set to indicate it is a gatherer/distributor block. The blocking system recognizes this flag and implicitly treats this as 10 blocks: The A–B, A–C, A–D, A–E and A–F distributor blocks (called that because cars are distributed from A to locations B–F, which are usually considered to be customer locations) and A–F, B–F, C–F, D–F and E–F gatherer blocks that pickup local traffic and gather it for further processing at F. It is possible this block has the same origin as destination. In practice, the train that carries the block may only stop at a subset of the traffic destinations, and the set of implicit block locations is the intersection of the traffic destination range and the train-route locations.

There is a second form of local service specification that gives more details on where within a customer location cars should be placed, typically at the zone-track-spot level. A zone-track-spot is a way to specify a specific siding at a specific customer, in effect a form of detailed addressing of locations within a larger station. An example is an automobile manufacturer that has specific locations for auto parts separate from locations for multilevel auto racks, but all of which is considered one station by the railroad's systems. Another example would be a mine that has some tracks in the same yard for various chemicals needed in the production of the bulk product and some tracks for loading the bulk product. The difference between this type of local block and the gatherer/distributor types of local blocks is that the zone-track-spot level blocks are usually describing movement within a station, while the more general local blocks describe movements between stations.

5.3 The Table-Based Blocking Systems OR Challenge

Given that most of the production blocking systems used by railways are table-based, and most OR-based approaches do not work well with tables, significant challenges arise in the creation of practical analytic and OR-tools to support blocking plan design. In short, while new plans can be developed using OR methods, these plans cannot be readily translated into a form usable by production blocking systems.

This situation gives rise to the need for several different types of algorithms and analytic techniques. These include ways to:

1. Translate algorithmic solutions into table-based solutions.
2. Develop incrementally focused optimization techniques.
3. Determine the quality of the current car routings in the existing table-based blocking rules, and suggest improvements.

To understand each of these issues, we must first understand in more detail the challenges of managing table-based approaches. Table-based blocking systems are easy to understand in concept, and work well in the sense that you get exactly what you specify in terms of car-to-block assignments. However, over time the rules can get very complex, and problems can arise.

One series of problems is that most changes to the blocking plan are based primarily on manual observations, and the personal knowledge of the operating plan by the persons making the changes:

- These observations tend to be localized in nature, which means that they often miss the network effect of changes. Because the system does not take into account the network impacts of a change, all changes by their very nature tend to represent local modifications to the blocking plan, unless the planner has a bigger picture perspective and acts upon that in a complete and thorough manner. As an example, the blocking tables that feed one yard can be modified independently to redirect traffic away from that yard to other system locations. This type of change may solve a local problem, but can overload the system at other locations or lead to inefficient routings. The results of this myopic view, when compared to a network-based view, are an increase in car-miles traveled, additional handlings, and potential delays due to unforeseen congestion incurred when transporting the current and rerouted traffic.
- The manual process can be efficient for small changes, but can also be very time intensive for large changes, which can manifest itself through slower response times to both planned and unplanned network disruptions. For example, if a line experiences an unplanned service outage, the planner must:
  - Identify all blocks that are affected by the disruption.
  - Manually identify acceptable reroutes for the affected traffic.
  - Manually enter the reroutes into the system by changing numerous rules at multiple locations.
  - Review the changes to ensure that (1) the reroutes are entered correctly, and (2) that the reroutes have the desired results.
- The overall process is by its nature very dependent on the skill level of the planner and can easily result in incomplete changes being made. For example, a complete job of introducing a new block requires that changes be made in not just the yard where the block is being added, but also at upstream yards. One must consider all of the traffic at the yard where the block is being added, and make sure that the rules for all of this traffic are changed, requiring many separate edits. Furthermore, by introducing a new block, it may make sense to reroute traffic to the yard where the block is being added, which requires both the vision to identify this traffic (which is non-obvious), and the need to change the blocking at various "up stream" yards to redirect this traffic. Finally, the downstream routes/blocking should be checked to make sure that the routes for the redirected traffic are efficient ones all of the way to destination.

One proposed solution, which can address many of these issues, is to replace the table-based blocking system with an algorithmic blocking strategy. This is what was done by Norfolk Southern and Canadian Pacific, and is under consideration at other railroads.

5.4 Algorithmic Blocking

The fundamental foundation of algorithmic blocking is the notion that given a set of blocks, one can find a shipment routing by finding a weighted shortest path across a network formed from the blocks. In this network, each block represents a link going from one yard to another. The cost associated with each block is generally a function of the end point yards, the physical lines traversed, the type of block, and the type of traffic being handled. The goal of the cost function is often not to represent true handling costs, but instead achieve an outcome that is consistent with current operating practices. For example, yard costs are often reflective of the bias of a railroad to use hump yards in preference to flat yards, and larger yards in preference to smaller yards.

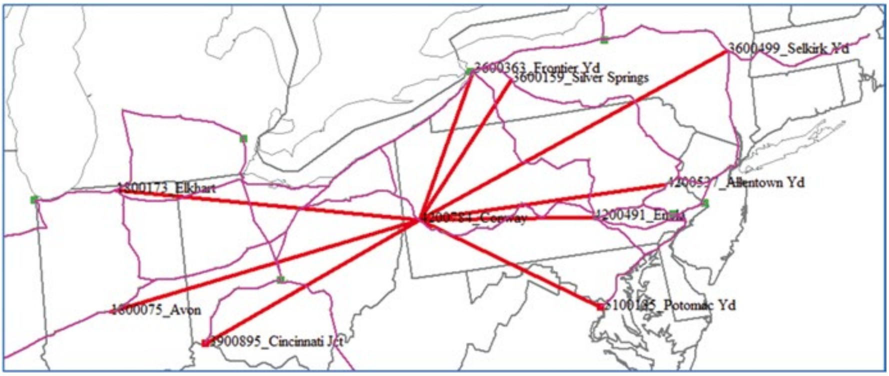
Once the blocks are defined, and the costs for each block are determined, a network can be created where the nodes are the starting and ending points of each block, and the links or arcs are the blocks themselves. By using a simple shortest path, we can then quickly determine the lowest cost routing for each shipment over a particular set of blocks.

There are still rules in an algorithmic blocking system. These rules are typically broken into two types, which are sometimes called “absolute” and “permissive.” An absolute rule acts much like the rules in a table-based system and specifies that the specific cars matching the rule must be placed in a specific block. Permissive rules simply specify which blocks could be used by a shipment, and are used to develop a list of potential or candidate blocks for moving a shipment. These rules also help to dictate the cost of using a particular block. For example, one might designate a block as being an intermodal block. If the shipment the system is trying to route using the algorithm is an intermodal car, then this block would be eligible for consideration. If the shipment was a general merchandise shipment, this same block would not be considered by the algorithm. In this way, the user can control the choices available to the algorithm when it selects blocks to move a particular shipment.

Both the absolute rules and the permissive rules are based on matching shipment attributes to the attributes associated with each rule. There are no limitations to the types of attributes that can be used, and these attributes are generally the same as those described above for the table-based system.

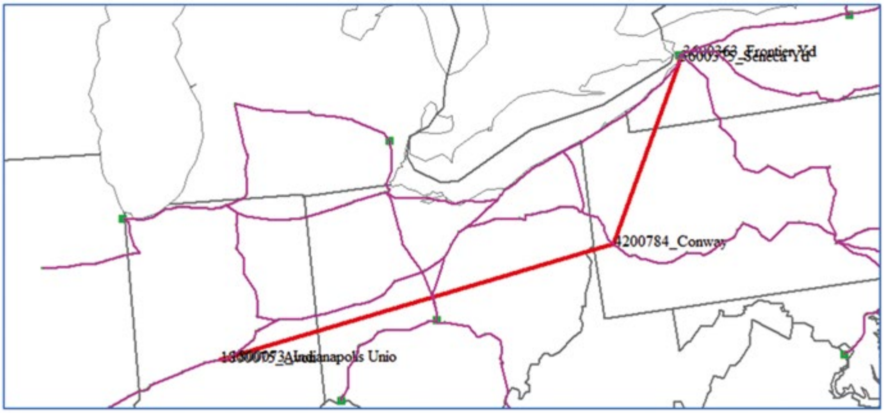
The most difficult part of defining a blocking plan is usually the specification of the car-to-block assignments through tables. The algorithm simplifies this tedious step, saving significant amounts of time and allowing large-scale blocking plan changes to be implemented more quickly and accurately than in a table-based system.

To understand the concept of a "network of blocks," please see the following diagram that shows all of the outbound blocks made from one location:



In the algorithmic process, each of these blocks is considered a network link. By combining this single location view with the blocks made at all other locations, we can create a complete network of blocking options. The costs of each link (or block) in the network is determined by the attributes of the shipment, the yard where the block is made, and various user controlled parameters. Only those blocks that are eligible to carry the car being routed are considered in this process. Where an absolute rule exists for a particular car at a particular yard, only one link would emanate from that yard.

Once formed, the algorithm can find a complete, end-to-end solution for routing a car across the blocks. Such a solution might look something like the illustration below that shows (although some details are hard to see) a traffic record from Indianapolis, IN to Seneca, NY taking four blocks: First is a local block to "Avon yard" near Indianapolis, second a block to Conway yard near Pittsburg, PA, third a block to Frontier Yard near Buffalo, and fourth a local block to Seneca yard.



In order to work, the blocking algorithm must know the destination of each block. This is a major difference when compared to the table-based blocking systems, and can be used to provide guidance to the trip planning system with respect to the validity of block-to-train assignments and the location of block swaps.

The real strength of this algorithm-based process is the ability to assign cars to blocks with a significantly reduced need for large tables specifying which cars are to travel in which blocks. When major changes need to be made in a table-based environment, the editing of the tables with their thousands of entries becomes a major barrier to the ability to undertake the change. Furthermore, in the planning environment such tables are a barrier to examining the full breadth of options available. In addition, the table-based approach is strongly influenced by the skill level and care taken by the analyst. While the algorithm-based approach cannot guarantee a specific outcome for each shipment being routed, it does assure that cars are routed consistently and efficiently. Furthermore, by adding in “absolute rules” the algorithm can be forced to produce specific outcomes when necessary.

Norfolk Southern and Canadian Pacific are the only railroads known to the authors to be using an algorithmic-based approach in their production systems. A number of railroads use algorithmic approaches in the planning process and to support optimization. At this time, to the best of the author’s knowledge, the translation of the algorithmic planning and optimization results into table-based solutions for use in the railroad production systems is a largely manual process.

5.5 Examples of Areas Presenting OR Challenges

The issues and analytic needs related to the blocking plan can be divided into several sub-problems or topics:

- *Blocking plan design*—typically an offline process that results in either incremental plan changes on a daily or weekly basis, or more sweeping changes on a more periodic basis.
- *Specialized blocking situations*—due to a combination of factors, there are many specialized situations that must be addressed by blocking systems. Examples include the need to specify blocking between railroads, separate service for different types of customers and lines of business, the need to specify how local services will be produced, and the management of capacities. When applying OR techniques, one often simplifies the problem in order to generate feasible solutions within acceptable computational limits. However, these simplifications often mean that either only a subset of the business can be modeled or optimized, or the solutions require significant manual adjustments to permit their use for actual operations. This is a significant ongoing limitation for most algorithmic and optimization-based blocking tools.
- *Blocking plan optimization*—a number of optimization methods have been developed, and applied with varying degrees of success (Ahuja et al. 2007; Van Dyke 1986, 1988; Bodin et al. 1980; Barnhart et al. 2000; Newton et al. 1998;

Newton 1996; Crainic et al. 1984; Gorman 1995; Keaton 1989, 1992; Yaghini et al. 2012). In addition to the citations provided, a number of organizations have successfully developed and deployed their own optimization models such as Norfolk Southern, CSX Transportation, and Oliver Wyman. The currently available techniques have a number of limitations. Generally, they are only for a single line of business (carload, intermodal, etc.), and they generally can only handle a subset of the problem. In particular, most of the existing methods do not do a very good job of handling the design of local blocking plans, and tend to only support generic carload blocks, and thus do not take into account a variety of special situations. Additional challenges arise when it comes time to adopt the solution, particularly with respect to translating the optimizer results into a set of table-based blocking rules. In the author's experience, this results in a strong need to manually review optimization results, and for using experienced planners to pick and choose from the optimization results those blocking changes that should be implemented. It also means that using an incremental approach to blocking plan improvement may be more effective than a "clean sheet" type approach.

- *Dynamic blocking concepts (time-based blocking)*—most current blocking systems focus on either use of rules-based routing, or algorithms that minimize a combination of distance traveled and handlings incurred. A number of time-based strategies are also possible (Kraft 1998), and some are being explored by a few railroads such as Norfolk Southern Corporation (INFORMS 2010) and BNSF Railway. These include both continued use of fixed routings of shipments, and also dynamic routing strategies. Both the dynamic routing approaches and fixed routing strategies are explored in Chap. 4. The fixed routing strategies attempt to factor in transit time into the static cost formulas along with distance and handlings. These static routing strategies try to reflect the available trains to move each block, and in some cases may reveal that a different routing for some shipments may produce better results in terms of total transit time, with no appreciable change in the other cost factors. As noted above, the routings for shipments are still fixed, but time will thus be added to the decision factors. Under dynamic routing strategies, the current status of each train relative to its carrying capacity is taken into account in deciding which sequence of trains and train-blocks should be used to advance the shipment on a real-time basis. These approaches are being explored, and used to some extent by at least two North American Class I railroads. In both cases the approach leverages a time–space variant of the existing shortest path algorithms described in this chapter.
- *Block-to-train assignment*—blocks are carried by trains from their origins to their destinations. Typically, the train design problem is treated separately from the block design problem for both historic and problem complexity reasons. This topic is primarily addressed in detail in Chap. 1 on train scheduling.
- *Execution support*—the systems and processes that are used in real-time to assign shipments to blocks and route the shipments to destination could benefit from stronger analytics and decision support to help guide real-time management decisions. This topic is examined in Chap. 4.

5.6 Semi-manual Blocking Plan Design Techniques

This section describes various techniques used to develop blocking plans. In practice, most blocking plans are developed incrementally, however this section will review both incremental and clean-slate approaches that have been done in practice. This section is devoted to semi-manual techniques that have been used routinely by railways; we delay until Sect. 5.8 the discussion of automated blocking plan design techniques using optimization methods.

5.6.1 Incremental Blocking Plan Design Techniques

One of the most common activities is to identify incremental changes to an existing plan that will improve overall performance of the plan, or address specific issues such as keeping workloads at a specific yard within its capacity limits. Thus, in these techniques we assume that there is already an existing blocking plan.

When determining what type of incremental changes to make, the first consideration is to establish the strategic reasons for making a change. Two examples are:

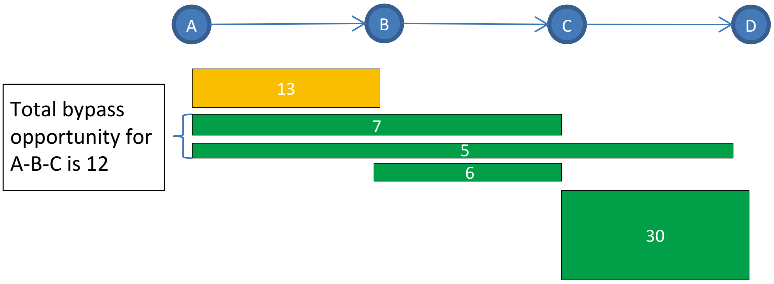
1. *Tuning an existing plan*: Traffic volumes have changed from when the existing blocking plan was created, and there is a need to fine tune the plan to better match the new levels of traffic.
2. *Change traffic volume at a yard*: There are various reasons that planners often have for wishing to increase or reduce the classifications per day made at a yard. A frequent reason is when traffic volumes change, a yard might become overloaded. In some cases, even a seemingly small change such as increasing from 2,000 classifications per day on average to 2,050 per day, proves to be a tipping point causing instability at the yard. Less common adjustments are during studies of yard expansions, where capacity is increased and the cost per classification at the yard declines. One part of the overall analysis on the cost/benefits to expanding the yard is to understand how the changes affect the overall network.

5.6.2 Tuning an Existing Plan

Two of the most common tools used in undertaking incremental, manual tuning of a blocking plan design are (1) the identification of bypass opportunities and (2) reviewing low-volume blocks.

A bypass opportunity occurs when two blocks in a sequence carry many of the same cars. For example, if block A–B carries 25 cars/day and block B–C carries 18 cars/day, and of these cars, there are 12 cars/day that travel on the A–B block and then the B–C block, we have a bypass opportunity of creating a block A–C to carry those 12 cars/day.

There can also be bypass opportunities that span several blocks, and in many cases the various bypasses can compete with each other for use of the same traffic. This is illustrated in the diagram below, where there are 13 cars/day that use the A–B block but not the B–C block, 7 cars/day that use the A–B–C sequence but not the C–D block, 5 cars per day that use the A–B–C–D sequence, 6 cars per day that use the B–C but not the A–B or the C–D blocks, and 30 cars/day that use the C–D but not the B–C block. This results in several bypass opportunities, one for 12 cars/day by creating an A–C block, one for 5 cars/day by creating an A–D block. However, if the A–D block is created, along with the A–C block, the A–C block would have only 7 cars/day instead of 12 cars/day.



Usually, a list of all the bypass opportunities is created, and filtered for larger opportunities and sorted either by cars per day or car-miles/kilometers per day. Further filtering is done to eliminate artificial opportunities caused by interchange and local blocks. Interchange received blocks are often a source of bypass opportunities, however to implement these requires the cooperation of the delivering railroad, which can require a significant commercial negotiation. For this reason, interchange bypasses are often handled as special cases or skipped in most analyses. Likewise, many local block bypass opportunities are not feasible due to operational and yard capacity constraints, and thus must also be ignored in the bypass analysis process.

Calculating bypass opportunities creates indications of where potential new blocks could or should be made. However, the analysis is only indicative. It takes further analysis before the planner can actually add a new block. For example, the planner needs to decide if yard A has the capacity to make an additional block. If not, the planner might need to eliminate a low-volume block that originates from A. Adding a bypass block may also substantially reduce the volumes of other blocks. Therefore the planner needs to be judicious in the application of bypass blocks, and generally the planner makes only a few changes before flowing traffic over the revised plan to obtain more precise estimates of block volumes.

Finding bypass blocks is the same whether the underlying blocking plan is algorithmic or table-based. However, implementing bypass blocks in a table-based plan is notably harder since the rules need to be setup to attach the right traffic to the new block. When the rules are mostly traffic destination based, then the yard relaxation algorithm (described in more detail later in Sect. 5.6.6) works well to automate the changes in the blocking rules related to which traffic destinations should be assigned to the new block. This same process can also be used to identify other traffic at the yard that may want to take advantage of the new block, where this additional traffic is not currently riding on the bypassed block.

The other tool used in refining blocking plans is fairly simple—identify zero and low-volume blocks as candidates for removal. In many cases, these low volume blocks need to exist to protect service to lightly used stations. Thus, removing low, but positive volume blocks, may result in some traffic being stranded if the removed block was the only way to reach the impacted location. In table-based systems, removing a block almost certainly dictates necessary changes in the rules for other blocks originating from that yard, since the block usually is intended for specific traffic destinations. In algorithmic blocking, removing a block is much easier, although it may still result in unmoved traffic. While removing a zero volume block will not result in any stranded traffic, there may be seasonal or periodic traffic for the location that must be protected. In these cases, the zero-volume block cannot be removed.

After removing a low-volume block, there needs to be a check to see if the circuity for the new traffic routings is too large.

5.6.3 Checking Circuity and Excessive Handlings

Two other techniques that are employed in developing a good blocking design are circuity and excessive handling analysis. Circuity analysis computes, for each traffic record, the ratio of the distance of the traffic route as given by the block sequence to the distance of the shortest path between the origin and destination of the traffic record. Traffic with large circuity (often thought of as being 1.2 or larger) is studied to see if changes in the blocking rules or adding a block are warranted. The studies usually take into account the volume of traffic, ignoring very small flows of high-circuity traffic.

Excessive handlings is the analysis of the number of classifications a shipment undergoes. As a rule of thumb, high-volume traffic with four or five intermediate classifications are usually examined to see if additional blocks or blocking rule changes are necessary.

When adding blocks to solve either circuity or excessive handlings issues, the planner usually looks at existing blocks to see if there are any simple reroute options that would solve the problem. These reroutes can be identified using the relaxation techniques described in Sect. 5.6.6, or in some cases through bypass analysis.

5.6.4 Change Traffic Volume at a Yard

Tools used for adjusting volume (workload) at a yard differ between algorithmic and table-based systems. Usually the planner has a target traffic volume in mind that should get added or taken away from a yard. With blocking plans that are table-based, the planner tries to identify a set of traffic whose expected volume is close to the target and then creates specific rules that address that set of traffic.

Unfortunately, the result of repeated tuning of the volumes through the yards can lead to complex rules, underutilized blocks and costly routings in some cases. The general idea is to go to "upstream" locations that feed traffic into a yard, and change the traffic routings for selected traffic from these upstream locations to use alternate yards than the one with an excessive workload. Of course, in making these adjustments, the impact on the yards to which the traffic is redirected must also be assessed, and the resulting circuity and handling levels must be assessed.

With algorithmic routing, planners can change the penalty (cost) of classification at the yard with too much traffic to influence the amount of traffic being switched at the yard. However, that has two problems:

- The changes in traffic volumes are often a step function with respect to the classification cost. For example, as the cost increases, the traffic stays flat for a while, then has a "jump" down and then stays flat until the next jump occurs. This is caused by groups of traffic changing their routings as various tipping points are reached in the relative cost of using the target yard compared to other yards. In some cases it can actually be hard to find the right cost parameter to use, and due to the step function there may be no "right" cost parameter.
- Planners are often loathe to change the classification cost because it might change many routings that they do not want to change.

Because of these two reasons, typically planners using algorithmic blocking systems use the same techniques as those using table-based blocking systems, which is to change the rules on the blocks at the upstream yards in a precise manner. However, they may use changes in the yard's penalty or cost as a way of identifying potential candidates for reroute.

5.6.5 Designing Blocking Plans Using a Clean-Sheet Approach

In the previous section, we discussed the use of bypass opportunities and removal of zero or low-volume blocks to take an existing blocking plan and improve it. To start without an initial blocking plan (variously called a clean-sheet, greenfield, cold start or zero-based approach), the usual practice is to generate a simple, but largely feasible initial block plan based on a yard hierarchy, and then incrementally improve it as discussed above. Usually at this step only algorithmic blocking is used, generally with blocks that have no specific traffic rules (but may be focused on specific lines of business). This allows a good plan to be formulated more quickly. Additional explanation of building an initial blocking plan and optimizing a plan is given in Sect. 5.8 below. Often in clean-sheet approaches, the local plan is not changed, and only the longer distance blocks are examined. Furthermore, interchange blocks to and from other railroads may be kept frozen, as these can only be changed through bilateral negotiations with each of the connecting railroads.

In a typical manual approach the steps are as follows:

1. Create an initial block plan. There are two strategies:
   (a) Take the existing blocking plan, and remove all of the non-local and non-interchange blocks, or
   (b) Use a hierarchical approach to generate a starting set of carload blocks that cover the movement of all of the traffic (see Sect. 5.8 below). This plan typically connects local or serving yards to larger regional and system yards on a nearest neighbor basis, and then connects regional yards to system yards, and system yards to each other.
2. Either exclude the non-carload traffic from this process, or use specialized logic to create initial blocks for other lines of business (unit train, intermodal, automotive, and grain).
3. Flow the traffic over this starting set of blocks using an algorithmic approach.
4. Review the plan, looking at circuity analysis, excessive handlings analysis, and bypass analysis to identify where new blocks could be added, and low volume analysis to identify blocks that might need to be eliminated. Examine yard volumes and the number of blocks made at each yard, and identify where adjustments to penalty costs to bring these yards into conformance with their capacities might be made.
5. Make some of the changes identified in step 4 above, and flow the traffic over the revised blocking plan.
6. Repeat steps 4 and 5, monitoring various key performance indicators (such as handlings, car miles, yard volumes) until the plan is satisfactory.
7. As appropriate, review and revise the local and interchange blocks as a separate review exercise.

The above process can be automated through an optimization approach, which is discussed in Sect. 5.8 below. Using a manual approach, it is the author's experience that a well-trained team can complete the steps above in a 1- to 2-week period for a large railroad. While optimization can produce an initial plan in a few days (including setup time), we have found that the resulting plan must still be manually reviewed and refined using some of the above steps in order for the plan to be acceptable to railroad management.

5.6.6 Tuning Table-Based, Traffic Destination Attribute Rules Using Relaxation

As indicated in Sect. 5.2, the most common type of table-based rule is the set of allowed traffic destinations for a block. Due to the manual nature of maintaining table-based rules, sometimes individual traffic movements are routed inefficiently and at higher cost than necessary due to using a less than optimal set of traffic destination rules on the blocks. This typically happens when the blocking plan is changed, but not every destination assignment is updated to reflect the change. This can happen when blocks are added or removed, or when rules are introduced to change the amount of traffic handled at a yard resulting in some traffic being purposefully routed non-optimally. Later there may be no need to route the traffic that way, but the change is forgotten and not undone.

A technique, *called rule relaxation*, can be applied to discover cases of non-optimal routing. Rule relaxation applies algorithmic-based routing to a set of existing blocks by ignoring traffic destination-based rules. This is easiest done for railways that use two levels of rules—where one level is based on all the blocking attributes and is placed higher in the rule priority order, and the other level is based strictly on traffic destinations. The concept is that algorithmic routing will find the least costly block sequence.

The broad-based approach to rules relaxation applies this approach to a large set of blocks and traffic at a network level. For example, we might apply this to all general carload traffic. All of the blocks in the table based plan would be reviewed using computer-based business logic, and each carload block identified. The carload traffic would then be flowed across the blocks using a modified algorithmic approach. Any complex rule would be retained, and applied to the traffic on an absolute basis. But for traffic not hitting such rules, an algorithmic approach would be used to flow the traffic using the carload blocks identified by the business logic. This will result in more optimized routings for some of the carload traffic, where the changes can be identified through a comparison to the routings produced by the pure table based approach (possibly using the triplet analysis described below).

This approach, while powerful, can be difficult to use. It can result in a large number of routing changes, which then need to be reviewed by the planners. Experience has shown that many of these changes are either of trivial value or unacceptable for operational reasons. The result is that the cost/benefit of this approach can be perceived as negative, or the process can simply overwhelm the planer. Furthermore there can be no automatic adjustment of the rules based on this kind of analysis due to the risks of unintended consequences, which means that the changes must be manually entered into the rules tables.

To understand the reason that more complex rules may need to be retained in the relaxation process, consider a yard with four blocks A–B, A–C, A–D, and A–E. The rules for the blocks are prioritized as follows:

1. If commodity is hazardous and traffic destinations are X, Y or Z, take block A–B.
2. If traffic destinations are X, Y or Z, take block A–C.
3. If commodity is hazardous and traffic destinations are P, Q or R, take block A–D.
4. If traffic destinations are P, Q or R, take block A–E.

The A–B block is for hazardous materials going to X, Y or Z, and based on the rule ordering the A–C block will implicitly be for non-hazardous material for traffic destinations X, Y and Z. Let us assume that going to C is a better, cheaper route for traffic destinations X, Y or Z compared to going to B. Under an open relaxation schema that discards the more complex rules, these four blocks will no longer be ranked relative to each other, because the rules will now be permissive. Now block

A–C would be able to take both non-hazardous and hazardous material, and since going to C directly is less expensive for traffic destinations X, Y or Z, it will naturally attract all traffic going to those locations. This illustrates the pitfalls of using completely open relaxation.

There is a second form of relaxation, called *yard relaxation*, for table-based blocking plans that examines and recommends traffic destination rules at a single yard. It is simple, but is well received by the planners and is considered quite valuable. It is primarily for railways that use traffic destinations as the primary blocking rule attribute. It also has the advantage of limiting the amount of information that needs to be reviewed by the planner, making it a much more understandable and approachable way to improve the plan.

In simple terms this is an exhaustive search approach. To work, one selects a specific yard (which we will call yard A). One then takes some set of candidate traffic movements and tests how each traffic movement would currently be routed from yard A, and how the traffic would perform if it used each of the other blocks that are made at yard A. The cost of the current routing is compared to each alternative, and the cases where improvements are realized by changing the routing are identified. As with other forms of relaxation, care must be taken not to test inappropriate cases such as putting an intermodal movement on a coal block. To protect against this, such relaxations are often limited to only carload traffic and only carload eligible blocks made at yard A are tested.

This approach can use either the existing traffic at yard A as the candidate traffic, or can generate a set of candidate traffic movements. The generated movements can have the advantage of providing test cases for all possible destinations on the railroad, supporting a more thorough review of the routing rules. In the generated case the user typically specifies a standard profile for the traffic movement attributes, such as a generic, loaded boxcar carrying a common, non-hazardous commodity.

In more mathematical terms, the process can be expressed as follows:

1. Let the yard in question be called A, and let the destinations of the blocks that originate from A be $\mathrm{BD} = \{\mathrm{B, C, D, ..., H}\}$. As noted above the set BD might be limited to only the carload blocks.
2. Execute a double loop—first for each possible candidate traffic destination (say *d*) then for each possible block destination in the set BD (say *r*).
3. Find the block sequence from *r* to *d*, and add to this the cost to go from A to *r*.
4. After going through each *r* in BD, find the block destination *r** whose cost from A to *r*, then *r* to *d*, is the smallest.
5. Assign the traffic destination *d* to block *r**, and repeat for the other possible traffic destinations.

At the end of the process, each block that originates from A has a set of preferred traffic destinations that can be compared to the current routings.

When the yard relaxation process suggests moving a traffic destination from one block to another, the process should calculate various metrics such as the total distance and total block sequence cost. This gives the user the ability to examine the proposed traffic destination assignment and make changes to the rules if necessary, potentially ignoring proposed changes that have only a small benefit.

5.6.7 Additional Methods for Testing Plans

In the previous section, two plan testing concepts were mentioned: circuity analysis and excessive handlings. In this section, we discuss other tests that a good blocking design should pass.

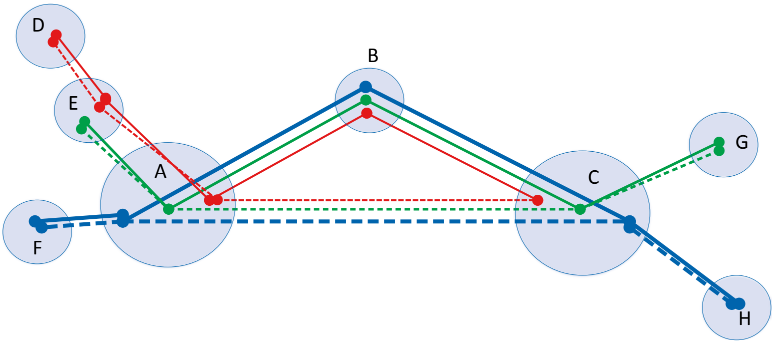
- *Unmoved traffic.* Most traffic—typically at least 98 % of the volume should have block sequences or routings. The large railways may have some traffic that cannot get a block sequence because no local block is defined either from an origin or to a destination, however operationally they usually know how to move this traffic should it occur.
- *Completeness.* The notion of completeness is that a blocking plan should be able to move any possible traffic that could at some point be tendered to the railroad. Even if all the existing or known traffic is moved, there is no guarantee that all elements of a future set of traffic records could be moved since the traffic set that is used for analysis (usually based on historical data) does not have all the combinations of origins, destinations and various other attributes that can arise. Railways sometimes have a "test traffic set" that they use for testing the completeness of the plan. It is generally composed of all origin/destination pairs for a generic shipment such as a standard, loaded box car, as well as all reasonable intermodal and automotive origin/destination pairs for the appropriate car types. This test traffic set may also have other records for specialty traffic cases.
- *Loops.* A common error in table-based blocking systems involves plans that generate block sequences with loops. For example, a traffic record from A to E might first take the A–B block, then the B–C block, then the C–D block, then a D–B block. At this point, the block sequence loops and keeps cycling. This would occur if, say, the D–E block has a lower priority than the D–B block and both could accept a traffic destination of E. Most block sequencing or routing procedures contain a test for loops, terminating the routing of individual shipments when loops are detected. Broader-based testing for loops can be done using the same "test traffic set" as is used for completeness testing (though loops are often caused by specialized rules for specific subsets of traffic).

5.6.8 Triplet Analysis for Blocking Plan Comparisons

Triplet Analysis is used to compare the block sequences from two different blocking plans for the same traffic set to understand fundamental routing differences between the plans. It works by examining the block sequence for each traffic record in the sample that has a different sequence between the two plans. Because there can be many individual traffic records that share the same routing difference, looking at individual traffic records can be time consuming and make it hard to identify the larger patterns. Triplet analysis attempts to identify the underlying patterns across multiple traffic records.

Triplet analysis has two primary values: It dissects and ranks the differences between two blocking plans, and it allows a user to examine selected differences and pick which ones to use. This is particularly important for using the current block optimization technology that cannot, by itself, produce a complete realizable plan but together with triplet analysis can produce useful modifications to an existing plan.

Consider traffic D–C, E–G, and F–H. Suppose in plan 1, the block sequences were (respectively) D–E–A–B–C, E–A–B–C–G, and F–A–B–C–H and in plan 2 the block sequences where D–E–A–C, E–A–C–G, and F–A–C–H. This is illustrated below, with the original block sequences in solid lines and the new ones with dashed lines.



All three traffic records have a sub-sequence of A–B–C in plan 1 and A–C in plan 2. Typically these routing differences occur for the same reason, hence grouping together these traffic records for analysis can be used to generate statistics that highlight the underlying differences between the two plans.

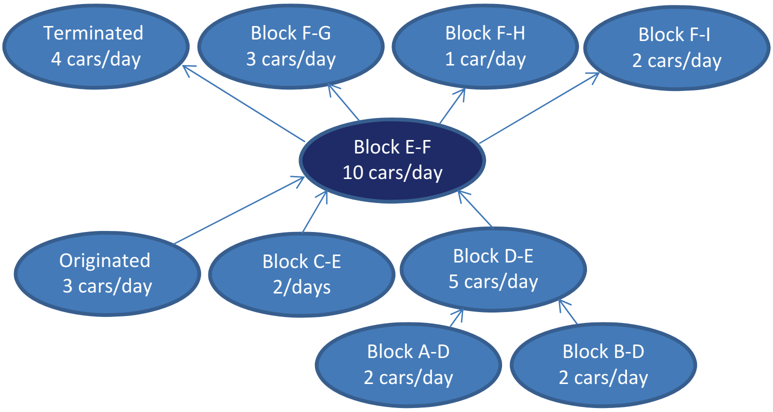
Triplet analysis has three components:

1. Identification of routing differences, as illustrated above.
2. Calculation of business statistics for each triplet such as the distance of the plan 1 route versus the plan 2 route, the car-miles/kilometers, the number of intermediate classifications, and the equivalent car-miles/kilometers when the intermediate classifications are converted to an equivalent distance metric. These statistics are critical to the ranking of the triplets.
3. Identification of which blocks occur only in plan 1, which blocks occur only in plan 2, and which ones are common. For example, if all three blocks (A–B, B–C, and A–C) are common, it suggests that the difference between the two plans is due to block-rule changes or change in the cost of classification at B.

It is called “triplet analysis” because the canonical example used when explaining how it works is the above example that involved three blocks, and in practice that is a common situation. Four blocks also commonly occurs (one route may be A–B, B–C, and the competing route is A–Z, Z–C) and sometimes more than four blocks.

5.6.9 *Tree View Analysis*

Tree views show how traffic flows through downstream and upstream blocks. Graphically, it is represented as illustrated below for the block E–F. It shows how the volume of block E–F flows from and into other blocks for a specified depth. The block D–E may have 20 cars/day, but only 5 cars/day subsequently go onto block E–F. Originations and terminations are usually shown for the E–F block and not the upstream or downstream blocks.



5.7 Specialized Blocking Situations

A number of specialized situations need to be factored into any blocking plan solution. These are discussed below.

- *Local Services*: Every railroad has a unique approach to the specification of local blocking. Many of the issues related to local services are discussed in Chap. 4, and to some extent in Sect. 5.2 above. In general, the most detailed rules in the blocking system are related to local blocking, and in some cases local blocking can represent 50 % or more of all the blocking rules. Both table-based blocking systems and algorithmic blocking systems require extensive rules to specify the local blocking because of the high degree of detail required to get shipments to the right customers on the right tracks. Furthermore, because many local block assignments occur at the destination of a trip, the algorithmic systems generally use a process similar to that used by the table-based systems for assigning the final yard-block code. At some railroads, the local delivery rules are maintained not in the blocking system, but in some kind of local service specification process that can be part of a separate local services database or part of the main train database. In these situations, the blocking system must either look at this external data source to determine the local blocking, or the blocking system must receive periodic rule updates that are derived from this external data source. From an optimization perspective, the local blocking is often treated as fixed, and the optimization focuses on the longer haul elements of the blocking plan.

- *Interchange blocks/run-through trains*: In some cases one railroad agrees to build blocks that "interchange" or go onto another railroad. For example, railroad A might agree to make five blocks for railroad B for interchange or handover at a specific location (or junction) in exchange for railroad B making five blocks for railroad A. Typically, the receiving railroad sends instructions to the delivering railroad that specify which shipments should go into each block. When these blocks are placed on a train that does not stop or get broken apart at the interchange, one gets a "run-through train." The biggest difficulties with interchange blocks and run-through trains are in maintaining accurate rules for assigning shipments to the right interchange blocks, and knowing in advance which block each shipment will be in when received from interchange. Problems with both issues can contribute to the generation of inaccurate initial trip plans on a real-time basis, as well as represent a challenge to the planning/modeling process. Optimization routines often get into trouble when they change the interchange blocks relative to the existing agreements with the other railroad in terms of either the number of blocks made or their content. From a blocking system perspective there are four things to consider:

  i. A back-up set of tables must be maintained for each pre-block the railroad makes that roughly mirrors the instructions that would otherwise be received from the foreign road (these instructions come through a data interchange process in North America known as the 419/420 message exchange process). These back-up tables ensure the continued functioning of the classification process in cases where the communications protocols fail, and support the modeling/planning process.
  ii. Each railway must maintain a definition for each pre-block it will ask a foreign railway to make, both to support the 419/420 process, and to provide back-up instructions for entry into the foreign road's computer systems.
  iii. Many railroads embed a special code on the waybill or movement record that reflects the pre-block assignment for both interchange received and interchange delivered traffic, often based on the 419/420 process, that can be seen and used by the blocking engine when doing its car-to-block assignments.
  iv. The railroad has the freedom to make pre-blocks in multiple locations on the railroad and the blocking system should be adjustable enough to optimize where traffic is classified for these blocks.

- *Handling of specific specialized services*: Most blocking systems are focused primarily on conventional carload traffic. Blocking can be entered into the system for a variety of specialized traffic such as grain, automotive, intermodal, and unit trains. Some services, such as unit coal traffic, do not fit very well to the traditional car scheduling and blocking paradigm. Others such as grain sometimes fall within the scope of the trip planning and blocking process, and in other cases do not. Finally, some services such as intermodal come pretty close to the trip planning/blocking process described above, but have complexities related to there not being a one-to-one correspondence between the railcars and the materials (boxes) being shipped. The result is a need for either special logic within the trip planning and blocking environments for each of these cases, or the exclusion of this traffic from the classical trip planning and blocking processes. The degree to which this can be done with any accuracy depends on a combination of business logic, data quality, and the way the blocking plan is designed. The degree of use varies widely for traffic such as grain and intermodal. The handling of these special cases is discussed in more detail later in this chapter.

- *Multi-location yards*: The concept of a multi-location yard is mostly specific to the rules-based blocking systems. Each set of rules is location-based. If one has many locations that have blocking rules, this can result in a huge number of rule sets and rules to maintain. In some cases there can be clusters of locations that will have similar or identical blocking. For example, a series of local yards all served by the same trains might have essentially identical blocking. To reduce the number of location-based rule sets and ease the manual maintenance process, the concept of the multi-location yard was developed. This is a situation where a set of locations all use the same set of blocking rules. To the extent there are differences between the locations, rules are created that use a "current location" attribute to restrict their applicability to only one location. These multi-location yards can increase the complexity of the blocking business logic in some cases, and there are indications that railroads are moving away from this concept. The multi-location yard concept is not used by algorithmic blocking systems such as the one used by Norfolk Southern and Canadian Pacific Railway.
- *Data clean-up issues*: A number of data sources are used as inputs to the blocking process, and many of them can have data issues associated with them. As a result, the blocking processes have a variety of mechanisms for correcting these data issues in order to improve performance. These include the ability to specify "variants" of spellings in the blocking rules, and "waybill correction" tables to standardize shipment attributes prior to their use by the blocking system.
- *Block assignment regeneration*: From the blocking plan generator's perspective, there is no specific requirement to "regenerate" shipment-to-block assignments. The blocking system simply processes requests when it is given the current location of the shipment, the targeted destination, and a set of attributes to be used in determining the appropriate shipment-to-block assignment. Based on that it provides back a block. External monitoring systems, including yard systems and trip planning systems, are responsible for determining when the shipment-to-block assignments need to be requested or regenerated.
- *Capacities*: In general, capacities of individual blocks are not considered in the design and maintenance of the blocking plan. In some cases capacities are considered during real-time execution of the plan with respect to specific trains and/or yards.

This subject is addressed in Chap. 4. During the design of a blocking plan, minimum volumes for individual blocks are often used to measure plan quality and as constraints on block formation. Maximum capacities of yards to handle railcars and maximum numbers of blocks that can be made at a yard are often considered during plan design and optimization. This use of capacities is explored further later in this chapter.

- *Hold Blocks*: Hold blocks are an important concept used to classify or sort a set of shipments into a group that does not have an outbound train, and thus requires manual intervention in the handling of the shipments. Essentially they can be viewed as a forced blocking plan failure. These hold blocks are used for a variety of purposes, but the most common is to collect shipments that will require manual intervention prior to onward movement. Common usages are for grain that may be assembled into solid or unit trains for onward movement, and for the collection of empty railcars. Hold blocks pose a particular challenge for algorithmic blocking systems that are focused on driving shipments to their destinations. They are often modeled as "regular blocks" from the algorithm's perspective, with a flag on them that indicates that they should not allow the routing process to progress once a shipment is assigned to such a block. This can also pose problems in statistical analysis as these shipments do progress to downstream locations during actual operations (after manual handling), and stopping their forward movement in the analysis process tends to understate railcar distance and handlings at yards.
- *Alternate Destinations*: In some cases the destination of a shipment can change depending on how it is routed. Four common examples of this are:

   i. Situations where a shipment is placed into "constructive placement" due to the inability of a receiver to accept the shipment.
   ii. Specification of en-route stop offs, where a shipment must go "via" a specific point for partial loading or unloading, or for actions such as cleaning or en-route weighing.
   iii. Substitution of alternate destinations relative to the "billing destination" found on the waybill.
   iv. Cases where a railroad has the option of delivering a shipment to an alternate interchange point to another railroad based on operational convenience.

   Each of these cases must be handled using special logic. The first three cases are the simplest. In cases (i) and (iii), the system typically has a way of "substituting" an alternate destination based on a table of some variety. This substitution is typically handled as a pre-process to the assignment of the shipment to a block, and thus has little impact on the blocking system design. For case (ii), logic is typically added to use the "via point" as the destination for the shipment until the shipment reaches that point, and then use the final destination thereafter. The last situation (iv) is the hardest, as this represents the possibility of dynamically changing the destination based on the circumstances. In table based systems, there are typically two solutions. One is to provide some kind of look-up table that sets the targeted destination based on the current location of the shipment. As the shipment advances to each location based on the blocking, the look-up table is consulted to see if an alternate destination should be used from that point forward. The second approach in table-based systems is to simply drive the shipment to a particular interchange point using the rules. In this situation, certain blocks are designated as "interchange blocks" and the shipment is treated as being complete whenever it is placed into an interchange block. For algorithmic systems, the standard option is to provide blocking choices to both interchanges, with a low or zero cost "phantom block" between the two interchanges that allows the shipment to reach its officially designated destination. Such phantom blocks can be somewhat challenging to specify and maintain, but appear to produce the desired result based on actual experience.

- *Re-hump Blocks*: in some cases during actual operations, the option to place a shipment into its block may not exist because no capacity is available to create the targeted block at the time the shipment is being processed, or the targeted block exists but is full. When this happens, the railcars are placed in a temporary block that will be switched into the targeted blocks at a later time. Such blocks are called "re-hump blocks" or "buffer blocks." While an important real-time operational consideration, and often needed when a yard makes more blocks than it has physical tracks, we will not address this issue in this chapter.
- *Cross-yard Blocks*: some yards are in reality compound facilities. For example, one yard complex might have separate yards for each direction, plus a local yard. These yards may not be modeled as a single location, but several separate, co-located facilities. When trains arrive, they often contain primarily shipments for a specific direction, and thus arrive at only one of these facilities. The blocking plan will then need cross-yard blocks to allow shipments to move between these facilities to reach the appropriate out-bound block. In algorithmic systems these cross-yard blocks are often set to have low costs so that such movements do not cause the yard to be avoided.
- *Directional constraints*: as noted in the discussion of the cross-yard blocks above, some yards make blocks primarily on a directional basis. This can pose a challenge during optimization of a blocking plan. If the destinations of blocks at the yard are limited to ones that are consistent with the yard's directional nature, this tends to ensure that only the most appropriate traffic is handled at the yard. To achieve this, the optimization algorithms are typically constrained to only consider the formation of out-bound blocks to appropriate locations. While one could also constrain the in-bound blocks that the yard can accept, this is often not required because the out-bound constraints will naturally limit what traffic will want to move through the yard.

5.8 Blocking Plan Optimization

In Sect. 5.6, we discuss strategies and tools that assist the planner in developing and assessing blocking plans. This section discusses strategies for automatic blocking plan optimization. A major theme in this section is that, at least at this time, there is no technique that will generate a blocking plan that passes all the real-world constraints so that the solution can be used unchanged. However, there are two important uses of blocking optimization that give significant value:

- It provides an excellent starting pointing for developing "zero-based" (or "clean sheet") operating plans.
- When used with good comparison tools such as the triplet analysis and tree-view analysis described in Sect. 5.6, it gives planners suggestions for incremental changes to the current blocking plan.

The main reasons why the current state-of-the-art for automatic blocking plan optimization is not able to develop final, usable plans include:

- The current blocking optimization models are for a single type of block (usually for "manifest" or "merchandise") with no differentiation for types of traffic. For example, the blocking optimization techniques cannot generate a set of blocks for automobile traffic (finished vehicles or parts) plus a set of blocks that allow both auto and manifest.
- Blocking optimization does not do a good job on "local" blocking for two reasons. The first is due to the need for significant use of rules to move the car the last mile. The second is that the yard capacity constraints are more complicated for local blocks since local traffic may be moved less than daily, or it could be that several local blocks might occupy the same track and be switched just before delivery.
- Blocking optimization by itself does not make strategic changes or trade-offs. For example, the railway might decide to change the function of a yard—say eliminate the hump, or change a flat yard to focus strictly on automotive traffic.
- A complete operating plan specifies the blocking plan, the train plan, and the assignment of blocks to trains. Often when developing the train plan, adjustments need to be made to the blocking plan to reduce block swaps, circuity of the blocks or changes to ensure trains are sufficiently filled out with cars.

5.8.1 Considerations That Automated Blocking Optimization Techniques Should Consider

So far, the main characteristics we have discussed for designing a blocking plan are:

1. Find a plan that allows all traffic to have a block sequence.
2. Minimize the sum of the costs of the block sequence (the cost is a combination of the distance the cars travel and the switching costs expressed as a distance penalty).
3. Limit the number of classifications made in a yard to fit the capacity of the yard.
4. Limit the number of blocks made in the yard to fit its capacity.

However, experience has shown that these criteria alone are not sufficient, and additional constraints need to be added. These include:

1. Progressive block size. Each block is given a minimum block size, and generally the block size increases with the block distance. For example, blocks traveling less than 100 miles may have a minimum block size of 5 cars, while blocks going greater than 500 miles might get a minimum block size of 20 cars. Large long distance blocks may become "anchor" blocks and have a train that carries them from origin to destination, perhaps with some minimal circuity so that it could process some other blocks along the way. Small long distance blocks would not have a train designed around them, and often would have to be carried using one or several block swaps, and hence are not desirable.
2. Directionality of blocks. Some yards, due to the physical track characteristics and presence of other nearby yards, are often constrained to make blocks that go in only a limited number of directions.
3. Local blocking. We already mentioned that block optimization does not work well for local blocks. In our experience, it is best to "roll up" the traffic to serving yards (the second smallest tier in the hierarchy—serving yards generally handle cars for several local yards), taking the local blocking plan design out of the optimization process. Such roll-ups also have the advantage of making the problem more compact, by reducing the number of locations to be considered, the total number of blocks in the plan, and the size of the traffic database. The traffic is reduced because the roll-up process reduces the number of unique origins and destinations for the traffic, allowing similar traffic records to be combined with each other.

5.8.2 Mathematical Representation of the Block Design Optimization Problem

There have been a variety of efforts to develop railroad blocking plan and railway operating plan optimizers dating back over many years (Ahuja et al. 2007; Van Dyke 1986; Van Dyke 1988; Bodin et al. 1980; Barnhart et al. 2000; Newton et al. 1998; Newton 1996; Crainic et al. 1984; Gorman 1995; Keaton 1989, 1992; Yaghini et al. 2012). While a number of these efforts have produced quite good mathematical statements of the problems, computational constraints have limited these formulations usability to solve real world problems. The consequence is that most practical solutions use some form of heuristic that includes subsets of the optimization formulation or other concepts discussed in this chapter.

Given the above qualification, here we state the optimization problem more formally, assuming that algorithmic blocking will be used as the source for obtaining block sequences.

5.8.2.1 Data

A set of yards $Y = \{1,2,\ldots n\}$, where n is the number of yards.

There is an underlying set of links L where $l \in L$ represents a directed link. We represent the tail as $t(l) = y_1$ and the head as $h(l) = y_2$. That is, the link goes from yard y_1 to yard y_2 and represents the physical track between these two yards. Usually there is another link that goes from yard y_2 to y_1. The graph (Y, L) is typically very sparse, with the number of links typically only slightly larger than twice the number of yards. Each link has a distance. We assume that the yards are connected: that for every pair of yards $(y_1, y_2) \in Y \times Y$ there exists a connected path of links $\{l_1, l_2, \ldots l_k\}$ where $t(l_1) = y_1, h(l_k) = y_2$ and $h(l_j) = t(l_{j+1})$ for $j = 1, 2, \ldots k-1$.

The set of all possible blocks is $\mathrm{B} = Y \times Y$, that is, B is all possible arcs between yards in Y. Denote the origin of the block $b \in \mathrm{B}$ as $o(b) \in Y$ and the destination $d(b) \in Y$. Each block $b \in \mathrm{B}$ has a distance $\omega(b)$ that is composed of finding the shortest distance path in the (Y, L) graph from the origin of the block to its destination. Note that one could substitute a "weighted distance" or cost for each link that is not the same as the physical distance in order to reflect "routing preferences" on the (Y, L) graph.

For each yard $y \in Y$, let B_y be the maximum number of blocks that can originate at y, and let C_y be the maximum number of railcars that can be switched at y. Note that this is a significant simplifying assumption in that the maximum number of blocks may be a "soft" number depending on the operating strategy for the yard, the mix of local versus longer distance blocks, and the total number of railcars that is handled at a yard. As noted earlier, we generally exclude the local blocks from the optimization problem, so that the maximum number of blocks would only reflect the longer distance blocks.

Let $M(\omega)$ = the minimum block size allowed for a block with distance ω.

Let T be the set of traffic. Denote the origin of the traffic $t \in T$ as $o(t) \in Y$ and the destination $d(t) \in Y$. The number of cars associated with a traffic record will be $w(t)$. This notation overlaps the notation for the block origin and destination, but it should be clear from the context when we mean block origin or traffic origin (respectively destination). It is generally assumed that this traffic has been "rolled up" to the serving yards, and excludes the local yards. Further, this formulation is a single commodity formulation, and as a result is generally limited to only carload.

5.8.2.2 Variables

The main variable represents the blocking plan. One way to describe it is as a set of binary variables $\delta_b \in \{0,1\}$ where $b \in \mathrm{B}$. If the block b is included in the block plan then $\delta_b = 1$, otherwise $\delta_b = 0$.

We also have the cost of a classification in yard $y \in Y$ as a non-negative variable P_y. This may appear strange to have the classification cost as a variable. It is natural to consider the very important classification cost to be fixed and known prior to the start of the optimization. In practice, this is the case and optimization algorithms typically assume the user has good initial values for the classification cost. However there may be circumstances when the classification cost needs to be adjusted during the course of optimization, and hence it will be considered for now as a variable. One example is when ensuring that the capacity of a yard in terms of the number of railcars being handled is respected in an optimal manner.

The block cost is the sum of the classification cost at the origin of the block and the distance of the block, denoted $c(b) = P_{o(b)} + \omega(b)$. P_y is non-negative, $c(b) \geq 0$. As noted earlier, the distance could use weighting factors to reflect routing preferences.

Given the set of active blocks $\hat{\mathrm{B}} = \{b \in \mathrm{B} \mid \delta_b = 1\}$, let the block sequence for a traffic record *t* be based on using algorithmic blocking; it is the shortest path in the graph $(Y, \hat{\mathrm{B}})$ based on the cost *c*(*b*) and denoted as $S\left(t \mid \hat{\mathrm{B}}, P\right) = \left(b_1, b_2, \ldots, b_{k_t}\right)$. In the block sequence, each $b_i \in \mathrm{B}$, and follows the usual rules for a path: $o(t) = o\left(b_1\right), d(t) = d\left(b_{k_t}\right)$, and $d\left(b_j\right) = o\left(b_{j+1}\right)$ for $j = 1, 2, \ldots, k_t - 1$. The notation is meant to explicitly show that the block sequence is dependent on the active blocks and the classification costs, and that the block sequence is an ordered-tuple and not an unordered set.

The cost of a block sequence, $C\left(t \mid \hat{\mathrm{B}}, P\right)$ is the sum of the costs of its components:

$$C\left(t \mid \hat{\mathrm{B}}, P\right) = \begin{cases} \sum\limits_{b \in S\left(t \mid \hat{\mathrm{B}}, P\right)} c(b) & S\left(t \mid \hat{\mathrm{B}}, P\right) \neq \varnothing \\ \infty & S\left(t \mid \hat{\mathrm{B}}, P\right) = \varnothing \end{cases}$$

Note that we do not have a cost for forming a block at a yard, only a cost for using a specific block sequence. Block formation costs are generally treated as zero, and instead we rely on the overall limit on the maximum number of blocks that can be made at each yard. The total cost of the solution is of course dependent on the volume of railcars using each sequence.

5.8.2.3 Constraints

All traffic must be moved:

$$\forall t \in T, S\left(t \mid \hat{\mathrm{B}}, P\right) \neq \varnothing \tag{5.1}$$

Number of blocks originating from a yard must be constrained:

$$\forall y \in Y, \sum_{(b \in \mathrm{B} \mid o(b) = y)} \delta_b \leq B_y \tag{5.2}$$

To show the number of classifications at a yard, we use the notation that the block sequence for t can be written as $\left(b_1, b_2, \ldots, b_{k_t}\right) = S\left(t \mid \hat{\mathrm{B}}, P\right)$. Using this notation, the constraint for the number of classifications at a yard is:

$$\forall y \in Y, \sum_{t \in \mathrm{T}} \sum_{j=1}^{k_t - 1} w(t) \cdot 1\left\{d\left(b_j\right) = y\right\} \leq C_y \tag{5.3}$$

Every block should have a minimum volume, based on the distance of the block:

$$\forall b \in \mathrm{B}, \sum_{t \in T} \sum_{j=1}^{k_t} w(t) \cdot 1\left\{b_j = b\right\} \geq \delta_b * M\left(\omega(b)\right) \tag{5.4}$$

A number of additional constraints can be introduced, but will not be explored further in this formulation. These include:

- Constraints to support directional activities at a yard, which can be implemented by limiting the set of blocks that can be considered from a specific yard.
- Constraints that require certain blocks to be made, or not made. One can think of this as fixing the integer variables for those specific blocks. An example is the fixing of interchange blocks.
- Constraints on the routing of specific traffic to use specific blocks, essentially fixing part of the path (block sequence) of certain traffic records.

5.8.2.4 Objective

The objective is to minimize total cost over all the traffic records:

$$\min_{\delta_b, P_y} \sum_{t \in T} C\left(t \mid \hat{\mathrm{B}}, P\right) \tag{5.5}$$

As noted earlier, we could introduce weighting factors on the distance costs, and have elected not to include block formation costs. Other formulations have also suggested making use (or non-use) of a yard a factor as well introducing a yard "opening" cost.

Optimization Techniques

There are three levels of techniques used for blocking plan optimization:

1. Automation of the techniques from Sect. 5.6—especially bypass opportunities and low volume block elimination.
2. Additional heuristics.
3. Advanced mathematical programming techniques.

Heuristic Approach

The heuristic approaches find blocking plans by seeking out opportunities to locally improve existing blocking plans by keeping most blocks fixed and only examining a limited number of changes at a time. These approaches rely on several ideas which are explained subsequently:

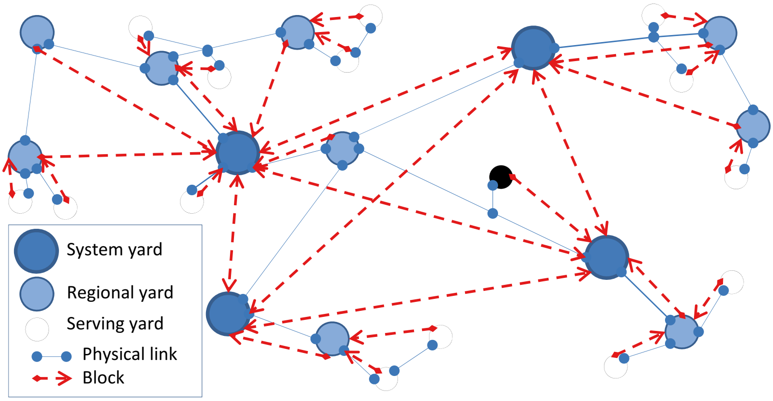
- The ability to create an initial blocking plan.
- Methods for iteratively improving blocking plans.
- The ability to quickly resequence and test out new blocking plans.
- Ability not to get stuck at a local optimum.
- The ability to change the yard penalties if classification capacity constraints cannot be otherwise met.

In many cases they allow for interim solutions that may be infeasible with respect to use of low volume blocks, the number of railcars handled at a specific yard, or the number of blocks formed at a specific yard. These constraints are generally respected in the final solutions, though there can be cases where the requirement that all traffic have a sequence may result in a violation of the low volume block constraint.

Initial Blocking Plan

The most common approach is to build an initial blocking plan based on a hierarchy of yards (or start with the existing plan used by the railroad). Yards can be usually categorized as local, serving, regional or system yards, with the cost per classification decreasing (and the number of classifications per day increasing) for each level of the hierarchy. The concept of the hierarchy is to build a set of bi-directional blocks from each local station to the closest serving yard, from each serving yard to the closest regional or system yard, from each regional yard to the closest system yard, and between all pairs of system yards.

The illustration below is an example where all four system yards have bi-directional blocks between them. Each regional yard has a single block to the closest system yard. Serving yards have a single block to the closest regional yard, with the exception of the serving yard that is in black which has a block to a system yard because that is closer than any regional yard.



One variant of the above is to allow connections from each non-system yard to more than one other yard that is higher in the hierarchy, provided it is within some prescribed distance and you do not need to pass through any other yard that is higher in the hierarchy to reach it. Because of the small number of blocks at other than system yards, this initial solution is usually feasible from a block formation perspective, but there likely will be too many blocks created at some system yards, and sometimes at regional yards as well. Note that the customer locations have been excluded from this process.

Iteratively Improve the Plan

There have been two general strategies for automation of improving an existing plan:

1. Iterative use of bypass opportunities to add potential new blocks, and low-volume analysis to remove blocks (Van Dyke 1986). In this approach, all bypass opportunities are calculated, then one or several of the top opportunities are taken. The bypass opportunities are given a score based on total car-distance, number of classifications, and a cost for violating the two types of yard capacities.
2. Iterative use of rebuilding the blocking plan for a single yard, and iterating through all the yards (Ahuja et al. 2007). This technique is an example of "very large scale neighborhood search." In this approach, for the given yard one block is entered at a time that is deemed the best block based on a score composed of car-distance and classifications. As each block is added, the block sequences are regenerated efficiently.

Both of these approaches rely on algorithmic blocking for determining block sequences. In turn, there is assumed a cost penalty for each classification at a yard, P_y. Often these cost penalties result in too many cars being classified at individual yards and therefore heuristics are used to adjust the cost penalties so that the algorithmic blocking meets the yard capacity.

Resequencing Quickly

Iterative algorithms rely on testing tens of millions—or more—of possible blocking plans. Each test requires evaluating the objective function as described in Eq. (5.5), which involves a full block sequence. Various authors have been reluctant to explain the tricks they developed to resequence quickly, although it is at the heart of the calculations. What is known is that they:

- Use all-pairs shortest path algorithms.
- Are able to restrict the block sequencing to a subset of the traffic at each iteration. One rule is based on logic such as if A–B is a new block, then it will never be used for traffic from B to A so no need to resequence that traffic. More generally, there are many traffic records that should never be classified at a particular yard because doing so would add an unacceptable amount of circuity.

Finding Global Optimum

The iterative techniques discussed above are not proven to be optimal. They stop when no further improvement can be found, but that does not imply optimality—rather it implies that the algorithms are not robust enough to seek better solutions and are stuck in what is known as a local optimum.

There are several methods used to try to move away from local optimum. Two methods, which are often combined, are:

1. Change the constraints (Eqs. 5.1–5.4) into penalties on the objective function. That allows, for example, more blocks than desired to originate from a yard. This may allow iterative algorithms to try out more possibilities then would otherwise be possible.
2. Add some type of randomness into the choice. For example, randomly allow a block into the solution that is economically not very good given all the existing blocks, but later on may prove useful. This is part of the concept of simulated annealing, which has been used in many instances to find better solutions.

There are other techniques that use randomness very successfully in a variety of iterative algorithms that could be applied here. Two popular ones are Tabu Search (Glover and Laguna 1997) and Genetic Algorithms (Simon 2013).

Changing Yard Penalties

Iterative algorithms always have an initial value for the classification cost P_y as described earlier. However, the cost may be too high to allow enough classifications at the yard, or too low, causing the yard to be overwhelmed. The model may not be able to build as many blocks as would be desirable at the yard because the limit on the number of blocks B_y is met well before the limit on the number of classifications C_y is met.

In these cases, the iterative algorithms need to set a trigger that, when over a number of iterations a yard is far away from the limits set, to adjust the penalties. While there is no precise methodology, it is occasionally necessary to make these adjustments to obtain an optimal block design. Typically this means treating the yard capacity constraints as soft (at least for the constraint on the number of railcars), because the violation of these constraints provides important information on how much to adjust the costs or penalties for using the yard.

Advanced Mathematical Programming

Bodin et al. (1980) were the first to produce a mathematical model to create blocking plans, followed by Newton (1996) and Newton et al. (1998). This was followed by Barnhart et al. (2000) that took the work a major step forward to solve blocking optimization problems of significant size. Their formulation found a near-optimal solution to a somewhat simplified version of the problem. They considered the main constraints—all traffic must obtain a block sequence (Eq. 5.1), the number of blocks (Eq. 5.2) and the number of cars classified at yard are limited (Eq. 5.3). They formulate the problem as a network-design integer program and use advanced mathematical programming techniques including Lagrangian decomposition, column generation, valid inequalities, and dual-ascent to solve the problem.

They start with a large number of potential blocks, and for each traffic flow they find a block sequence within those potential blocks, such that all the block sequences taken together meet the two yard constraints and provide minimal total block sequence cost.

This approach has several issues, however, that need more investigation before it can be used solve real-world blocking problems:

- A necessary constraint for developing a realistic blocking plan is that each block should have a minimum block size (Eq. 5.4). This constraint is not found in their model and while it could be easily placed in their model, it will significantly complicate their Lagrangian decomposition approach.
- The block sequences found may not achieve the routing consistency produced by algorithmic or simple table-based rules because it finds a block sequence for each traffic flow that is governed by capacitation limits on yards. In their case, each traffic flow has a different origin/destination combination. One traffic flow may have a block sequence A–B–C–D–E, another may have a sequence A–B–F–D–G. The inner sequence for the first flow is B–C–D, but it is different (B–F–D) for the second flow because the switching capacity at yard C is met by the first flow, so it needed to alternatively route the second flow through F. Algorithmic blocking in this case will not generally allow two different inner sequences. It is possible to use table-based rules to achieve this outcome, but there will be inconsistencies in the tables—what is the block sequence for traffic from B to D? Is it through C or through F? This may not be a significant issue in practice, but needs to be examined.

  It is possible in their solution to also send half a shipment from B to D via C, and half via F, which also violates using algorithmic or table-based blocking plan designs.
- The authors claim that one part of their approach uses a simplified objective function of only minimizing classifications, and not the total cost of a block sequence. They use this special objective to speed up part of their algorithm. This objective function most likely produces additional circuity.

Despite these issues, we strongly encourage researchers to continue the efforts of using advanced mathematical programming techniques for solving the blocking design problem.

5.9 Additional Considerations

In thinking about the issues related to blocking plan design, optimization, and shipment routing, there are a number of other issues to be considered:

- *Planning Versus Execution Systems*: planning systems and execution systems have different objectives and needs. Real-time systems generally treat the blocking plan as static. The core question they seek to answer is "given that a shipment is at location X, what block should it be assigned to out-bound from X." To answer this question, the system will either use a rule-based look-up process, or an algorithmic routing process. In the planning environment, the goal is more complex. In plan maintenance mode, the systems must support creation, testing, and maintenance of the blocking plan to support the execution systems. To do this, the planners need access to "what if" capabilities, ways to identify possible plan problems and possible plan improvements, tests for plan completeness, and projections of workloads. In addition, planners are likely to periodically take a deeper look at the blocking plan, and seek ways to identify potential broader plan improvements through use of optimization or other improvement techniques.
- *Traditional Problem Separation of Blocks Versus Trains*: At present, most optimization and design strategies separate the blocking plan from the train plan, or approach the problem in an iterative manner. Under this approach, the blocking plan is designed first. The train plan is then created based on the blocking plan. As part of the train design process, issues with the blocking plan may be identified, and used to see if the blocking plan can be improved to yield a better overall solution when the trains are taken into account. This separation is done for two reasons. First, it is dictated in part by the complexity of the problem and the associated difficulties in solving the joint problem. Second, the blocking plan design remains a largely manual process. Even with the use of optimization, the optimizers are only used as a source of ideas or suggestions for plan design, and the final plan usually represents a process of manual review of the optimization results and the selective adoption of the best ideas from the optimization into the final plan. The consequence of this is that wholesale optimization of the blocking plan on a joint basis would be unlikely to produce a result that would be used in the real word, and might be too complex to support manual review. To the extent that joint optimization is possible, this is generally limited to allowing the system to change only a limited number of blocks, both to ensure that the core blocking plan is protected in the optimization process and to make manual review simpler. Such joint optimization strategies are explored in more detail in Chap. 1 on train schedule design.
- *Location-based Routing Control Versus Shipment-based Control*: Under the blocking systems described above, when a railcar is moved, it does not own its own routing plan. Instead, at each location the shipment visits, tables and other systems are examined, and based on the content of these tables, the next location

for the shipment is determined. Thus, the routing plan is "location centric" and not "shipment centric." This had significant advantages in an environment with limited communications, and no fully defined, centralized, computerized operating plan. Each location could have a "blocking book" or "routing guide" and know what to do with each shipment without having to consult with a central authority. Even today, this approach has advantages when shipments are misrouted, or fail to connect to their expected train, because it supports a straight forward way to determine what to do with the shipments. Going forward, it may become more common to instead take a broader network view of the routing process, and then tie the resulting routing to the shipment. When a shipment is processed at a yard, it would then be assigned to a block (or train) not based on local routing instructions, but based on the routing instructions owned by the shipment. A fallback solution will still be required when a shipment falls off its planned routing. This has a number of advantages, including the ability to support reservation type systems, customize routings for individual shipments/customers, and provide a foundation for supporting a dynamic, capacitated routing process.

5.10 Opportunities

Hopefully the reader has gained an understanding of the blocking problem, and the strengths and weaknesses of current approaches to the problem from this chapter, as well as an understanding of where future research and development is needed. While there are many facets to the problem, the authors would like to point out some specific areas for future research below:

(a) Classification table generation problem: as noted extensively in this chapter, most production blocking systems use tables to direct the classification of shipments. Most optimization tools and efficient block sequencing tools use non-table-based algorithms. The reliable translation of these algorithmic routings to table-based solutions that are maintainable and acceptable to railroad planners remains a major challenge that is largely unmet. The authors participated in one such effort that produced a mathematically perfect translation, but was not acceptable to the railroad due to the complexity of the rules that were produced. This complexity resulted in an increase in the total number of rules, made the rules difficult to maintain on a manual basis going forward, were difficult for the planners to understand, and were too different from the historic rules to be acceptable to the planners.

(b) Multi-commodity optimization: most current optimization strategies are single commodity, and cannot take into account the differing needs of each line of business served by the railroad, and the cross-over effects of some traffic operating in dedicated, specialized services and some traffic "falling into" the general carload network. Planning for the movement of grain traffic, which can move in both dedicated trains and in the carload network provides a prime example of this problem.

(c) Joint train/blocking plan problem: trains can be viewed as serving the purpose of moving the blocks in the blocking plan. However, if the blocks cannot be efficiently bundled into trains of reasonable size and complexity, the blocking plan itself can prove to be impracticable. As a result planners typically follow an iterative process where issues in the design of the trains may cause them to make changes to the underlying blocks. While some solutions for train design are capable of suggesting limited changes to the blocking plan, we ultimately would like to see solutions that are of a more integrated nature.

(d) Reservation/capacity management concepts: at present the authors are only aware of one or two railroads on a world-wide basis that use a train level reservation approach to the movement of shipments. Such an approach has the potential to support advanced capacity management concepts that might be able to produce lower cost solutions, improved service reliability, and better overall network management. These concepts are explored in Chap. 4 on car scheduling and simulation.

(e) Local service design: we have repeatedly pointed out that the local service design problem is generally handled manually, and on a separate basis from the more system level blocking plan problem. Tools and techniques for improving the local plan would be very beneficial, particularly given the large percentage of total trip costs associated with local service.

References

Ahuja RK, Jha KC, Liu J (2007) Solving real-life railroad blocking problems. Interfaces 37: 404–419

Barnhart C, Jin HH, Vance P (2000) Railway blocking: a network design application. Oper Res 48(2):1–12

Bodin LD, Golden BL, Schuster AD, Romig W (1980) A model for the blocking of trains. Transport Res 14B:115–120

Crainic TG, Ferland JA, Rousseau JM (1984) A tactical planning model for rail freight transportation. Transport Sci 18:165–184

Glover F, Laguna M (1997) Tabu search. Kluwer Academic Publishers, Norwell, MA

Gorman MF (1995) An application of genetic and Tabu searches to the freight railroad operation plan problem, INFORMS spring meeting

IBM. The optimization of global railways, http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/smarterrail/

Keaton MH (1989) Designing optimal railroad operating plans: Lagrangian relaxation and heuristic approaches. Transport Res 23B:363–374

Keaton MH (1992) Designing railroad operating plans: a dual adjustment method for implementing Lagrangian relaxation. Transport Res 26A:263–279

Kraft ER (1998) A reservations-based railway network operations management system, Ph.D. Dissertation, Department of Systems Engineering, University of Pennsylvania, Philadelphia, PA, UMI Order # 9829930

Kwon OK, Martland CD, Sussman JM, Little PD (1995) Origin-to-destination trip times and reliability of rail freight services in North American railroads, Transportation Research Record, No. 1489, pp 1–8

Little PD, Kwon OH, Martland CD (1992) An assessment of trip times and reliability of boxcar traffic, proceedings of the transportation research forum, 34th annual meeting, vol 1, Arlington, VA

Newton HN (1996) Network design under budget constraints with application to the railroad blocking problem, Ph.D. Dissertation, Auburn University, USA

Newton HN, Barnhart C, Vance P (1998) Constructing railway blocking plans to minimize handling costs. Transport Sci 32(4):330–345

Norfolk Southern Corporation. Next generation car routing system at Norfolk Southern. http://www.informs.org/content/download/239255/2274025/file/SC1.pdf

Railway Age (2014) Obituary for guerdon sterling sines, 1928–2014, railroad computer systems pioneer, railway age, 5 Sept 2014. http://www.railwayage.com/index.php/news/guerdon-sterling-sines-1928-2014-railroad-computer-systems-pioneer.html

Simon D (2013) Evolutionary optimization algorithms. Wiley, Hoboken

Van Dyke CD (1986) The automated blocking model: a practical approach to freight railroad blocking plan development. Transport Res Forum 27:116–122

Van Dyke CD (1988) Dynamic management of railroad blocking plans. Transport Res Forum 29:149–152

Yaghini M, Seyedabadi M, Khoshraftar MM (2012) A population-based algorithm for the railroad blocking problem. J Ind Eng Int, SpringerOpen, 8:8 doi:10.1186/2251-712X-8-8

Chapter 6
Crew Scheduling Problem

Balachandran Vaidyanathan and Ravindra K. Ahuja

6.1 Introduction

The crew scheduling problem (CSP) involves assigning crew to trains, while satisfying a variety of Federal Railway Administration (FRA) regulations and trade-union work rules. Train crew work together to move a train from its origin to its destination. As the train travels over its route, it goes through numerous crew districts. In each crew district, the train is manned by an engineer and a conductor who are qualified to operate the train within that district. The objectives of crew scheduling are therefore to assign crew to the trains, while minimizing the cost of operating trains, improving crew quality of life, and satisfying all FRA regulations and work rules.

The crew scheduling problem is a difficult problem to solve because the deployment of crew on trains is governed by many regulations. Crews cannot be assigned outside their crew districts and they need to have minimum rest between assignments. Each crew has a home location and an away location, and there are rules that govern how often a crew must return to its home location. If a crew is detained at an away location for more than certain duration, the railroad needs to pay detention costs. Further, crew need to be assigned to trains in a First-In-First-Out (FIFO) manner. Also, the number of incoming trains and outgoing trains may be imbalanced, which may necessitate crew deadheading on trains or repositioning via taxi so that they may be available to work at a different location. All these constraints and decisions make the problem hard to solve.

B. Vaidyanathan (✉)
FedEx Corporation, 1000 Ridgeway Loop Road, Suite 500, Memphis, TN 38120-4045, USA
e-mail: bala.vaidyanathan@gmail.com

R.K. Ahuja
Optym, 2153 SE Hawthorne Boulevard, Gainesville, FL 32641, USA
e-mail: ravindra.ahuja@optym.com

© Springer Science+Business Media New York 2015

B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3_6

Several researchers have worked on airline and passenger rail crew scheduling (for example, Barnhart et al. 1994, 2003; Caprara et al. 1997; Chu and Chan 1998; Freling et al. 2004). Most of the railroad crew scheduling literature is related to European and Asian railroads; these settings do not have the FIFO requirements and are therefore very different from that in North America. The two articles that have been written specific to the North American railroad crew scheduling are due to Gorman and Sarrafzadeh (2000) and Vaidyanathan et al. (2007). Gorman and Sarrafzadeh (2000) used dynamic programming to solve CSPs where the districts are single-ended (all crew have the same home location); single-ended districts are the simplest crew district configuration. Vaidyanathan et al. (2007) developed a crew scheduling model that works for double-ended and other complicated crew district configurations; their work reports the most comprehensive crew scheduling model to date. Hence, the mathematical model and the solution approach described in this chapter are based on Vaidyanathan et al. (2007), though the rest of the paper deals with crew scheduling in general.

6.2 Background on Crew Scheduling

This section gives an overview of the CSP and defines some of the terminology needed to understand the problem. It also gives an overview of some of the typical regulations which govern crew management.

6.2.1 Terminology

Crew District: The railroad's network is divided into numerous *crew districts*; a crew district constitutes a subset of terminals. Each crew district is a geographic corridor over which trains can travel with one crew. A typical network for a major railroad in the U.S. is divided into as many as 200–300 crew districts. As a train follows its route, it goes from one crew district to another, picking up and dropping off crew at *crew change terminals*.

Crew Pools: Within a crew district, there are several types of crew called *crew pools* or *crew types*, which may be governed by different trade-union rules and regulations. For example, a crew pool may have preference over the trains operated in a pre-specified time window. In some cases, a crew pool consisting of senior crew personnel is assigned only to pre-designated trains so that crews in that pool know their working hours ahead of time.

Home and Away Terminals: The terminals where crews from a crew pool change trains are designated as either *home terminals* or *away terminals*. The railroad does not incur any lodging cost when a crew is at its home terminal. However, the railroad has to make arrangements for crew accommodation at their away terminals. A crew district with one home terminal and one away terminal is called a *single-ended crew district*. The other type of crew district is a *double-ended crew district*, in which more than one terminal is a home terminal for different crew pools. Some of the other crew district configurations are crew districts with one home terminal and several away terminals, and crew districts with several home terminals and corresponding sets of away terminals.

Crew Detention: Once a crew reaches its away terminal and rests for the prescribed hours, the crew is ready to head back to its home terminal. However, if there is no train, then the crew may have to wait in a hotel. According to the trade-union rules, once a crew is at the away terminal for more than a pre-specified number of hours (generally 16 h), the crew earns wages (called *detention costs*) without being on duty.

Crew Deadheading: This refers to the repositioning of crew between terminals. A crew normally operates a train from its home terminal to an away terminal, rests for a designated time, and then operates another train back to its home terminal. Sometimes, at the away terminal, there is no return train projected for some time, or there is a shortage of crews at another terminal. Thus, instead of waiting for train assignment at its current terminal, the crew can take a taxicab or a train (as a passenger) and deadhead to the home terminal. Similarly, the crew may also deadhead from a home terminal to an away terminal in order to rebalance and better match the train demand patterns and avoid train delays.

On-duty and Tie-up Time: When a crew is assigned to a train, it performs some tasks to prepare the train for departure, and hence crews are called on-duty before train departure time. The time at which the crew has to report for duty is called the *on-duty time.* Similarly, a crew performs some tasks after the arrival of the train at its destination, and hence crews are released from duty after the train arrival. The time at which the crew is released from duty is called *tie-up time*. The duty duration before train departure is referred to as *duty-before-departure* and the duty duration after train arrival as *duty-after-arrival.* Hence, the total duty time (or *duty period*) of a crew assigned to a train is the sum of the *duty-before-departure*, the *duty-after-arrival*, and the travel time of the train.

Duty Period: In most cases, duty period of a crew assigned to a train is the total duration between the *on-duty time* and the *tie-up time*. In some cases when a crew rests for a very short time at an away location before getting assigned to a train, the rest time and the duration of the second train may also be included in the duty period of the crew.

Dead Crews: By federal law, a train crew can only be on duty for a maximum of 12 consecutive hours, at which time the crew must cease all work and it becomes *dead or dog-lawed.*

Train Delays: When a train reaches a crew change location and there is no available crew qualified to operate this train, the train must be delayed. Train delays due to crew unavailability are quite common among railroads. These delays are very expensive and can be reduced significantly through better crew and train scheduling.

6.2.2 Regulatory and Contractual Requirements

Assignment of crews to trains is governed by a variety of Federal Railway Administration (FRA) regulations and trade-union rules. The regulations vary from district to district and from crew pool to crew pool. Some examples are listed below:

- Duty period of a crew cannot exceed 12 h.
- When a crew is released from duty at the home terminal or has been deadheaded to the home terminal, they can resume duty only after 12 h of rest (10 h rest followed by 2 h call period) if duty period is greater than 10 h, and after 10 h of rest (8 h rest followed by 2 h call period) if duty period is less than or equal to 10 h.
- When a crew is released from duty at the away terminal, they can typically resume duty only after 8 h rest.
- Crews belonging to certain pools must be assigned to trains in a FIFO order.
- A train can only be operated by crews belonging to pre-specified pools.
- Every train must be operated by a single crew.
- Crews are guaranteed a certain minimum pay per month regardless of how much they work.

Figure 6.1 gives an example of the decision process that needs to be followed by railroad crew planners.

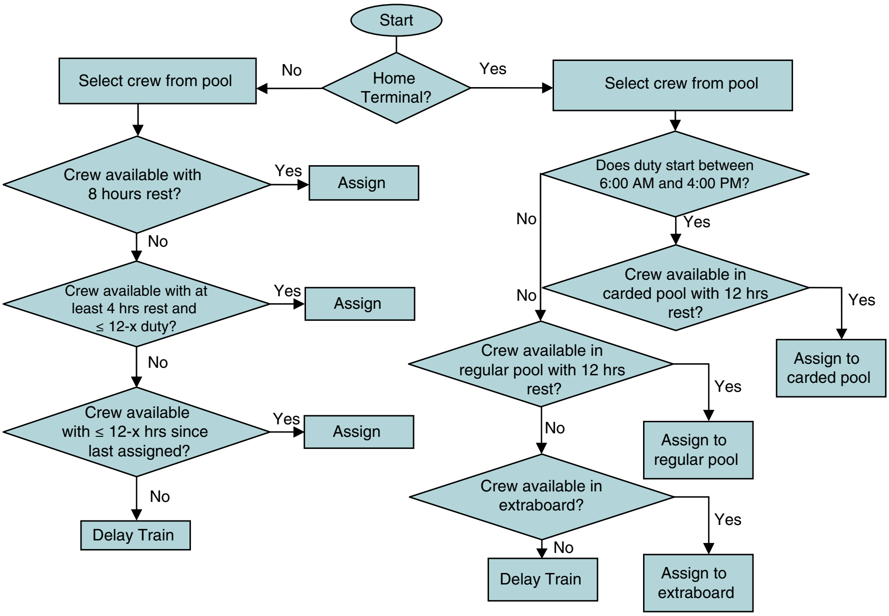

Start
Home Terminal?
No
Yes
Select crew from pool
Select crew from pool
Crew available with 8 hours rest?
Yes
Assign
No
Crew available with at least 4 hrs rest and ≤ 12-x duty?
Yes
Assign
No
Crew available with ≤ 12-x hrs since last assigned?
Yes
Assign
No
Delay Train
Does duty start between 6:00 AM and 4:00 PM?
No
Yes
Crew available in carded pool with 12 hrs rest?
No
Yes
Assign to carded pool
Crew available in regular pool with 12 hrs rest?
Yes
Assign to regular pool
No
Crew available in extraboard?
Yes
No
Delay Train
Assign to extraboard

Fig. 6.1 An example of crew scheduling decision tree

6.3 Mathematical Models for Crew Scheduling

We now describe the mathematical formulation of the crew scheduling problem. Since crews do not work outside their crew districts, this means that the problem can be solved as an independent problem for each crew district. We first describe the inputs that are required to define the problem. Then, we describe the network that is used to model the problem. Finally, we describe the mathematical formulation and solution approaches.

6.3.1 Model Inputs

The inputs that go into the mathematical formulation of the crew scheduling problem are:

- Train Schedule: The train schedule provides information about the departure time, arrival time, on-duty time, tie-up time, departure location, and arrival location for every train in each crew district it passes through.
- Crew Pool Attributes: This includes the home location, the away locations, minimum rest time, and train preferences for each crew pool.
- Crew Initial Position: This provides the position of each crew at the beginning of the planning horizon, and includes the terminal at which a crew is released from duty, the time of release, the number of hours of duty done in the previous assignment, and the crew pool of the crew.
- Train-Pool Preferences: The train-pool preferences specify the set of trains that can be operated by a crew pool.
- Away Terminal Attributes: This includes the rest rules and detention rules for each crew pool at each away terminal.
- Deadhead Attributes: This specifies the travel time by taxi between two terminals in a crew district.
- Cost parameters: Cost parameters are used to set up the objective function. They consist of crew wage per hour, deadhead cost per hour, detention cost per hour, and train delay cost per hour.

6.3.2 Space–Time Network Construction

The CSP is solved as a separate problem for each crew district. The schedule of crew is modeled as the flow of commodities on a space–time network (refer to Ahuja et al. (1993) for more about networks). Each node in the network corresponds to a crew event and has two defining attributes: location and time. The events that are modeled while constructing the network are departure of trains, arrival of trains, departure of deadheads, arrival of deadheads, initial positions and availability of

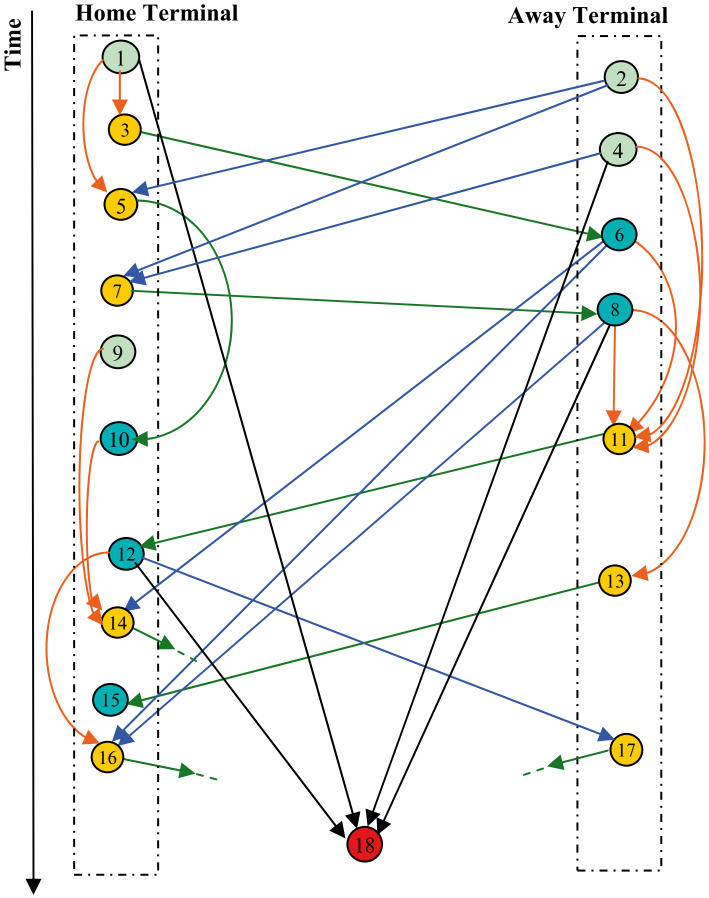


Fig. 6.2 Space–time network for a single-ended district with a single crew type. *Node legend*: *green* (supply), *blue* (arrival), *yellow* (departure), *red* (demand). *Arc legend*: *green* (train), *orange* (rest), *blue* (deadhead), *black* (demand)

crew, and end of the planning horizon. Figure 6.2 presents an example of the space–time network in a crew district (for the sake of clarity, this network only represents a subset of all the arcs).

For each crew, a supply node whose time corresponds to the time at which this crew is available for assignment, and whose location corresponds to the terminal from which the crew is released for duty is created. Each supply node is assigned a supply of one unit and corresponds to a crew. The network also has a common sink node for all crews at the end of the planning horizon. This sink has no location attribute and has the time attribute equal to the end of the planning horizon. The sink node has a demand equal to the total number of crew in the district.

For each train l passing through a crew district, a *departure node*, l', is created at the first crew change terminal and an arrival node, l'', is created at the last crew change terminal in the crew district. Each arrival or departure node has two attributes: place and time. For example, *place* (l')=*departure-station* (l) and *time* (l')=*on-duty-time* (l); and similarly, *place* (l'')=*arrival-station* (l) and *time* (l'')=*tie-up-time (l).*

Train arc (l', l'') is created for each train l connecting the departure node and arrival node of train *l*. *Deadhead arcs* are constructed to model the travel of crew by taxi. A deadhead arc is constructed between a train arrival or crew supply node at a location and a train departure node at another location. All the deadhead arcs which satisfy the contractual rules and regulations are created. *Rest arcs* are constructed to model resting of a crew at a location. A rest arc is constructed between a train arrival node or a crew supply node at a location and a train departure node at the same location. Rest arcs are created in conformance to the contractual rules and regulations. All rest arcs which satisfy the contractual rules and regulations are constructed. Since the contractual regulations are often crew pool specific, deadhead arcs and rest arcs are created specific to a crew pool. Finally, *demand arcs* are created from all train arrival nodes and crew supply nodes to the sink node. Each arc in the network has an associated cost equivalent to the crew wages, deadhead costs, or detention costs, as the case might be. All contractual requirements other than the FIFO constraint are easily handled in the network construction.

So far, the network does not model the scenario when qualified crews are not available for assignment to a train, which causes train delays. Train delays are modeled by the construction of additional arcs. To do this rest arcs and deadhead arcs which do not honor the rest regulations are also constructed and flows on these arcs are penalized to ensure that flows on these arcs occur only when qualified crews are not available for assignment. If the solution contains nonzero flows on these arcs, it implies that the associated train will be delayed until crew becomes qualified for train operation. Since the delay of a train could have propagating effect in the availability of crews in subsequent assignments, it is assumed that the crew assigned to a delayed train has sufficient slack in the rest time at the train arrival node to make it qualified for subsequent assignments.

6.3.3 Mathematical Formulation

The CSP is formulated as an integer multi-commodity flow problem on the space–time network described in the previous section. Each crew pool represents a commodity. Crews enter the system at crew supply nodes, travels on a sequence of connected train, rest, and deadhead arcs before finally reaching the sink node (Table 6.1).

Decision Variables

x_c^l: Flow of crew pool $c \in C$: On each train arc $l \in L$.

x_d: Flow on deadhead arc $d \in D$.

x_r: Flow on rest arc $r \in R$.

Table 6.1 Notation

N	Set of nodes in the space–time network	i_c^+	Set of outgoing arcs specific to crew pool c at node i
L	Set of train arcs in the network, indexed by l	i_c^-	Set of incoming arcs specific to crew pool c at node i
D	Set of deadhead arcs in the network, indexed by d	Ar	Set of arcs on which flow will violate FIFO constraint if there is flow on rest arc r
R	Set of rest arcs in the network, indexed by r	f	Total number of available crew
A	Set of arcs in the space–time network, indexed by a	M	A very large number
G (N, A)	Space–time network	c_c^l	Cost of crew wages for crew pool $c \in C$ on train arc $l \in L$
N_s	Set of crew supply nodes	c_d	Cost of deadhead arc $d \in D$
N_d	Sink node	c_r	Cost of rest arc $r \in R$
C	Set of crew pools in the system, indexed by c	tail(l)	The node from which arc l originates
i^+	Set of outgoing arcs at node i	head(l)	The node at which arc l terminates
i^-	Set of incoming arcs at node i		

Objective Function

$$\text{Min} \sum_{l\in L}\sum_{c\in C} c_l^c x_l^c + \sum_{d\in D} c_d x_d + \sum_{r\in R} c_r x_r$$

Constraints

$$\sum_{c\in C} x_l^c = 1, \quad \text{for all}\, l \in L \tag{6.1}$$

$$\sum_{a\in i^+} x_a = 1, \quad \text{for all}\, i \in N_s \tag{6.2}$$

$$\sum_{a\in N_d^-} x_a = f \tag{6.3}$$

$$x_l^c = \sum_{a\in \text{tail}(l)_c^-} x_a, \quad \text{for all } l \in L, c \in C \tag{6.4}$$

$$x_l^c = \sum_{a\in \text{head}(l)_c^+} x_a, \quad \text{for all } l \in L, c \in C \tag{6.5}$$

$$\sum_{r'\in A_r} x_{r'} - M\left(1 - x_r\right) \le 0, \quad \text{for all } r \in R \tag{6.6}$$

$$x_l^c \in \{0,1\} \text{ and integer}, \quad \text{for all } l \in L, c \in C \tag{6.7}$$

$$x_d \in \{0,1\} \text{ and integer}, \quad \text{for all } d \in D \tag{6.8}$$

$$x_r \in \{0,1\} \text{ and integer}, \quad \text{for all } r \in R \tag{6.9}$$

Constraint (6.1) is the train cover constraint, which ensures that every train is assigned a qualified crew to operate it. Constraint (6.2) ensures flow balance at a crew supply node. Constraint (6.3) ensures the flow balance at the sink node. Constraints (6.4) and (6.5), respectively, ensure flow balance at train departure and arrival nodes. Constraint (6.6) ensures that the crew assignment honors the FIFO constraint. Constraints (6.7)–(6.9) specify that all the decision variables in the model are binary. The objective function is constructed to minimize the total cost of crew wages, deadheading, detentions, and train delays. Note that the detention and delay costs are taken into account while calculating the cost of rest arcs.

Most crew districts have two terminals, and a typical train schedule has around 500 trains running in 2 weeks in a crew district. Each crew district could have two to four crew types and around 50 crews. Therefore, the space–time network could have around $50+2\times500=1{,}050$ nodes. The number of deadhead arcs is typically around 25,000, and the number of rest arcs is around 100,000.

Since the number of rest arcs for a typical problem is of the order of 100,000, and as each rest arc has one FIFO constraint, the number of FIFO constraints in the model is around 100,000, which is very large. Also, these constraints spoil the structure of the problem and a direct approach using commercial solvers to solve the CSP suffers from intractability and does not converge to a feasible solution even after several hours of computation. However, the integer programming problem with FIFO constraints relaxed (*Relaxed Problem*) can be solved to optimality within minutes. In the next section, we describe efficient methods to solve the CSP.

6.3.4 Solution Methods

6.3.4.1 Successive Constraint Generation (SCG)

The SCG algorithm is very simple. The algorithm works by iteratively pruning crew assignments which violate the FIFO constraints from the current solution of a more relaxed problem. First, the relaxed CSP without any FIFO constraints is solved. Then, the algorithm checks for violations of the FIFO constraint. If there are no violations, then the optimal solution to the CSP has been determined, and the algorithm terminates. If there are FIFO violations, the algorithm adds the violated constraints and resolves the problem. This procedure is repeated until an optimal solution that does not violate the FIFO constraints is found.

6.3.4.2 Quadratic Cost-Perturbation (QCP) Algorithm

While the SCG is an exact algorithm, the running time of this algorithm could be quite high. The cost perturbation-based algorithm described in this section is a heuristic but works extremely well in practice. This algorithm penalizes FIFO violations, so that the FIFO constraints do not need to be explicitly considered while

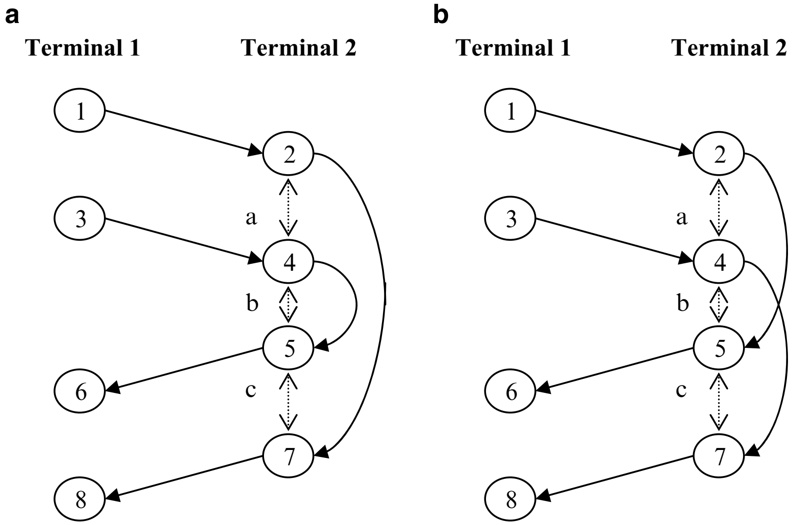


Fig. 6.3 Illustrating the FIFO assignments. (**a**) Invalid assignment. (**b**) Valid assignment

solving the problem. In other words, the costs of arcs are perturbed by a small amount so that the solution to the relaxed CSP is automatically FIFO compliant.

The cost perturbation strategy is presented through the illustration shown in Fig. 6.3 for the case when there is only one crew pool type. In case (a), crew assignments are made in a non-FIFO manner, and in case (b), the assignments are made in a FIFO manner. Consider the case when crews are detained at the Terminal 2. Then, due to the nature of detention costs, the cost of the assignment (b) would definitely be less than or equal to the cost of assignment (a), and hence the solution to the relaxed CSP would honor FIFO constraints. On the other hand, suppose all the rest arcs had a cost of zero; then both the assignments would have the same cost, and the relaxed CSP would have no cost incentive to choose assignment (b) over assignment (a). Thus, a solution to the relaxed problem may violate the FIFO constraints. In order to provide an incentive to the relaxed CSP to choose case (b) over case (a), the cost assignments on rest arcs are perturbed.

The cost perturbation scheme that is used is a function of the duration of rest arcs. Suppose that the time duration between events corresponding to nodes 2 and 4, 4 and 5, and 5 and 7 are a, b, and c, respectively. Consider a cost assignment which is proportional to the square of the duration of rest arcs. The constant of proportionality is represented by k (k is set to a very small value).

Then, cost of assignment

$$\begin{aligned}(\mathrm{a}) &= k\left(\text{duration arc }(2,7)\right)^2 + k\left(\text{duration arc }(4,5)\right)^2 = k\left(a+b+c\right)^2 + kb^2\\ &= k\left(a^2+2b^2+c^2+2ab+2bc+2ca\right),\end{aligned}$$

and cost of assignment

$$(\mathrm{b}) = k\left(\text{duration arc } (2{,}5)\right)^2 + k\left(\text{duration arc } (4{,}7)\right)^2 = k\left(a+b\right)^2 + k\left(b+c\right)^2$$
$$= k\left(a^2 + 2b^2 + c^2 + 2ab + 2bc\right).$$

The cost of assignments in case (b) is less than that in case (a). Hence, when the rest arcs have zero costs, the quadratic cost perturbation scheme gives FIFO compliant assignments, without having to explicitly add FIFO constraints to the model.

The solution time of QCP is comparable to that of the relaxed CSP. As reported in Vaidyanathan et al. (2007), the QCP method produced solutions with objective function values almost the same as those for the relaxed CSP. This implies that FIFO constraints can be satisfied with little or no impact on the solution cost. Thus, QCP can be used to obtain excellent quality solutions very fast. Due to its attractive running times and solution quality, this method has the potential to be used in both the planning and the real-time environment.

6.4 Applications of the Model

The crew scheduling model has many applications in the tactical, planning, and strategic environments, and some examples are provided in this section.

6.4.1 Tactical Benefits

The model has several benefits in the tactical scheduling environment such as:

- Assignment of crew to trains: The output of the model gives the assignment of crew to trains.
- Recommend which crews to place in hotels and which crews to deadhead home: When a crew arrives at an away terminal, the crew callers have to decide whether the crew should deadhead back home or go to a hotel for rest. The model can be used to mathematically look ahead and evaluate the trade-off between different costs such as crew wages, deadhead cost, detention costs, and rest violation costs.
- Minimize train delays due to shortage of crew: Train delays are potentially very costly because the delay of a train may lead to the unavailability of crew to operate another train in the future and may have a negative domino effect on network-wide operations. By creating several deadhead arcs while constructing the space–time network, the possibility of train delays is reduced.
- Disruption management: The model can be used as a tool to bring back disrupted operations to normalcy. Suppose at some point in time the operations are disrupted. The current state or snapshot of the system gives us the location of each crew and the hours of duty already done. Using this information and the information about the future train schedule, the model can be used to optimally re-assign crew to trains.

6.4.2 Planning Benefits

The essence of the crew planning problem is to determine how many crews should be in each crew pool. Railroads typically solve the pool sizing problem based on historical precedent and rules-of-thumb, through negotiation with the union, and by trial and error. The network flow model can satisfy the need for a structured approach that captures all of the considerations, quantifies the various costs, and recommends the best way to define and staff crew pools. Some of the applications of the model in the planning environment are:

- Develop and evaluate crew schedules: The crew scheduling model can be used to compare the current crew schedule used with the model-generated schedule on the basis of several criteria such as average rest time at the home location, average rest time at the away location, average deadhead time, etc. By suitably changing the model cost parameters, schedules with different characteristics can be obtained.
- Size of crew pools: The crew scheduling model can be used to study the impact of varying the crew pool size on the solution quality. For example, suppose the objective is to minimize the number of crew used. While formulating the problem, large cost incentives can be given to flow on the demand arcs from crew supply nodes to the sink node.

6.4.3 Strategic Benefits

Strategic management involves development of policies and plans and allocating resources to implement these plans. The timeframe of strategic management extends over several months or even years. Strategic crew problems include forecasting future head-count needs and evaluating major policy changes such as negotiating changes to trade-union rules or changing the number and location of crew change points on a network. The model can be used to quickly calibrate efficient frontiers for each crew district and show what number of crews minimizes the sum of train delay costs and crew costs.

Some applications of the CSP in the strategic environment are:

- Determining the number of crew districts and territory of crew districts: The model can be used to re-optimize and test different crew district configurations. For example, suppose crew district 1 operates trains between location A and location B, and crew district 2 operates trains between location B and location C. The model could be used to evaluate the benefit of merging all three stations into a single crew district.
- Effect of changing crew trade-union rules: The crew scheduling problem is a complex optimization problem due to strict trade-union rules related to crew operation. The change of any of these rules will face a lot of resistance from the labor union. At the same time, change of any of these rules has the potential to impact crew costs substantially. Using the crew scheduling model, the impact of changing the trade-union rules on the crew cost can be evaluated.

- Forecasting crew requirement: The model can be used to forecast crew requirement by running it with a very large number of available crew. Since the crew supply is more than what is required, many crews will directly flow from the crew supply to the sink node. The total crew supply minus the number of unused crews will give an idea of the number of crews required based on the forecasted train schedule.

References

Ahuja RK, Magnanti TL, Orlin JB (1993) Network flows: theory, algorithms, and applications. Prentice Hall, NJ

Barnhart C, Johnson EL, Nemhauser GL, Vance PH (2003) Crew scheduling. In: Hall RW (ed) Handbook of transportation science. Kluwer Academic Publisher, Norwell, MA, pp 493–521

Barnhart C, Johnson EL, Anbil R, Hatay L (1994) A column generation technique for the long-haul crew assignment problem. In: Ciriano T, Leachman R (eds) Optimization in industry: volume II. Wiley, England, pp 7–22

Caprara A, Fischetti M, Toth P, Vigo D, Guida PL (1997) Algorithms for railway crew management. Math Program 79:124–141

Chu CK, Chan CH (1998) Crew scheduling of light rail transit in Hong Kong: from modeling to implementation. Comput Oper Res 25:887–894

Freling R, Lentink RM, Wagelmans APM (2004) A decision support system for crew planning in passenger transportation using a flexible branch-and-price algorithm. Ann Oper Res 127:203–222

Gorman MF, Sarrafzadeh M (2000) An application of dynamic programming to crew balancing at Burlington Northern Santa Fe Railway. Int J Serv Tech Manag 1:174–187

Vaidyanathan B, Jha KC, Ahuja RK (2007) Multicommodity network flow approach to the railroad crew-scheduling problem. IBM J Res Dev 51:325–344

Chapter 7
Empty Railcar Distribution

Michael F. Gorman

7.1 Introduction

Each year in North America, approximately 30 million carloads are shipped via rail in “general merchandise” or carload service (AAR 2012). In each case, the railroad must deliver a rail-owned empty railcar (such as a box car, gondola, or hopper depending on the commodity) to the origin of the shipper to begin loading. (This process does not apply to private fleets owned and managed by the shipper, as is common for some car types such as tank cars.) After the loaded railcar is delivered to the shipper’s destination and emptied, the rail car is released back to the railroads’ custody and the cycle begins again. The challenge of repositioning a multitude of rail-owned railcars to various origins is known as the empty railcar distribution problem.

The empty railcar distribution problem is complicated by a number of considerations, including the specificity of the wants and needs of the customer (such as capacity and door height), rail ownership and rent paid to other railroads for use of their cars (known as “foreign car hire”), and the distance and time the empty must move over. Further, orders are received and cars released unpredictably, so the problem is constantly changing throughout the day. Finally, there are a number of “soft” trade-offs such as the desire for timely delivery (not too early and not too late) and customer car preference.

Effective solution of the problem is extremely valuable to the rail industry. First, customer service can be improved. Second, cars spend less time empty and more time loaded. Third, with effective empty railcar distribution fewer cars are needed, and those that are in service travel fewer empty miles, producing lower wear and tear on cars per load handled. Fourth, the train space required for moving empties is reduced, effectively expanding capacity for loaded movements, and saving locomotive fuel.

M.F. Gorman (✉)
Department of Operations Management, University of Dayton School of Business, OH, USA
e-mail: mgorman1@udayton.edu

© Springer Science+Business Media New York 2015

B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3_7

Railroads have reported saving tens of millions of dollars per year and billions of capital avoidance from the implementation of railcar distribution systems (see Gorman et al. 2010, 2011; Narisetty et al. 2008).

These tactical railcar systems have been extensively applied to general merchandise traffic, which accounts for about 20 % of all rail traffic. The systems are described in detail for the most of the remainder of this chapter. Automotive railcar distribution follows similar rules, but is managed differently in U.S. rail, as described at the end of this chapter. Coal and grain typically move in "unit" trains, cycling between origin and destination collectively as a train set. Intermodal railcar distribution has different requirements because shippers do not order or load the railcar, but rather the container. (This problem is discussed in the intermodal chapter of this book.)

7.2 Background on Empty Railcar Distribution

7.2.1 Local Distribution and Shipper Pools

Before centralized information systems became commonplace in the rail industry, railcar supply was managed locally. A pool of cars was managed locally and allocated among local shippers. Decentralized control led to inefficiencies such as hoarding behavior and regional shortages. Rail mergers have created larger and more complicated rail networks, creating the need for more sophisticated methods. Improved information systems have allowed for centralized car tracking information. "Shipper pools" (sets of cars with similar characteristics that were dedicated to a shipper) were still used to manage car supply after centralization, but such constraints on car usage vastly reduced flexibility and did not allow for more efficient assignments.

7.2.2 Rules-Based Transaction Processing Systems

Early equipment distribution systems were rules-based transaction processing systems. As a car was released by a consignee, or an order for equipment placed by a shipper, various criteria (such as car type, dimensions, capacity, and ownership) were checked and a car was assigned to an order. In the case of a car becoming available and no orders for that type of car present in the system, a generic "flow order" was used to get the car moving in the general direction of the demand for such cars. This expert-system style of rules codified the knowledge and heuristics used by car distributors to manage car supply and fill orders. Importantly, it automated a labor-intensive task.

However, these systems were lacking in a number of ways. First, the copious rules had to be managed and changed as very seasonal shipping patterns changed. Effective rules are hard to create, and harder to keep up to date when shipping patterns shift. Second, a heuristic system had rules that worked in general, but often failed to manage the fleet well in specific instances because of the sequence dependency of the execution of such rules. For example, a car might receive a rule-based assignment to an order 500 miles away from the shipper, and subsequently another car could become available only 50 miles away from the shipper, but the first car would remain on the order. This assignment would have been reversed if the near-by car was released first, demonstrating another problem; the execution of the rules was highly sequence dependent. Car distributors could manually override such poor assignments, but because the volume of transactions was high, often such inefficiencies would go unnoticed.

7.2.3 Nonintegrated Optimization Systems

Early attempts at optimization of railcar distribution were not integrated with the transactional systems. (Published examples include Jordan and Turnquist 1983; Turnquist 1986; Turnquist and Markowicz 1990.) Typically, a week's worth of actual and forecasted orders were optimally allocated according to an objective such as minimizing total miles of empty car movements, subject to customer service constraints. The problem was formulated as a transportation problem in which supplies of empty cars are assigned to customer orders, minimizing the distance the cars travel, among other considerations. Such a system showed potential improvements over rules-based systems because of their more global view of the problem.

However, these optimization programs generally did not achieve anticipated benefits. The weekly forecast was, of course, subject to error. Often, the optimal results were out of date long before they could be used, and worse, recommendations could be wrong because of errant forecasts and execution failures. Finally, model results would be implemented in the transactional system, causing a large number of assignments to be manually entered. As a result, nonintegrated optimization was not a successful attempt at introducing optimization to empty car distribution.

7.3 Current Day Integrated Real-Time Optimization Systems

In the late 1990s and early 2000s, railroads began investing in integrated, near-real time optimization systems. CSX Railroad implemented its "Dynamic Car Planning System" (DCP) in 1997; BNSF developed its "Equipment Distribution Optimization" (EDO) system in 2000 (Gorman et al. 2010). The Union Pacific developed its system in 2003 (Narisetty et al. 2008).

7.3.1 Model Inputs

The systems have remarkably similar characteristics; below we describe the common components found in such systems: Car supply, shipper orders, marginal shipping cost factors, and customer preferences.

7.3.1.1 Car Supply: Actual and Predicted

The primary source of car supply is the location, date and time of release of an empty car from a consignee. In some cases, equipment is moved from the consignee's location to storage locations in anticipation of future orders.

Forecasted supply is also often used for predicting future anticipated supply. In some cases, empty equipment is in transit, and its "location" is the next location, date and time the car is planned to be available when the train is at the next yard where the car is switched. In each of these cases, the actual car and its full set of attributes (dimensions, etc.) are used in matching the car to customer orders. Often, cars are interchanged between railroads, and future supply is predicted to be delivered from the other railroad where they meet. In lieu of information shared from the "foreign" railroad, interchange volumes of general equipment types are forecasted based on historical patterns and general equipment attributes.

7.3.1.2 Car Orders: Actual and Predicted

Car orders are placed by shippers at the time they plan a loaded shipment. Railroads request (but do not always receive) sufficient lead time to plan empty economical and on-time deliveries, usually 1 or 2 weeks in advance. Some railroads request more advance time on orders for longer term planning, others supplement actual car orders with statistical forecasts based on historical patterns. Such forecasts are often simply moving averages with some day of week and time of year seasonality. Orders are notoriously hard to predict, so actual car orders are vastly preferred. Often, such orders are aggregated into large geographic regions, and storage yards are used as the center of aggregation. Thus, a car that is assigned to a forecasted load (planned for a storage yard) is superseded by an actual car order.

7.3.1.3 Shipper Preferences

Shipper preferences such as maximum allowable early and late delivery, specific physical car requirements (capacity, door heights, and other attributes), and allowable substitutes are kept to balance customer car needs with the need for some flexibility in meeting orders. Hard requirements act as constraints on allowable equipment assignments to orders; preferences are included as a component of the cost of car-to-order mismatches.

7.3.1.4 Cost Parameters

Railroads consider a number of hard dollar costs for empty car distribution, including empty car mileage, travel time costs (including car rents for use of foreign railroad cars), and car handling costs for switching between trains at yards. Shipper preferences are also used in costing for capturing the soft service costs of empty car assignment. Within allowable car assignments, slightly early, late and mismatched cars are assigned a service cost. All of these costs are applied to the feasible railcar-to-order pairings and combined into a single cost coefficient in a costing module for use in the optimization model.

7.3.1.5 Operational Information

Service times from empty supply locations to shipper origins based on train service helps to identify the feasibility of assignments of empties to orders from a timing perspective.

7.3.2 Model Framework

7.3.2.1 Model Preprocessing

The complexity of train operations is simplified through preprocessing. Allowable matches are found by matching car attributes to customer requirements, and checking service feasibility based on empty availability date, customer order date, and the service time between the two locations. In this way, the complexity of rail movements and operations is reduced to core information needed by the model: where is the car, when will it be available, and how long does it take to get to candidate destinations. Each car is considered for assignment to orders for which it meets the customer service criteria.

7.3.2.2 Model Formulation

Problem preprocessing allows the empty railcar assignment problem to be solved via a transportation problem or transshipment problem formulation. Because the two formulations are similar, only the transportation problem is shown below. A detailed comparison of the two formulations can be found in Gorman et al. (2011).

We define a as the vector of permanent car attributes such as car type and ephemeral attributes such as next available date and available location. We define b as the vector of attributes on a customer order, including specific requirements on car type as above, and other shipper attributes such as location, priority, date equipment required, customer preferences on acceptable substitute equipment, acceptable early or lateness, forecasted or actual order, etc. We let A be the set of cars in the planning period and let B be the set of orders in the planning period. We define S_a, $a \in A$ to be the number of planning period cars with particular attribute vector a, and D_b, $b \in B$ be the number of planning period orders with particular attribute vector b. As described earlier, the set of attributes a and b includes not only physical car attributes and customer requirements (e.g., car type, dimensions, etc.), but also the date and location of each car and order.

We define Φ as the set of allowable pairings (*a*,*b*), with $a \in A$ and $b \in B$ to assign a car with a vector of attributes *a* to an order with a vector of attributes *b* established in preprocessing. Φ limits the number of decision variables, x_{ab}, considered by the model by eliminating parings of cars to orders that are not acceptable. Φ is not only based on the customer's car acceptance profile which defines allowable assignment of a car of attributes *a* to a customer car order with requirements *b*, but also the feasibility of the railroad to deliver the car from its location and available date within an acceptable time window of the customer's desired date to the customer at a given location.

As discussed in the previous section, the hard and soft costs of any allowable assignment in Φ of supply to demand (whether actual cars and orders, or forecasted groups of car types and order types) are established via preprocessing, and are included in the single cost coefficient, c_{ab}.

In order to assure feasibility of any model run regardless of the data, a phantom supply source (*r*) and super sink (*l*) are created prior to optimization. The source and sink capacity are calculated prior to model formulation. The number of supply units at *r*, $R = \sum_{b \in B} D_b$, and the total demand at *k*, $K = \sum_{a \in A} S_a$. Thus, source and sink volumes meet all customer and car attribute requirements (the source node is connected to all demands and the sink node is connected to all supply) so that all supply and demand constraints are met with equality: $R + \sum_{a \in A} S_a = K + \sum_{b \in B} D_b$. By definition, every model run is feasible because if necessary all orders can be met by node *r* and all cars can be sent to node *k*.

The transportation problem formulation is given by the optimization model in Eqs. (7.1)–(7.6).

$$\text{Min} \sum_{ab \in \Phi} c_{ab} x_{ab} + \sum_{a \in A} C_k x_{ak} + \sum_{b \in B} C_r x_{rb} \tag{7.1}$$

Subject to:

$$\sum_{b \in B} x_{ab} + x_{ak} = \mathrm{S}_a \forall a \in A,\ (a,b) \in \Phi \tag{7.2}$$

$$\sum_{a \in A} x_{ab} + x_{rb} = D_b \forall b \in B,\ (a,b) \in \Phi \tag{7.3}$$

$$x_{rk} + \sum_{b \in B} x_{rb} = R \tag{7.4}$$

$$x_{rk} + \sum_{a \in A} x_{ak} = K \tag{7.5}$$

$$x_{ab} \geq 0, \text{ and integer } \forall a \in A, \forall b \in B \tag{7.6}$$

The vectors a and b on empty equipment and customer orders contribute to the cost of each assignment, c_{ab}. To the extent that a customer might accept a car that is not a perfect match or not delivered on the exact want date, the cost coefficient c_{ab} is increased accordingly. The total costs of assignments are minimized through optimal assignments x_{ab}, which is a nonnegative integer variable (7.6).

The flow of cars from each supply node to demand locations or the super sink must equal the supply at each node (7.2), and all customer orders of each type must be met from allowable supply or the super source (7.3). Sizable penalties of using phantom cars or car storage (C_k and C_r) are used to discourage flows directly from source to sink. The cost parameter C_r explicitly captures the cost of not meeting a customer order with the decision x_{rb} to supply the order from a phantom car source, r. Similarly, C_k captures the cost of the decision x_{ak} to not use car and moving it to a super sink location, k. Constraints (7.4) assure that all units of supply at the super source flow to demand nodes or to the sink, and constraints (7.5) assure excess supply and super source cars flow to the sink. In the case of a balanced network, $x_{rk} = R = K$.

7.3.3 Model Output Post Processing

The result of the model run is a set of car to car order assignments. The highest priority assignments are actual cars and orders, but the car order-specific assignments are supplemented by forecasted cars and orders.

In the case of oversupply, the model sends cars to a super sink. Car distributors "flow" cars into regions where they are needed or to storage facilities when the optimization model does not have a use for them. Such a flow generally is a function of a forecast or experience. Cars are flowed to storage yards. If no order is received, the cars remain in storage as supply. In the case of a deficit, car distributors must prioritize orders and ration cars between orders. This delicate balance is based on customer priorities and equitable treatment.

No model is perfect; specific operational complexities not known to the model (such as yard configurations which affect desirability of pulling cars in various locations), or information discrepancies (missing reportings), and the like cause distributors to make revisions to model outputs. Where the car distributors formerly worked with rules to allocate all cars, they now focus only on problematic exceptions. The exceptions are managed in the form of car assignment instruction overrides in an environment similar to the rules-based system described earlier.

7.3.4 Systems Integration

A critical component to the success of equipment distribution systems is a deep integration with operational systems. To overcome challenges of early optimization-based methods, the equipment distribution optimization engines must be deeply integrated with other production systems. Recall early attempts' solutions became "stale" as unexpected events occurred, and manual translation of model solutions into car movement instructions was laborious. Updated information with automated translation make integrated equipment distribution systems both more efficient and effective. See Gorman et al. (2010) for more details.

7.3.4.1 Optimization Engine: Customer Car Order System

The optimization engine receives live car orders (and cancelations) from customers in near real-time to ensure the engine is considering the most up-to-date demand information. This includes both individual car orders, as well as customer order preferences (car types, acceptable earliness, and lateness) described above.

7.3.4.2 Optimization Engine-Transactional Equipment Distribution System

Model results are communicated to the field via movement instructions through tight integration of the rules-based system. Model assignments are translated into car movement instructions consistent with the previously described rules, and the optimization model simply provides assignments through the rules-based system. In fact, in many cases, the rules-based system is still in use as a safety net for unallocated cars as they become available. It is worthy of note that, while the optimization engine may have a network-wide plan for current and future empty cars, only the empty cars that require a decision are acted upon operationally. Thus, the transactional rules-based system provides a means to implement the network solution one car at a time.

7.3.4.3 Transactional Equipment Distribution System: Car Movement Management and Tracking System

Car movement management and tracking systems allow railroads to monitor key operational events on the network that spur management action. Two examples are when a customer releases an empty car after it is unloaded (a "release empty" event) and another is when another railroad sends an empty car back to its owning railroad (an "interchange" event). Both of these events constitute new supply for equipment distribution to assign to customer orders; these are "trigger events" for the transactional equipment distribution system to provide disposition for an empty car. These events are automatically transferred to the equipment distribution system so that the equipment distribution system has current information on car supply. But, this is just one example of the deep integration of the car management and equipment distribution systems.

Once the origin–destination pair of the empty car is established by the model, it is translated to an empty car movement instruction. This origin–destination pair is then transferred to the car movement system for the automated creation of a “trip plan” for the empty. A “trip plan (see Ireland et al. 2004; Ahuja et al. 2007) is generated for the car. The chapter on Car Scheduling in this book provides more specifics on the development of trip plans. Similar to an itinerary in air travel, the trip plan maps the sequence of trains from origin to destination to get the car to destination with appropriate cost and service. This trip plan is a live version of the more static “operational information” (described above) that is used on input to the model for the original model preprocessing to determine the timing feasibility of empty car assignments.

As cars move across the network and, critical “events” are tracked (such as “In yard,” “On train,” etc.) so the progress of the move can be monitored. Generally, if cars move according to their trip plan, they are on time and will meet the timing for the customer’s request. Cars that are in jeopardy of being late can, at a minimum, be managed by exception by equipment managers, or at a minimum, status updates given to the customer. But, more importantly, such event information can be integrated directly into the optimization model to optimize the network as critical events occur.

7.3.4.4 Optimization Model: Operational Systems: Decision Making Process Integration

A critical insight into the trip plan helps drive empty assignment flexibility, improved dynamic decision making, and reduce costs: At yards where cars are sorted between trains and reblocked, there is little or no incremental costs of changing to which block a car goes (Gorman et al. 2011). The car can easily be reassigned at any yard where it is reblocked. As such, the empty car may be considered “available supply” just like a release empty or interchange empty event. Thus, the empty car optimization can consider empty cars on assignment for reassignment simply by treating those empty cars as available supply, and their orders as open orders in need of a car. Any changes that have taken place since the last optimization run (i.e., new car releases, cars break down, orders are made and canceled) can be taken into account, and the entire network reoptimized. This capability allows the static optimization model to incorporate dynamic information.

The optimization model can be solved frequently because of the simple and efficient formulation. In fact, the optimization itself is a small fraction of the time to resolve the network because of the data retrieval and transfer times between systems. Railroads typically reconfigure the network ever 10–30 min so that the optimal results are “fresh.” The reoptimized network also considers assignments that car distributors have “locked” in place (for example, to ensure a delivery of a particular car to a particular customer), and these will not be changed; they are treated as a hard constraint. Through near continuous resolving the network problem, a current best solution is always available. Though future events might modify the optimal solution, the solution and resulting empty car distribution instructions are automatically calculated and generated; obviating two key problems of prior systems by generating a network optimal solution and automating the communication of that plan to operations.

7.3.5 Reported Benefits

These systems are among the most success examples of the application of operations research. Railroads have claimed dramatic benefits of such systems, based on a reduction of empty car miles (7–15 %), improved customer order fulfillment and customer satisfaction, and very high return on investment.

For examples, CSX railroad reports approximately a $50 million per year benefit from their system, and BNSF has estimated $13 million; their systems cost approximately $3 to $5 million. The Union Pacific reports a 35 % return on investment, but does not report dollar amounts.

CSX also estimates that based on higher utilization of its rail car fleet, it has avoided purchasing additional $1.4 billion dollars in railcars to support its base of business.

The U.S. public also benefits from reduced truck traffic from reduced pollution, road congestion and the like; based on the CSX diversion of road traffic to rail, that benefit is approximately $50 million per year.

7.3.6 Other Implementation Considerations

7.3.6.1 User Acceptance

Car distributors must go through a big change in the way their work is conducted when such systems are implemented. Railroads report a number of strategies to improve user acceptance and model adherence. First, railroads spend copious time setting model cost and constraint parameters to improve model solutions, though balancing soft and hard costs and constraints is an ongoing challenge. Finding key modeling advocates who also know the problem domain helps build acceptance. Finally, rail customers can be uncomfortable with switching cars on their orders; only through extensive communications, changes in policy, and improved performance can customers be persuaded.

7.3.6.2 Model Thrashing

One concern facing repeated model optimization runs is possible “thrashing” from model run to model run. Thrashing occurs if model recommendations change regularly between runs, resulting in operational confusion or lack of trust in model results. Railroads limit thrashing in a number of ways. First, they leverage automatic locking or freezing of assignments as the empty car is near the customer (for example, 72 h from delivery). Second, model releases assignment information on a "need to know" basis; that is, though the model may have a number of different possibilities for an empty car, a decision is only communicated when disposition is required at interchange, empty release, or at an intermediate yard. As a result, the flexibility afforded by repeated optimization results in relatively infrequent (5–10 %) decision thrashing. Yet, the changes that are made are of great economic value.

7.3.7 Other Modeling Considerations

7.3.7.1 Endogenizing Stochasticity

One approach to addressing the inherent stochasticity facing this problem is to endogenize it within the optimization methodology. As reported above, railroads have solved a deterministic problem repeatedly as the input data change. An alternative might be to endogenize the stochasticity and solve a stochastic model, as is done in Topaloglu and Powell (2006) using an approximate dynamic programming approach. That approach is reportedly under development at Norfolk Southern. While endogenizing stochasticity has potential, specifying the form of that stochasticity can be problematic, and the complexity of modeling and implementation grows.

7.3.7.2 Including Blocking Costs in Empty Car Assignment

When cars move in collections (known as blocks), the handling cost of each car falls. To the extent that empty car assignments can consider such handling considerations, car sorting and handling can be reduced. As noted above, US freight rail organization and processes separate the assignment and routing decisions organizationally, separating these decisions; thus, such a modeling paradigm would not be appropriate. However, Joborn (1995), Holmberg et al. (1998), and Joborn et al. (2004) explore methodologies to exploit economies of scale for repositioning multiple empty cars in the same group in large blocks, effectively combining the assignment and routing decision. This line of work strives for improved equipment distribution methods for the Swedish National Railway using a deterministic capacitated multicommodity time–space network. They address uncertainty of delivery time by explicitly modeling empty capacity of train routes; resulting in better empty delivery reliability.

7.3.8 Other Areas of Application in Rail

To this point, this chapter has focused on the distribution of traditional mixed merchandise rail freight cars (e.g., box cars, condoles, etc.). Some modeling efforts have been made in intermodal and automotive as well.

In intermodal, Powell and Carvalho (1998a, b) approach intermodal flat car distribution using approximate dynamic programming. The problem is different in intermodal, as the network has fewer nodes, and individual cars are not ordered by customers (rather, they carry customer's trailers and containers).

The automotive industry has much higher concentration of shippers, therefore, shifts in shipping patterns can be a large to railroads' operations and car management. In automotive railcar management, in an attempt to reduce the empty miles traveled by empty automotive railcars between destinations of loaded shipments and origins of subsequent shipments, railroads have created a common pool of automotive railcars that are shared among railroads, and are managed by a jointly owned subsidiary, TTX Corporation. While this arrangement greatly increases the options for railcar assignment to loads (and with the increased options comes lower empty miles), it creates a challenge for the fleet sizing and management of the railcars amongst competing organizations with disparate objectives. Because of the limited size of the network (relatively few origin and destination nodes), the distribution problem is simpler. Thus, Sherali and Tuncbilek (1997) and Sherali and Maguire (2000) discuss the modeling challenges of developing fleet size strategies and help equitably distribute cars and allocate empty repositioning costs amongst the shippers. In this case, annual forecasts are developed, along with estimated monthly fluctuations. Car allocations are a function of relative demand; empty car costs are programmatically distributed amongst the participants based on the results of a time–space network, a series of cost allocation rules, and negotiated agreement amongst the carriers.

References

Ahuja RK, Jha CK, Liu J (2007) Solving real-life railroad blocking problems. Interfaces 37(5):404–419

Cordeau J-F, Toth P, Vigo D (1998) A survey American Association of Railroads (AAR). www.aar.org, 2012

Gorman MF, Acharya D, Sellers D (2010) CSX railway cashes in on optimization of empty equipment distribution. Interfaces 40(1):5–16

Gorman MF, Crook K, Acharya D (2011) North American freight rail industry real-time optimized equipment distribution systems: state of the practice. Transport Res C 19:103–114

Holmberg K, Joborn M, Lundgren JT (1998) Improved empty freight car distribution. Transport Sci 32(2):163–173

Ireland P, Case R, Fallis J, Van Dyke C, Kuehn J, Meketon M (2004) The Canadian Pacific Railway transforms operations by using models to develop its operating plans. Interfaces 34(1):5–14

Joborn M, Crainic T, Gendreau M, Holmberg K, Lundgren J (2004) Economies of scale in empty freight car distribution in scheduled railways. Transport Sci 38(2):121–134

Joborn M (1995) Empty freight car distribution at Swedish Railways – analysis and optimization modeling. Ph.D. Thesis, Department of Mathematics, Linkoping University, Sweden

Jordan WC, Turnquist MA (1983) A stochastic, dynamic network model for railroad car distribution. Transport Sci 17(2):123–145

Narisetty AK, Richard J-P, Ramcharan D, Murphy D, Minks G, Fuller J (2008) An optimization model for empty freight car assignment at Union Pacific. Interfaces 38(2):89–102

Powell WB, Carvalho T (1998a) Dynamic control of logistics queueing networks for large-scale fleet management. Transport Sci 32(2):90–109

Powell WB, Carvalho T (1998b) Real-time optimization of containers and flatcars for intermodal operations. Transport Sci 32(2):110–126

Sherali, Hanif D., and Cihan H. Tuncbilek. “New reformulation linearization/convexification relaxations for univariate and multivariate polynomial programming problems.” Operations Research Letters 21.1 (1997):1–9.

Topaloglu H, Powell WB (2006) Dynamic programming approximations for stochastic, time-staged integer multicommodity flow problems. Informs J Comput 18(1):31–42

Turnquist MA (1986) MOVE-EM: a network optimization model for empty freight car distribution. School of Civil and Environmental Engineering, Cornell University

Turnquist MA, Markowicz BP (1990) TIMS/ORSA conference presentation, Vancouver, BC

Chapter 8
Network Analysis and Simulation

Carl Van Dyke, Marc Meketon, and Bruce W. Patty

8.1 Introduction and Background

Railroad operations are complex processes incorporating several different decisions in order to move a railcar from one location to another. These decisions are often made separately without the ability to easily understand the impact of one decision on another. For example, if a new train is added, and another removed, will the expected connections of the traffic from those trains to subsequent trains still be acceptable, will the network capacity still be sufficient, and will the yards be able to handle the changes in workload? The role of the network simulation capability is to allow analysts to understand how all these disparate pieces fit together, primarily in the context of evaluating operating plan designs and contingency planning.

Network simulation models have evolved over the years and have gained increasing importance, especially as railroads have embraced the need to operate a "scheduled railroad." In this chapter, we will discuss several of these models, as well as describe some examples of how they have been used over the years.

C. Van Dyke (✉)
TransNetOpt, Princeton, NJ, USA
e-mail: carl@cvdzone.com

M. Meketon
Oliver Wyman, Princeton, NJ, USA
e-mail: marc.meketon@oliverwyman.com

B.W. Patty
Veritec Solutions, San Rafael, CA, USA
e-mail: bpatty@veritecsolutions.com

© Springer Science+Business Media New York 2015

B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3_8

8.1.1 Planning and Simulation

The designers of railway operating plans have a number of goals centered on minimizing overall costs while maximizing service. Costs are a function of both direct operating costs and asset or capital costs. Operating costs typically focus on the operation of road and local trains, and the switching of rail cars at both intermediate locations and at customer sites or interchanges with other railways. Asset costs focus to some extent on the use of infrastructure, such as classification yards, but primarily focus on overall rail car fleet requirements and locomotive requirements. In general, asset costs are heavily driven by the velocity with which rail cars and locomotives move, and the total number of locomotives required to operate a plan. For an overview of how all of these considerations come together in the planning process see Ireland et al. (2004) and Stewart (1980).

It is well understood that there are peaks and valleys in traffic levels through the course of a week, and over longer periods of time. While some of this variation is predictable, within bounds there is also a fairly stochastic behavior to this variation. From a cost perspective one never wants to design anything, including a railroad operating plan, for the peak day. On the other hand, an operating plan designed to carry exactly the long-term average, or less than this average, will likely produce unsatisfactory results in terms of service and potentially lost business. Thus, it is the authors' experience that railroad operating plans typically have some excess capacity built into them, perhaps at the 10–20 % level, and the designers must ensure that the plan can handle typical variations in volume.

Customer service is also important, and must be factored into the plan. With some exceptions, such as intermodal, railroad customers consistently report that reliable service is at least as important as fast service (Allen et al. 1985; Roberts and Holcomb 2012; McGinnis 1990; Ballou 2004). Reliability is determined by many factors, including plan complexity, plan adherence, and the extent to which the plan contains sufficient capacity to handle the expected volumes.

Simulation is used in the planning process to address these issues. At the heart of most simulations is the basic trip planning logic described in Chap. 4. The goal of these simulations is to model over a period of time the movement of each shipment, and the associated movement of the trains carrying these shipments. The simulation of locations, including classification yards, is a natural element of this process. The ultimate simulation would also take into account the impacts on locomotive and railcar requirements, and potentially crew requirements as well.

For the purposes of this discussion, simulations can be broken into several varieties:

- Deterministic simulations with fixed plans and no capacity restrictions.
- Deterministic uncapacitated simulations with probabilistic connections.
- Capacitated simulations with fixed plans.
- Capacitated simulations with dynamic plan elements.
- Full Monte-Carlo, capacitated simulations.

Common to all of these simulations is the use of a traffic database. The essence of this traffic database generally does not change based on the type of simulation. The database is made up of a series of shipments, with each shipment having a specific start time and date (referred to as a "release time") over the course of the simulation period. Each traffic record typically includes sufficient information for the shipment to be properly handled by the simulator (origin, destination, load/empty status, car type, commodity, etc.), as well as any other information needed for analysis purposes. The traffic database can be based on an extraction from history, or can represent a forecast. Because the traffic database must include all movements (loaded and empty), and be specific at the origin–destination level, the transformation of forecasts into an adequate traffic database can represent a significant challenge. While the topic of creating a traffic database is beyond the scope of this discussion, the authors wish to note that in their experience the development of a quality traffic database can be as challenging, or even more challenging, than the creation of the simulation or encoding of the operating plan.

The discussion on different types of simulations starts with the following assumptions:

1. That a traffic database is available that contains shipments with release times for the full period of the simulation.
2. That the operating plan has been encoded into a suitable database.
3. That capacities for specific blocks and trains have been encoded, if required.

Given the above, each type of simulation is described in Section 8.2 below.

8.1.2 Other Types of Simulations

There are many other roles for simulation which are not addressed in this chapter. For example, there are line capacity simulations (or dispatch modeling), which are addressed in Chap. 3. Simulations can also be used to estimate the detailed needs for locomotives, train crews, railcars, as well as the best way to use rail yards. Some of these topics are touched on briefly in this chapter, and a number are examined in more detail in other chapters found in this book.

Simulations do not always need to look at multi-day time horizons. One example of this is the ability to estimate train sizes as the trains are designed. This type of simulation provides an ability to give rapid, interactive feedback to the train designer as the trains are created, and is examined in a later section of this chapter.

8.2 Types of Network Level Simulations

As noted above, there are at least five types of network simulations, each of which is described below. To fully understand the basics of computing the car schedules which underlie the simulation process, please see Chap. 4 on Car Scheduling.

The Car Scheduling and Train Planning chapters should also be reviewed to understand the process of specifying and applying train level capacity limits. Finally, Chap. 5 on Blocking should be reviewed to understand the process of routing the railcars from yard to yard.

8.2.1 Uncapacitated Deterministic Simulations with Fixed Plans

This is the simplest type of simulation. Under this approach the operating plan is typically represented for a full week, and a set of traffic is provided that represents the shipments that are expected to move over the course of this week, including a release time and date when each shipment becomes available for movement. It is important to note that if a train operates on more than one day of the week, each specific instance of the train will be separately represented in the simulation (e.g., train 301 of Monday, train 301 of Tuesday, etc.).

Because the simulation is uncapacitated, each shipment will follow its "default" trip plan, which also means that each shipment can in effect be treated as independent of all other shipments. This fact allows for simple use of multi-threaded logic in the simulation to increase computational performance. Thus, the simulation proceeds by generating the expected trip plan for each shipment using the logic described in Chap. 4 on car scheduling. These trip plans are saved to a database. Over the course of the simulated week, some traffic may have trip plans that extend into the second week. Because every week is treated as having the same schedule in the simulation, a wrap-around strategy is used so that traffic that uses a train in the second week has those volumes assigned to the matching train in the first week for workload estimation purposes. This principal is also applied to the estimation of yard workloads. Using this approach eliminates the need for a warm-up period in the simulation.

Each trip plan can be reorganized into a set of train leg records and a set of connection records. Sorting the train leg records by specific train instance allows each train to be profiled. Sorting the connection records by each location allows each yard to be profiled. While there are no stochastic aspects to this approach, it does show the "natural" or "unconstrained" workloads that should be expected at each yard over the course of the week, and unconstrained sizes of each train, and can be used to adjust the plan to ensure it satisfies the design goals.

One example of this type of simulation is the USRA/SRI network model developed for the formation of Conrail (Siddiqee and D'Esopo 1977, 1975; Siddiqee et al. 1975). The USRA/SRI model was later adapted for use at a number of other railroads both inside and outside North America. Another example is the uncapacitated simulation option within the MultiRail model developed in part by the authors Van Dyke and Meketon (Oliver Wyman). Within this type of simulation, the operating plan is fixed, and no trains are annulled, added, or have their times changed. All scheduled connections are made. A number of other simulation models are believed to exist that use similar principals, including one developed for BNSF Railway.

8.2.2 Uncapacitated Deterministic Simulations with Probabilistic Connections

As part of the Freight Car Utilization Project, a research program jointly sponsored by the U.S. Federal Railroad Administration and the Association of American Railroads, the concept of the "PMake" function was developed (Martland 1982). PMake stands for the "Probability of Making a Connection." The focus of this research work was to understand the causes of variability or unreliability in the transit times of rail cars. For conventional car load traffic, the critical element was found to be the consistency with which inbound railcars departed on the expected outbound trains when making connections at rail yards. The PMake function can be used to capture the consistency with which these connections can be achieved.

Using the principle of the PMake function, a simulation model called the Service Planning Model (SPM) was developed (McCarren and Martland 1980; Van Dyke 1981). This simulation model is in many respects similar to the uncapacitated, deterministic simulation with fixed plans described above. The major difference is that each shipment was in effect broken apart into multiple trip plans based on a set of probability functions. For example, if the probability function specified that there was an 80 % chance of making the first connection at a yard, then a trip plan would be generated for that connection, and 80 % of the traffic would be assigned to the connection. The remaining 20 % would be held back, to make a later connection. If for the next connection at the same yard, there was a 90 % chance of making the connection, then a second trip plan would be generated using that connection, and 18 % of the traffic would be assigned to the connection (90 % of the 20 % that missed the first connection).

This process would then be repeated until all of the traffic was assigned to an outbound connection or some maximum time had elapsed. In this way a trip plan distribution can be developed for each shipment, and the likely level of service for customers can be projected. These various trip plans can also be used to project train sizes and yard workloads over the course of the week, so that the same benefits as the previously described uncapacitated, deterministic simulation with fixed schedule can also be realized, with the added benefit of also having better quality transit time and reliability predictions. The authors found that in practice railroads were sometimes uncomfortable with this approach as many railroaders felt it set ambiguous expectations as to field performance. They preferred the deterministic simulations from a communications perspective in terms of telling the field to try to achieve 100 % of the targeted connections.

8.2.3 Capacitated Simulations with Fixed Plans

Railways generally accept that capacity limits will cause some traffic to be delayed. In the simplest strategy for handling peak volumes, the excess traffic is simply held for a later train that has capacity (called "rolling" traffic). For example, if Thursday is a peak day, and the amount of traffic to be carried exceeds a train's capacity, then the excess railcars are held and put on Friday's train. Under this approach, train capacities are largely treated as fixed, and trains are not added or dropped based on volume variability. Instead traffic is simply delayed to a later train that does have capacity. In some cases the planner might schedule additional trains on certain days of the week to handle the excess capacity, but this is done on a fixed plan basis, not a dynamic plan basis. An underlying assumption in this approach is that the overall capacity of the plan is greater than the average volumes to be handled by the trains.

Simulating such situations requires an event-based simulation where multiple traffic records are processed in parallel, and thus a different approach is required compared to the simpler simulations described above. In general, the operation of the actual trains is treated as deterministic—once a train departs, it will arrive at its subsequent route locations as scheduled. As a result, one can design the simulation to be focused on locations, not the operation of the trains, where the events that are tracked consist of the releases of traffic at the origins, the arrival and departure of trains, and the arrival of shipments at their destinations. In reality, the simulation can be further focused only on train departure events, as these are the only critical decision points (once on a train, the traffic can be immediately advanced to the location where it gets off the train, since the movements of the trains are fixed).

Thus, the core concept is to advance shipments from location to location in accordance with the trip planning principles described earlier, keeping a list of all traffic awaiting each outbound train at each yard. As the clock advances, each train departure is identified, and the list of traffic that could ride on the next departing train is processed (once traffic is placed on a train, it is assumed it will not be removed until it reaches its planned set-off point). If the available traffic is less than the capacity of the departing train, then it is all placed on the train, otherwise, business rules are applied to determine which traffic will depart and which traffic will be delayed to a later train. The concepts related to the specification and management of the capacities are discussed in detail in Chap. 4 on car scheduling.

This type of simulation is effective in showing that a plan has adequate capacity, and the extent to which traffic will be delayed due to capacity constraints. In general, yards are treated as having no capacity limits in this approach. While some customer service impacts can also be accessed through this approach, the impact on specific customers cannot be readily assessed by such simulations due to the unpredictability of when capacity constraints will occur in reality. The degree of impact on customers also depends on the extent to which the simulation accurately reflects the decision process on which traffic will be advanced and which traffic will be delayed when capacity constraints are encountered. The authors Van Dyke and Meketon have developed several simulations of this type as part of the MultiRail planning platform. A version of this approach was also used in the Conrail Network Analysis Model (CNAM), developed by American Airlines Decision Technologies for use by Conrail in the early 1990s, with involvement by the author Bruce Patty.

A warm-up period is required in this type of simulation, which generally needs to be somewhat longer than the longest trip time experienced by any shipment.

8.2.4 Capacitated Simulations with Dynamic Plan Elements

In actual railway operations, a number of actions are taken to adjust the train plan to reflect the impact of volume variations:

- Extra trains are added.
- Some trains are annulled or consolidated together.
- Some trains are operated only when sufficient traffic exists.

By extending the event-based simulation described above, it becomes possible to incorporate dynamic train operations into the simulation. By tracking the volumes of traffic accumulating at each yard for each outbound train, business logic can be developed to trigger changes to the train plan. Adding extra trains, or annulling selected trains, is fairly straight forward provided the extra trains are simply copies of existing trains, and the overall train plan remains fixed.

Of more interest is the creation of "demand" driven trains. In the authors' experience, while the North American industry has generally moved to a fixed schedule-based approach, a number of railroads such as those in the former Soviet Union continue to operate on a demand driven or "tonnage" basis. Under this scenario the train and blocking plans remain fixed in the sense that there is only one way to route each shipment, but the departure time of the trains becomes variable. The concept is that a train will not depart until it becomes full. To simulate this, one must designate which trains are to be demand driven, and the event-based simulation must include a time-based check to determine when sufficient railcars have accumulated to trigger the operation of the train. This can also be combined with a maximum delay time that causes the train to be operated even if not full, once the most delayed shipments being held for a train exceed some amount of time. In practice, the authors have observed that trains are often filled out with small quantities of other traffic going in the same general direction, rather than further delay the departure of the train. This can be simulated using the concept of a fill block, which is discussed in Chap. 4 on car scheduling. The MultiRail planning platform contains a simulation mode that supports the above scenario (Oliver Wyman).

The authors believe that the above approach could prove useful in the planning of North American railroads through the simulation of unit trains and solid trains that are effectively demand triggered. For example, while the volume of unit train traffic between a pair of locations may be known, it is not possible to specify a fixed schedule to move this traffic. At the planning level one can only specify the route and make-up of the unit trains, and that they will have an expected frequency over a relatively long period of time such as a month. However, during the simulation the volumes available for each unit train can be tracked, and the movement of the trains triggered by sufficient traffic being available for each train.

This approach could also be used for certain other types of operations, such as the movement of large lot grain shipments. For example, two or more lots of cars from separate loading points are sometimes combined into a single train for a common destination in the management of grain traffic. Typically, a fixed plan moves the grain cars to a consolidation point near the shipment origin locations. There the groups of cars are examined and some are released for movement in the general carload network, and others are combined into solid trains. The simulation approach described above could be modified with appropriate business logic to reflect such operations. Other examples where demand triggered trains might prove useful in a simulation include the movement of empty grain cars and autorack cars. The authors are not aware of a current simulation that uses this approach for this purpose.

8.2.5 Full Monte-Carlo Capacitated Simulations

The next step beyond the simulation approaches described above is to develop a full Monte-Carlo capacitated simulation. Under this approach, stochastic elements are introduced to reflect variations in train operations, variations in yard operations, and potentially even events like periodic bad ordering of cars. The authors are not aware of any successful examples of such simulations at a network level, though some attempts have been made to create them (Allman 1966a, b; Bellman 1967; Wilson and Hudson 1970; Petersen and Fullerton 1975).

With one exception related to the simulation of crew districts, it is unclear to the authors what the advantage and purpose would be of creating such a simulation. Our general assumption is that the primary purposes of simulating an operating plan are (1) to determine the quality, robustness, and cost effectiveness of a specific operating plan, or (2) to determine the overall costs and resource requirements associated with a future scenario based on a traffic projection. We believe that the above described approaches are more than adequate for these purposes, and see relatively little incremental benefit from a highly detailed, Monte-Carlo type simulation.

The authors built a full Monte-Carlo simulation of train movements for a single crew district to estimate minimum crew pool size requirements. In this case, significant stochastic effects due to trains leaving late and non-deterministic levels of unit trains have measurable effects on crew pool sizes needed to meet a given service level. Section 8.3.1 has additional details.

While prior attempts at building such simulations have not proved effective, with the continuous advance of computer capabilities, it is the authors' view that it would certainly be feasible to construct such a simulation at this point in time.

8.3 Resource Estimation

One of the goals for performing a network simulation is to estimate resources such as the number of crews, locomotives and cars needed to operate the plan, as well as yard workloads. There are two common approaches that are used for estimation:

1. Use a sophisticated, detailed model that simulates or optimizes assets. For example, given the train schedule and the estimated train volumes derived from the network simulation, use a locomotive model that plans which locomotives should be assigned to which train.

2. Use historically derived ratios of key statistics, such as the ratio of the number of locomotives to the gross-ton-miles. The network simulation estimates the gross-ton-miles, and then the historical ratio is used to estimate the total locomotive requirements.

8.3.1 Estimation of Crews

Railways generally have two types of crew assignments: pooled crews and assigned crews. In the first case, there is a "pool" of crews available to take trains within a crew district, and they are called in a round-robin manner. A specific crew could take different trains on different days. There are many rules that affect which crew is called for which train that will not be discussed here.

There are two types of detailed models that have been used in the industry for estimating pooled crew resources, plus a simpler approach based on the results of a network simulation or assessment of the trains operated in crew district.

The first type of detailed model is a simulation model that simulates the crew calling process for both assigning crews to trains and for determining when to deadhead a crew from an away location to their home base. The simulation begins with no-crews, and adds crews one by one as they are needed. These types of models can handle a variety of random effects such as late trains and crews unexpectedly taking time off—called "marking off" in North America. This approach can also model crew rest times.

The second detailed model is a multi-commodity network flow model that cycles crews and determines the deadheads. See, for example Vaidyanathan et al. (2007). The optimization approach is more suited for tactical planning and less suited for estimating the number of crews because it does not account for various stochastic issues which lead to sizable increases in the number of crews needed.

However, it is most common in the industry to use simple historical ratios of the number of crews per crew start. The number of crew starts comes directly from the operating plan or network simulation results and hence this is easy to apply. Because it is based on historical ratios, it implicitly accounts for the various random effects such as late trains and mark-offs.

The detailed models are sensitive to the specifics of the operating plan such as the timing of the trains and the overall balance of trains within the crew district. A crew district is balanced if the number of trains going in one direction is the same as the number of trains in the other direction. Imbalanced schedules lead to additional deadheads of crews.

By contrast to the pooled crews discussed above, assigned crews are where specific crews are pre-assigned to trains. Freight railroads often have this for local crews where the schedule is predictable and cyclic. An example is where crew 101 takes local train L321 each weekday, starting at 0800. While the exact set of locations that train will visit could vary each day, the crew running the train does not vary.

In many cases estimating the number of assigned crews can be done by inspection because the assignments are simple due to a crew being assigned a single train per day. In situations when crews are assigned to multiple trains per day, it might become complex. An extreme example, although not applicable to North American freight railroads, is planning crews for commuter railroads where the crews may take four to six trains per day. That is solved by sophisticated column generation techniques that are very similar to the well-published airline crew scheduling algorithms. It is fairly common in European freight operations for a train driver to operate more than one train in a day.

See Chap. 6 on crew planning for further discussion of this topic.

8.3.2 Estimation of Locomotives

To our knowledge, there are two general types of detailed models for locomotive estimation:

1. A multi-commodity network flow algorithm loosely based on airline fleet assignment models (Hanes et al. 1995) which models trains over a time-space network, similar to the dynamic car scheduling network (see Chap. 4 on car scheduling), but where the integer variables represent the number of each type of locomotive to place on a train. Complications include modeling coupling/de-coupling between the locomotives (many railroads discourage breaking apart locomotive sets), constraints on the combinations of locomotives that could be used at the same time on a train, and where to use light-engine-moves (where trains are composed entirely of locomotives that needed to be repositioned from one yard to another). See also Luo and Meketon (1997).
2. Approximate dynamic programming (Powell et al. 2014) that simulates the movement of locomotives over time and runs a series of small optimizations to locally optimize the locomotive assignment. There is a feedback loop that examines the proposed solution, which allows the system to adjust the cost parameters and then regenerate the solution. Admittedly, this is the highly simplified explanation.

The multi-commodity network flow algorithm tends to find solutions that use significantly fewer locomotives than what the railroad needs, primarily because it has a perfect forecast of all the trains and their exact timing over the week. These models are calibrated to have artificially large minimum connection times so that the locomotive fleet count estimates are closer to reality.

The approximate dynamic programming model also requires extensive calibration, although the “simulation” part could directly model the train schedules that deviate from the operating plan.

In many planning situations, railroads use historical ratios for estimating locomotive requirements instead of the more sophisticated optimization tools. One railroad the authors are familiar with uses the following historical ratios:

1. Locomotive days per 1,000 ton-kilometers.
2. Percent of time a locomotive is active, but not pulling a train (such as during a light-engine-move).
3. Percent of time a locomotive is in maintenance.

These factors are then used to estimate total locomotive requirements for a specific plan. Often these factors are broken out by train type, and geography to improve the accuracy of the estimation. An obvious example is to tie the estimation process to train weight to approximate the number of locomotives on each train, and train hours to estimate the total train-related locomotive hours needed by a plan. The results of a network simulation can directly feed into this process.

See Chap. 2 on locomotive planning for further discussion of this topic.

8.3.3 Estimation of Railcar Requirements

To our knowledge, the majority of estimation techniques used by North American railroads for railcar fleet requirements are based on historical ratios and not based on any detailed simulation or optimization models. The main reason for this is that a significant driver of fleet requirements is the amount of time a car spends at a client location being loaded or unloaded, which is generally not captured by simulation models. Instead these load and unload times are estimated based on historic performance, often using car movement history data.

Within a network simulation model there is a traffic file that has car counts by car type and commodity. This traffic data reflects the number of cars needed for the shipments, but not the number of cars needed overall. The usual methods for estimating railcar fleet requirements rely on counts of the number of loaded movements (and sometimes empty movements) combined with separate estimates of the typical car cycle:

- Time to load the car at the shipper.
- Time to move the car to the consignee.
- Dwell time at the consignee while unloading.
- Time to return the empty car to either the next load or a staging area.
- If at a staging area, wait time for next shipper demand, plus the time to move the car to the demand location.
- Additional time allowances for activities such as repairs, cleaning, and storage.

Network simulations can directly estimate the time spent moving the railcar from shipper to consignee including the dwell times at the intermediate yards, as well as the time spent moving the empty cars (providing empty car movements are in the traffic file). But the time spent at the shipper and the consignee, as well any time dwelling while empty at a staging area, being cleaned, repaired, etc., are not captured by the network simulation and need to be estimated using historical car movement data.

Once all these quantities are known, then the total car days can be calculated. The usual way is to take each traffic record, with its car count, use the network simulation results to calculate the transit time, add half the historical time to load and unload at each end of the movement, plus any factors to allow for cleaning, repair, etc., to estimate the total car-days for that movement. The sum of all the car-days divided by the simulation time (usually 7 days) is the average number of cars needed.

Planners often prefer to determine the peak number of cars needed, and not just the average number of cars required. One method for calculating peak usage is to use a histogram of railcar usage over the course of a week created by layering each traffic movement (plus allowances to load and unload) together, as illustrated in Fig. 8.1. The blue bars in the below diagram represent the railcars in transit, which comes directly from the network simulation. The yellow patterned bars represent the time the car spends being retained at the client or storage yard, and is developed from historical data. Below the bars is the histogram that illustrates the number of cars in use over time. The quantiles of car usage could then be derived from the histogram—the 100 % quantile is the absolute peak usage for the week, the 50 % is the median (and in our experience is within 0.1 % of the average). The 90th percentile is commonly used for estimation.

The below chart (Fig. 8.2) shows actual data from a large railroad, although the numbers have been scaled for confidentiality, and the railcar type is not displayed. The 90 % peak usage is 32,039 railcars.

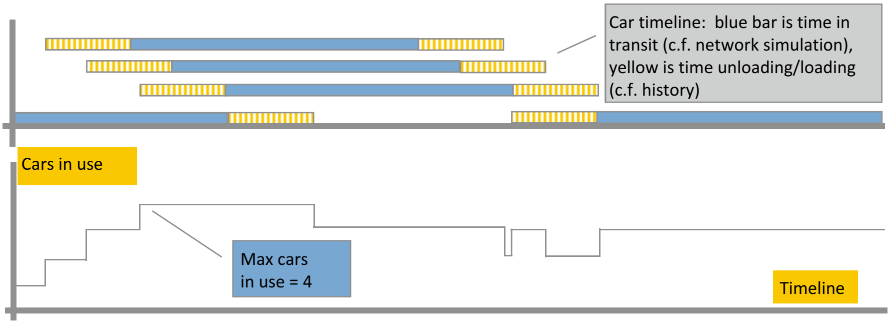

Car timeline: blue bar is time in transit (c.f. network simulation), yellow is time unloading/loading (c.f. history)
Cars in use
Max cars in use = 4
Timeline

Fig. 8.1 Sample peak usage report

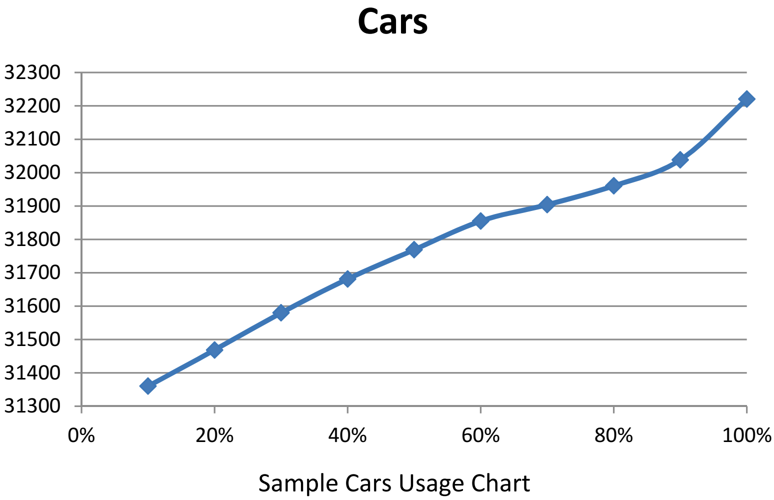


Fig. 8.2 Sample cars usage chart

Usually the railcar estimation process will be for aggregated car fleets. For example, there may be an estimate for “tank cars” but not for specific kinds of tank cars.

8.3.4 Estimation of Yard Workloads

There are three primary metrics used to estimate the switching activity workload at a yard:

- The number of cars that are switched between inbound and outbound trains.
- The number of block swaps. This is usually not dependent on the number of cars associated with a specific block swap.
- The number of cars that are re-humped. This happens when a car needs to be switched more than once in a yard. A typical example occurs for yards that make road blocks and local blocks. Since there may be many local blocks, the cars from inbound trains going to outbound local blocks may first be switched to a track that is for a collection of several local blocks. Then at specific times of the day these cars are switched again into more specific local blocks. In general, when a yard makes more blocks than they have tracks to hold the blocks, some set of cars get switched into a holding track and later re-switched.

The number of cars that are switched could be estimated only using the block sequence as explained in Chap. 5 on blocking. Or it could come from the network simulation, which can also provide a day-by-day/hour-by-hour estimation of the yard workload. The number of block-swaps could be somewhat estimated by the analytical methods related to train volume estimation discussed later in this chapter, but most often is derived directly from the network simulation.

There are two avenues for estimating the number of cars being re-humped. The easiest, and most practical, is to apply historical ratios. If in the past 10 % of the cars are re-humped, assume that ratio will be effective in the future.

The other way is if the plan not only specifies the outbound block and train, but specifies which track in the yard is responsible for holding the cars that form that outbound train-block. The network simulation could then be enhanced to reflect the need to re-hump cars. When classifying the cars from an inbound train, if the outbound train-block’s track is not available or does not have capacity to hold the car, then the car would be sent to a hold track for re-humping. This might also happen when cars are delayed to later trains in the capacitated simulation. Albeit this is complex, but such approaches have been studied in Europe, where it appears that re-humps play a larger role in the classification process. It is not unusual for an European yard to have 15 tracks and make 30 blocks, which means that not all blocks are able to exist at all times.

It should be noted that in addition to estimating the car switching activities at a yard, the network simulation (or the base plan) can also be used to reflect a number of other aspects of yard workload:

- Number of trains arriving/terminating at the yard by time-of-day and day-of-week.
- Number of trains being made-up at the yard/departing the yard by time-of-day/day-of-week.

- Connections that must be made between in-bound and out-bound trains and the time available to make such connections.
- Expected inventory levels at the yard based on the in-bound and out-bound train patterns and connections that must be made.

The above information can be used in a variety of ways. One is as an input to a detailed yard simulation designed to test the feasibility of an operating plan. A second is as an input to a resource requirements model to determine the necessary staffing levels at a yard over the course of a week, including railcar inspection and repair personnel, switch engine and associated crew requirements, ground crew requirements, and overall supervisory requirements. For more information, please see Chap. 9 on Yard and Terminal Simulation.

8.4 Roles of Network Simulation

Network Simulation tools allow analysts to perform "what if" studies in support of various strategic or tactical initiatives. In this section, we describe a few of these occasions and how network simulation models have been of great value.

8.4.1 Mergers

Since the merger of the railroads in the northeastern US in the 1970s that resulted in the formation of Conrail, network simulation models have been used to support merger studies. This includes both the process of evaluating the economic benefits of a merger in order to determine the price that should be offered to purchase another railroad, as well as putting together a comprehensive operating plan that supports a cohesive operation across the newly combined railroads. One of the most recent mergers, and perhaps the most complex, was the acquisition of Conrail by both CSX and Norfolk Southern. This acquisition was especially complex because Conrail's network was to be split up between CSX and Norfolk Southern. Therefore, not only did both Norfolk Southern and CSX have to determine how to incorporate new terminals, rail lines, and customers into their networks, in many cases they did not have the option of handling the business the way it had been handled before.

All three of the authors of this chapter participated to some degree in the modeling of the merger/acquisition of Conrail by NS and CSX. MultiRail was used by both railroads in order to evaluate alternative train schedules and blocking plans. The two mergers that had taken place immediately prior to the Conrail acquisition were the acquisition of Southern Pacific by Union Pacific, and the merger of Burlington Northern and the Santa Fe railroads. Both of these efforts encountered significant operational issues when the merged railroads began to operate as one. For months, service was hampered and customers were impacted. Both NS and CSX wanted to do everything they could to avoid such situations, so they relied heavily on network simulations to help them identify what might happen under varying situations so that revisions to the plans could be made to avoid problems.

Not only were network simulations used to identify “watch outs” for the Conrail acquisition, they were also used to develop revised train schedules and blocking plans for traffic “touching” the newly acquired territories.

Rail planners used the results of the simulation runs to determine the number of cars that were assigned to each block in order to identify opportunities to improve the plan. For example, consider the case of a block at Atlanta that had 60 cars assigned to it and another which only had five. Rail planners would look at the destinations of the cars in the 60-car block to determine if service could be improved by splitting out some of the traffic in the 60-car block into a new block. At the same time they would evaluate the five-car block to see if there were alternate methods to handle this traffic, so the freed up blocking capacity could be used to form a new block.

As a concrete example, prior to the acquisition of part of Conrail by NS, traffic heading from Boston to Atlanta might be placed into a block at Boston to be interchanged with NS at some location, like Hagerstown, where NS would then reclassify the cars into an Atlanta block. After the acquisition, NS might want to build an Atlanta block in Boston, if there was sufficient traffic to warrant such a block. This would eliminate the intermediate handling at Hagerstown and reduce transit time. Similarly, NS might want to build a Boston block in Atlanta to avoid intermediate handlings for freight headed to Boston. However, not all terminals had sufficient capacity to build blocks to all newly acquired destination terminals on what was previously a different railroad. Rail planners used network simulation models, primarily MultiRail, to assist them in evaluating alternative revised operating plans for the new traffic and the new network.

They also needed to determine which trains should carry these new blocks and develop new block-to-train assignment rules. Train capacities had to be taken into account, so it was necessary to perform network simulation runs to determine what the estimated train sizes would be for a given scenario.

The effort to develop revised operating plans took several months, but it helped to avoid the magnitude of issues that had plagued the earlier mergers.

8.4.2 Network Modifications

Another role for network simulation models is to determine the impact of modifications to the rail network. Typical modifications include closing a terminal, opening or expanding a terminal, or adding more capacity to a rail line. For example, when American Airlines Decision Technologies (AADT) developed the Conrail Network Analysis Model (CNAM) for Conrail, one of the first analyses to be performed using the model was the closure of the Enola terminal, across the Susquehanna River from Harrisburg, PA.

The Enola terminal was an older terminal and had outserved its purpose, at least in the mind of several people at Conrail. But, before they completely closed the terminal, they wanted to be able to determine if their ideas for how to handle the traffic that used to be classified there would work. CNAM, unlike most network simulation models, had the ability to take terminal and train capacities into account during a model run. If more cars were assigned to a train than the capacity of the train, then the model would not allow some of the cars to be moved by that train. Rules were put in place that dictated the creation of "extra sections" or "second sections" that could move the overflow traffic, or the traffic would be shifted to the next scheduled train that carried the blocks with the delayed cars.

While rail planners had to determine on their own the various strategies that they wanted to test regarding how the traffic that used to be classified at Enola was to be handled, the model could report on what would happen as a result of each revised operating plan. Results were stored in an Oracle database, so comparisons between model runs could be performed by comparing the events in the event table. This could be done at a terminal level, a train level or even at the individual car movement level.

8.4.3 Emergency Situations or Special Circumstances

Occasionally, railroads need to modify their service plans in response to unexpected situations. In these cases, it is often helpful to be able to simulate the revised plan to make sure that nothing has been overlooked before the revised plan is put into place. One example of such a circumstance took place in conjunction with the Atlanta Olympic Games in 1996. Shortly prior to the games, the decision was made to divert any hazardous materials being moved by the railroads operating in the vicinity of Atlanta. CSX and Norfolk Southern both needed to revise their operating plans, especially any exception logic in their blocking plans that would route hazardous material near Atlanta. While this was quite an effort, especially for CSX that did not have algorithmic blocking in place, both railroads needed to determine what this would do to the volumes on trains and at the terminals to which the traffic was diverted.

Other situations where similar capabilities are beneficial include major derailments, bridge closures, or major rail construction projects.

8.5 Average Day Analysis

One of the leading reasons for simulating an operating plan is to understand the workloads the plan will experience by day of week, and ensuring that the plan contains sufficient, but not excess, capacity. Implicit in this process is that the plan overall includes sufficient capacity over the course of a typical week to handle all of the expected traffic. If this is not the case, then the queues of delayed traffic would grow ever larger in the simulations, or the simulation would be forced to ignore the capacity restrictions at some point.

The result is that there is a need in the design process to understand if sufficient capacity exists on an average basis as the plan is created to handle the expected traffic. Due to the time delays in running a simulation, and the need for a complete plan when doing a simulation, a faster mechanism is needed to estimate train sizes as the trains are created and modified that does not require that the plan be complete.

One way to address this need is to use average day analysis to estimate the sizes of trains on an interactive basis. This is in effect a simplified version of the trip planning process. Three approaches to average day analysis are presented below.

8.5.1 Uncapacitated Average Day Analysis

The authors Van Dyke and Meketon have developed simple, uncapacitated train size estimation processes for use in MultiRail and other software tools that can be used effectively during the design process to provide immediate feedback on expected train sizes as trains are defined by the planners (Ireland et al. 2004). The basic approach leverages the trip planning concepts. As discussed in Chap. 5 on blocking, given the destination of each block, it is possible to route all traffic across a set of blocks, generating the "block sequence" for each traffic record, and based on this an estimate of the volume of traffic that will be carried by each block.

As discussed in Chap. 1 on train scheduling, using the specification of each train we can determine which blocks are carried by each train. Given the estimated volumes for each block, and the block-to-train assignments, we can estimate the expected size of each train per run of that train. Consider the case of a train AB1 that goes from A to B that carries an A–B block that has an estimated 350 cars per week assigned to it. If train AB1 is the only train to carry the A–B block, and it carried only this individual block, and it operates 7 days per week, then the average train size would be expected to be 350/7 or 50 cars per run of the train. Now consider a case where the train only operates 5 days per week. This would mean that the cars per run of the train would increase to 350/5, or 70 cars per run. If the train AB1 carried other blocks, the contribution of these other blocks to the train size would also be taken into account, and the overall estimated train size would be adjusted accordingly.

A more complex case would be to introduce a train AB2 that also carried the A–B block from A to B. If trains AB1 and AB2 both operated 7 days per week, then the volume per run of each train would become 350/14, or 25 cars per run. However, if one train operated 7 days per week, and the other 5 days per week, we would need to adjust the cars per run of each train to 350/12, or 29.2 cars per run. Variations of this approach can be used to also handle block swaps.

Overall, this estimation approach has the advantage of being simple and fast, thus providing a means to quickly estimate train sizes as blocks are assigned to trains and train frequencies changed. Note that this approach impements a strict a frequency-based allocation, with no bias in how it allocates the traffic to each run of the trains carrying the block. Thus, it cannot directly handle issues related to capacities, designation of some trains as secondary for moving a particular block, and other factors that might impact the relative volumes on trains. See the discussion on capacitated average day analysis in Sect. 8.5.2 below for examples of how to take a more sophisticated approach.

This approach also does not handle variations of train size due to the timing of train schedules and the time-of-week when shipments are released. In the example above with trains AB1 and AB2 each running 7 days per week, if train AB2 departs shortly after AB1, then AB1 would have siphoned off most of the traffic and the train AB1 may will receive much more than 25 cars/run while train AB2 would have received that much less.

The authors have worked with many railroads that find use of average day analysis for their planning works reasonably well. There are only a couple of railroads that prefer to use volume estimation based only on using trip plans due to the variations of how traffic is assigned to multiple competing train-blocks based on the train schedules and shipment release times. They prefer to use full trip plans, estimate the volume for each run of the train, and then average the volumes over each run of the train. The trade-off, as mentioned earlier, is that the computation time goes from nearly instantaneous using the average day analysis to perhaps 30 min using the trip plan simulation approach.

8.5.2 Capacitated Average Day Analysis

The authors have also developed algorithms for undertaking average day analysis in the presence of capacitation. Here are three examples we have seen from different railroads:

Example 1 Railroad has a block (AB) from A to B, and two trains that go from A to B: one is a road train (RAB) that is a direct, non-stop train, and the other is a local train (LAB) that makes intermediate stops between A and B to pick-up or set-off cars. The local train also carries the AB block, but it is treated as an overflow block—traffic should be placed on the train-block only if the road train is at capacity.

Example 2 Railroad has a block AB from A to B. There is one train (RAB) per day from A to B. There is an intermediate, major yard C that is somewhere between A and B. If the traffic on the AB block exceeds the capacity of the RAB train, then the excess traffic is routed on the road train RAC from A to C, block swapped at C, and then carried on the road train RCB.

Example 3 Railroad has a road train RAC whose route is A–B–C. It carries two blocks, AC and BC. The BC traffic is deemed "hot" and is desired to ride only on the RAC train, while the AC block has alternative trains that can carry the block. While some traffic on the AC block can ride the RAC train, there must be room on RAC for all the BC traffic. This is an example of reserving capacity on a train for "downstream" blocks.

In this section, we describe a mathematical model that uses linear programming to develop average day analysis when trains and train-blocks are capacitated. The authors have found in practice that this runs very quickly—less than one second—for medium size railroads.

It is important to note that the below analysis does not roll traffic from one day to the next day—rather it only reallocates traffic from one train to another, where both trains carry the same yard-block. In more detailed simulations, the capacitation limits may force some traffic to leave on the same train as originally intended, but a day later. The analysis below does not handle this roll over, rather only the case when traffic has essentially a primary train symbol or train-block and when that reaches capacity the traffic could roll to a different train symbol.

Consider the following overly complex set of trains and train-blocks (this example is artificially complex to illustrate the considerations that must be made in the model).

A block from A to E has daily volume of ten railcars per day. That is 70 railcars per week. The following trains carry this block:

Train	Pickup	Setout	Runs per week	Uncapacitated volume per run of the train	Capacitated volumes, no priorities	Capacitated volumes with priorities
101	A	E	5	7	7	10
102	A	B	2	7	7	4
103	B	E	2	11.5	10	6.57
104	A	C	3	7	7	4
105	C	B	3	3	2	1.71
106	C	D	4	3	3.75	1.71
107	D	E	6	2	2.5	1.42

Fig. 8.3 Data used in capacitated average day example

Graphically this looks like:

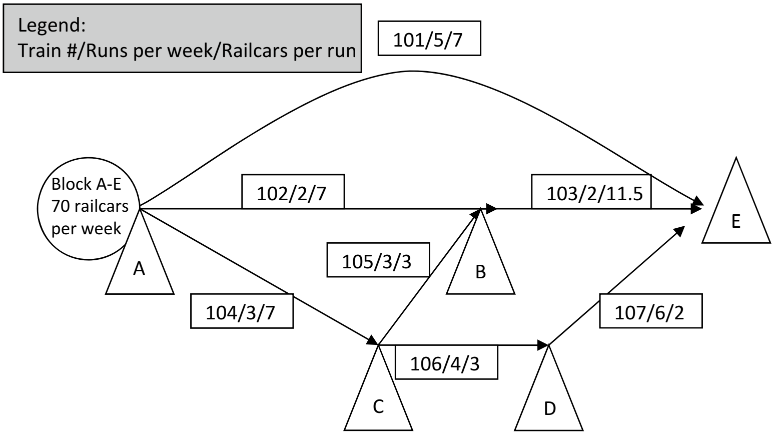


Fig. 8.4 Graphical illustration of the train-blocks used in capacitated average day example

We will limit the mathematical formulas to this one example, and we only consider capacitation limits on the number of cars. In actual application, limits are usually on both length and gross weight, not on cars, which increases the number of constraints and variables. We will discuss other implications of constraining by both length and volume a little later.

Let $X_{t,b,p,s}$ = the (non-negative) volume assigned to train *t* carrying block *b* with pick-up at location *p* and set-off at location *s*. In the example above, we have seven of these variables: $\{X_{101,AE,A,E}, X_{102,AE,A,B}, X_{103,AE,B,E}, X_{104,AE,A,C}, X_{105,AE,C,B}, X_{106,AE,C,D}, X_{107,AE,D,E}\}$.

We need to obey block-swap conditions at locations B, C and D—cars in equals to cars out. This involves both the volumes on the train-blocks and the runs per week each train-block makes:

$$\begin{aligned} &2X_{102,\mathrm{AE,A,B}} + 3X_{105,\mathrm{AE,C,B}} = 2X_{104,\mathrm{AE,A,C}} \\ &3X_{104,\mathrm{AE,A,C}} = 3X_{105,\mathrm{AE,C,B}} + 4X_{106,\mathrm{AE,C,D}} \\ &4X_{106,\mathrm{AE,C,D}} = 6X_{107,\mathrm{AE,D,E}} \end{aligned} \tag{8.1}$$

We need to have the total volume on block AB to be allocated to the three train-blocks that pick-up at A:

$$70 = 5X_{101,\mathrm{AE,A,E}} + 2X_{102,\mathrm{AE,A,B}} + 3X_{104,\mathrm{AE,A,C}} \tag{8.2}$$

The volumes cannot be negative, so we have the constraints:

$$\begin{aligned} &X_{101,\mathrm{AE,A,E}} \geq 0, X_{102,\mathrm{AE,A,B}} \geq 0, X_{103,\mathrm{AE,B,E}} \geq 0, X_{104,\mathrm{AE,A,C}} \geq 0, \\ &X_{105,\mathrm{AE,C,B}} \geq 0, X_{106,\mathrm{AE,C,D}} \geq 0, X_{107,\mathrm{AE,D,E}} \geq 0 \end{aligned} \tag{8.3}$$

We are allowed to place limits on the size of the train-blocks and the size of the trains. Placing limits on the train-block size is easy. If we say that all train-blocks must have no more than 10 cars, we would write:

$$\begin{aligned} &X_{101,\mathrm{AE,A,E}} \leq 10, X_{102,\mathrm{AE,A,B}} \leq 10, X_{103,\mathrm{AE,B,E}} \leq 10, X_{104,\mathrm{AE,A,C}} \leq 10, \\ &X_{105,\mathrm{AE,C,B}} \leq 10, X_{106,\mathrm{AE,C,D}} \leq 10, X_{107,\mathrm{AE,D,E}} \leq 10 \end{aligned} \tag{8.4}$$

In this example, the trains only have one train-block, so limiting the size of the train-blocks is the same as limiting the size of the trains. In more complex situations, the trains carry several train-blocks, and at each pick-up location the sum of the train-block volumes for the train-blocks that are picked up at the location or are still transiting on the train must be constrained. For example, if we assumed that train 105 starts at a location called F, and carries the FB train-block, then we would have one more variable $X_{105,\mathrm{FB,F,B}}$ and then if train 105 could never carry more than 15 cars, we would add $X_{105,\mathrm{FB,F,B}} + X_{105,\mathrm{AE,C,E}} \leq 15$. Ordinarily we would also add a constraint $X_{105,\mathrm{FB,F,B}} \leq 15$ that limits the size of train 105 when it departs F, but that is not truly needed here.

There are occasions where the capacity for the train size varies along its route, and this can also be handled easily.

There are many solutions for these constraints—there are seven variables and four equality constraints, leaving three degrees of freedom. In the next step, we use linear programming to pick an appropriate solution. We begin by demanding that if the capacities were not limiting factors then we should obtain the same solution as we would from the usual average day analysis. That means that we want the volumes per run of trains that emanate from a location to be the same. In this case, it means:

$$\begin{aligned} X_{101,\mathrm{AE,A,E}} &= X_{102,\mathrm{AE,A,B}} \\ X_{101,\mathrm{AE,A,E}} &= X_{104,\mathrm{AE,A,C}} \\ X_{105,\mathrm{AE,C,B}} &= X_{106,\mathrm{AE,C,D}} \end{aligned}$$

We now have seven variables and seven equations.

However, these four equations could lead to infeasibilities—if the capacities were the bottlenecks, then we would not expect all the volume to be equal per run—some volume would be moved from one train-block to another train-block with the same pick-up location. Rather, we create constraints that will lead the model to try to equalize volumes, but has flexibility to create unequal volumes if necessary. We write the constraints as:

$$\begin{aligned} &\frac{X_{101,\mathrm{AE,A,E}} + X_{102,\mathrm{AE,A,B}} + X_{104,\mathrm{AE,A,C}}}{3} - X_{101,\mathrm{AE,A,E}} = r^{+}_{101,\mathrm{AE,A,E}} - r^{-}_{101,\mathrm{AE,A,E}} \\ &\frac{X_{101,\mathrm{AE,A,E}} + X_{102,\mathrm{AE,A,B}} + X_{104,\mathrm{AE,A,C}}}{3} - X_{102,\mathrm{AE,A,B}} = r^{+}_{102,\mathrm{AE,A,B}} - r^{-}_{102,\mathrm{AE,A,B}} \\ &\frac{X_{101,\mathrm{AE,A,E}} + X_{102,\mathrm{AE,A,B}} + X_{104,\mathrm{AE,A,C}}}{3} - X_{104,\mathrm{AE,A,C}} = r^{+}_{104,\mathrm{AE,A,C}} - r^{-}_{104,\mathrm{AE,A,C}} \\ &\frac{X_{105,\mathrm{AE,C,B}} + X_{106,\mathrm{AE,C,D}}}{2} - X_{105,\mathrm{AE,C,B}} = r^{+}_{105,\mathrm{AE,C,B}} - r^{-}_{105,\mathrm{AE,C,B}} \\ &\frac{X_{105,\mathrm{AE,C,B}} + X_{106,\mathrm{AE,C,D}}}{2} - X_{106,\mathrm{AE,C,D}} = r^{+}_{106,\mathrm{AE,C,D}} - r^{-}_{106,\mathrm{AE,C,D}} \end{aligned} \tag{8.5}$$

$$\begin{aligned} r^{+}_{t,b,p,s} &\ge 0 \\ r^{-}_{t,b,p,s} &\ge 0 \end{aligned} \tag{8.6}$$

The variables $r^{+}_{t,b,p,s}$ and $r^{-}_{t,b,p,s}$ are the "residuals" and are non-negative. The difference $r^{+}_{t,b,p,s} - r^{-}_{t,b,p,s}$ represents either the positive or negative difference between the volume for train τ, block b, at pick-up location p and set-off location s and the average volume for all the train-blocks associated with block b at pick-up location p. The sum of the residual variables, $r^{+}_{\tau,b,p,s}+r^{-}_{\tau,b,p,s}$ is the absolute value of the difference. This is guaranteed if the linear program uses a positive cost coefficient for the residuals. For example, consider the following linerar program:

$$\begin{aligned} \min_{\{\{X_{\tau,b,p,s}\},\{r^{+}_{\tau,b,p,s}\},\{r^{-}_{\tau,b,p,s}\}\}} & r^{+}_{101,\mathrm{AE,A,E}} + r^{-}_{101,\mathrm{AE,A,E}} + r^{+}_{102,\mathrm{AE,A,B}} + r^{-}_{102,\mathrm{AE,A,B}} + r^{+}_{104,\mathrm{AE,A,C}} \\ & + r^{-}_{104,\mathrm{AE,A,C}} + r^{+}_{105,\mathrm{AE,C,B}} + r^{-}_{105,\mathrm{AE,C,B}} + r^{+}_{106,\mathrm{AE,C,D}} + r^{-}_{106,\mathrm{AE,C,D}} \end{aligned}$$

Subject to constraints (8.1)–(8.6).

The solution to this linear program appeared in the above table in the column “Capacitated volumes, no priorities.” If we ignored the capacity constraints (8.4) then the linear program obtains the usual average day solution that appears in column “Uncapacitated volume per run of the train.”

There are further refinements that need to be made:

1. Ensure that the linear program also gives results even if they break capacity limits. This is to ensure that the technique is robust—it will always provide solutions. This makes it similar to the average day analysis that will always provide a solution.
2. Train-blocks could have a priority, in keeping with the examples at the top of this section.
3. Develop a notion of “fill” blocks—blocks that could exceed their stated train-block capacity if there is still unused capacity on the train but all higher priority blocks are filled as much as they can be.
4. Allow capacity to be stated in both length and gross weight.

After this section we explore a more definitive and complete statement of train-block priorities and fill blocks which could be used for full simulation analysis. For average day analysis, we will only discuss a simpler version of priorities on train-blocks and fill train-blocks.

8.5.2.1 Achieving a Robust Train Volume Formulation

There could be examples where the capacity limits in the above described model will lead to an infeasible formulation. A simple example occurs if the total block volume of 70 increased to 101. Equation (8.2) would then change to:

$$101 = 5X_{101,\mathrm{AE},\mathrm{A},\mathrm{E}} + 2X_{102,\mathrm{AE},\mathrm{A},\mathrm{B}} + 3X_{104,\mathrm{AE},\mathrm{A},\mathrm{C}}$$

Since Eq. (8.4) puts upper bounds of 10 on all the variables, we have

$$5X_{101,\mathrm{AE},\mathrm{A},\mathrm{E}} + 2X_{102,\mathrm{AE},\mathrm{A},\mathrm{B}} + 3X_{104,\mathrm{AE},\mathrm{A},\mathrm{C}} \leq 50\times 10 + 2\times 10 + 3\times 10 = 100$$

Which implies that $5X_{101,\mathrm{AE},\mathrm{A},\mathrm{E}} + 2X_{102,\mathrm{AE},\mathrm{A},\mathrm{B}} + 3X_{104,\mathrm{AE},\mathrm{A},\mathrm{C}}$ could never be 101.

However, we could still want to produce volume estimates. That means that we need to make capacity limits soft constraints. For example, we could change $X_{101,\mathrm{AE},\mathrm{A},\mathrm{E}} \leq 10$ to $X_{101,\mathrm{AE},\mathrm{A},\mathrm{E}} \leq 10 + s_{101,\mathrm{AE},\mathrm{A},\mathrm{E}}$ where $s_{101,\mathrm{AE},\mathrm{A},\mathrm{E}}$ is a new, non-negative variable with a high penalty cost added to the objective function. It says that we could violate the capacity constraints for train-blocks, but only at a high cost of doing so.

Typically, the penalty cost for the s variables would vary depending on how the user wants to model train priorities and is discussed below.

8.5.2.2 Train-Block Prioritization

As discussed in the examples, capacitation for average day analysis usually has a "primary" train-block that should get as much traffic as possible until it hits its capacity limit, and a "secondary train-block" that accepts the overflow traffic.

As an example, suppose that train 101 carrying the AE block from A to E is the high priority train-block, and the other two train-blocks from A, train 102 carrying the AE from A to B and train 104 carrying AE from A to C, are the overflow blocks.

To implement this, we could change the first three equations in (8.5) from:

$$\frac{X_{101,\mathrm{AE,A,E}}+X_{102,\mathrm{AE,A,B}}+X_{104,\mathrm{AE,A,C}}}{3}-X_{101,\mathrm{AE,A,E}}=r^{+}_{101,\mathrm{AE,A,E}}-r^{-}_{101,\mathrm{AE,A,E}}$$
$$\frac{X_{101,\mathrm{AE,A,E}}+X_{102,\mathrm{AE,A,B}}+X_{104,\mathrm{AE,A,C}}}{3}-X_{102,\mathrm{AE,A,B}}=r^{+}_{102,\mathrm{AE,A,B}}-r^{-}_{102,\mathrm{AE,A,B}}$$
$$\frac{X_{101,\mathrm{AE,A,E}}+X_{102,\mathrm{AE,A,B}}+X_{104,\mathrm{AE,A,C}}}{3}-X_{102,\mathrm{AE,A,C}}=r^{+}_{104,\mathrm{AE,A,C}}-r^{-}_{104,\mathrm{AE,A,C}}$$

to:

$$\frac{X_{102,\mathrm{AE,A,B}}+X_{104,\mathrm{AE,A,C}}}{2}-X_{102,\mathrm{AE,A,B}}=r^{+}_{102,\mathrm{AE,A,B}}-r^{-}_{102,\mathrm{AE,A,B}}$$
$$\frac{X_{102,\mathrm{AE,A,B}}+X_{104,\mathrm{AE,A,C}}}{2}-X_{104,\mathrm{AE,A,C}}=r^{+}_{104,\mathrm{AE,A,C}}-r^{-}_{104,\mathrm{AE,A,C}}$$

And we would need to put a penalty on $X_{102,\mathrm{AE,A,B}}$ and $X_{104,\mathrm{AE,A,C}}$ in the objective function so that the solver will favor putting traffic on $X_{101,\mathrm{AE,A,E}}$ first, and using the other two train-blocks as overflow—ensuring both those train-blocks having the same volume if possible.

In accordance with having a robust formulation, there would be a relaxation of the constraints on the capacity of the blocks:

$$X_{101,\mathrm{AE,A,E}}\le 10+s_{101,\mathrm{AE,A,E}}$$
$$X_{102,\mathrm{AE,A,B}}\le 10+s_{102,\mathrm{AE,A,B}}$$
$$X_{104,\mathrm{AE,A,B}}\le 10+s_{104,\mathrm{AE,A,C}}$$

Generally, a higher penalty cost would be used for $s_{101,\mathrm{AE,A,E}}$ than for $s_{102,\mathrm{AE,A,B}}$ or $s_{104,\mathrm{AE,A,C}}$ to ensure that the primary train-block would not overflow if at all possible, but allowing the secondary train-blocks to overflow first if necessary.

8.5.2.3 Fill Blocks

A fill block is a lower priority train-block that has a capacity limit that is allowed to be exceeded as long as all the capacity requirements of the higher priority train-blocks have been met. In general, due to constraints (8.1) and (8.2), all traffic must flow on the train-blocks—no traffic is left behind. And hence there is no need to explicitly model fill blocks.

8.5.2.4 Capacitation by Length and Gross Weight

The prior discussion has focused strictly on cars, but we must consider length and weight capacities and allocations as well. The volume measurements on a block usually are calculated in cars per week, length per week, and gross weight per week. It is assumed that if, say, one-fifth of the weekly volume runs on a particular train-block, that ratio applies to all three measures.

For that reason we need to change the definition of the main variables and use the concept of ratios of train-block volumes to total block volume instead of the number of cars. Specifically, we let the main variables be $\left\{\tilde{X}_{\tau,b,p,s}\right\}$ where $\tilde{X}_{\tau,b,p,s}$ is the ratio of the volume assigned to train-block (τ, b, p, s) to the total block volume (in this example it was 70 cars). We need to redefine Eq. (8.2) which distributes out the block volume:

$$1 = 5\tilde{X}_{101,\mathrm{AE},\mathrm{A},\mathrm{E}} + 2\tilde{X}_{102,\mathrm{AE},\mathrm{A},\mathrm{B}} + 3\tilde{X}_{104,\mathrm{AE},\mathrm{A},\mathrm{C}} \tag{8.2$'$}$$

If for this block the 70 cars per week had an average length of 50 feet per car and an average gross weight of 80 tons per car, and if we wanted to limit the train-block to 750 tons and 600 feet, we would modify the first capacity constraint in (8.4) to be

$$\begin{aligned} 50 \times 70 \times \tilde{X}_{101,\mathrm{AE},\mathrm{A},\mathrm{E}} &\le 600 \\ 80 \times 70 \times \tilde{X}_{101,\mathrm{AE},\mathrm{A},\mathrm{E}} &\le 750 \end{aligned}$$

Other comments:

- The equations in (8.5) that compare the average train-block volume from a common pick-up location to the individual train-block volumes from that pick-up location could use an average weighted by the runs per week. In our example, instead of using

$$\frac{X_{101,\mathrm{AE},\mathrm{A},\mathrm{E}} + X_{102,\mathrm{AE},\mathrm{A},\mathrm{B}} + X_{104,\mathrm{AE},\mathrm{A},\mathrm{C}}}{3}$$

we might use

$$\frac{5X_{101,\mathrm{AE},\mathrm{A},\mathrm{E}} + 2X_{102,\mathrm{AE},\mathrm{A},\mathrm{B}} + 3X_{104,\mathrm{AE},\mathrm{A},\mathrm{C}}}{10}$$

8.6 Future Directions and Opportunities

Advances in performance and memory of desktop computers will allow simulation models to include capabilities in the future that will make the results more useful. These include:

1. Graphical representations of simulations—while this has been used effectively for modeling individual terminals, this has not yet been employed for network simulations. This capability will allow users to view the flow of trains across the network, build up of cars at terminals, and potential congestion points on main lines. This may make it easier for network planners to understand the impact of plan changes on the network as opposed to standard reports. This is especially true if the changes in the results from one scenario to another can easily be displayed graphically rather than in reports.
2. Incorporation of meet/pass planning logic—most network simulation approaches have considered running times between terminals as a constant or as a function of train size. They generally have not incorporated meet/pass planning logic to add delay to trains en-route due to the need to pull into a siding and allow another train to pass. With improved computer performance, it will be possible to include such complex calculations and still keep runtimes at a reasonable level.
3. Embedded optimization—while some network simulation models have incorporated relatively simple optimization tools like shortest paths to determine block routings, they have not incorporated more complex optimization models like those used to determine locomotive scheduling, crew scheduling, or terminal task scheduling. Typically, simple rules are used if these resources are modeled at all within the network simulation. With improved computer performance, it will be possible to incorporate more sophisticated models, such as those described elsewhere in this book, within the network simulation model. Not only will this improve the results from the network simulation model, but it may also provide a more powerful way to understand the true network benefits of using these kinds of optimization models.
4. Quicker evaluation of alternative operating strategies—with improved computer performance, network planners may become more interested in using network simulation models as a way to test out various operating strategies. For example, planners could change the blocks being made at various terminals, rerun the simulation model, and quickly identify what changed from the previous evaluation. While this can be done now, the faster the response time, the more likely it is that network planners will use simulation in an interactive mode.
5. Dynamic train operations with equipment cycling—current simulations focus primarily on the carload train operations, and do not examine the impact of equipment cycles on train operations. In reality, the majority of traffic on many railroads is carried in unit trains such as utility coal trains or grain shuttle trains, which are highly dependent on the cycling of equipment. For example, a grain shuttle train remains intact as it cycles between various grain elevators and unload points such as ports. Similarly, a coal unit train will cycle as a train set

between one or more mines and one or more power plants. To truly understand the resource requirements of a plan, there needs to be a way to translate a given level of demand (loads or tons to be moved) into a forward prediction of train movements and associated resource requirements (locomotives, crews, train sets). The specifics of the train schedules used in the simulation may not be as important as the resultant predictions of carrying capacity and resource needs.

Increased interest and demand for operating a scheduled, service-sensitive railroad will result in an increased reliance on the use of network simulation models to test out changes in operating strategies, blocking plans, train schedules, and terminal capabilities. Railroads will insist on having a deeper understanding as to how these kinds of changes will not only potentially reduce costs but may also impact service before they are enacted.

References

Allen WB, Mahmoud MM, McNeil D (1985) The importance of time in transit and reliability of transit time for shippers, receivers, and carriers. Transport Res B 19(5):447–456

Allman WP (1966) A computer simulation model of railroad freight transportation system. Proceedings of 4th international conference on operations research, Wiley, pp 339–351

Allman WP (1966) A network simulation approach to the railroad freight train scheduling and car sorting problem, Ph.D. Thesis, Northwestern University

Ballou RH (2004) Business logistics: supply chain management. Pearson Prentice Hall, Upper Saddle River, NJ, p 14

Bellman JA (1967) Railroad network model. Second international symposium on the use of cybernetics on the railways, Montreal, Canada. Proceedings edited by International Union of Railways (UIC), Paris, France

Hanes CA, Barnhart C, Johnson EL, Marsten RE, Nemhauser GL, Sigismondi G (1995) The fleet assignment problem: solving a large-scale integer program. Math Program 70:211–232

Ireland P, Case R, Fallis J, Van Dyke C, Kuehn J, Meketon M (2004) The Canadian Pacific Railway transforms operations by using models to develop its operating plans. Interfaces 34(1):5–14

Luo M, Meketon M (1997) A train scheduling model for reducing locomotive requirements. INFORMS National Meeting, San Diego

Martland CD (1982) PMAKE analysis: predicting rail yard time distributions using probabilistic train connection standards. Transport Sci 16(4):476–506

McCarren JR, Martland CD (1980) The MIT service planning model. Studies in Railroad Operations and Economics. Massachusetts Institute of Technology, vol 31 Cambridge, MA, USA

McGinnis MK (1990) The relative importance of cost and service in freight transportation choice before and after deregulation. Transport J 30(1):12–19

Oliver Wyman MultiRail Enterprise Edition, http://www.oliverwyman.com/insights/publications/2012/mar/multirail-enterprise-edition.html#.VFZwRfnF8pg

Petersen ER, Fullerton HV (1975) The railcar network model, Canadian Institute of Guided Ground Transport, Queen's University, Kingston, ON, Canada, CIGGT Report No. 75-11

Powell W, Bouzaiene-Ayari B, Lawrence C, Cheng C, Das S, Fiorillo R (2014) Locomotive planning at Norfolk Southern: an optimizing-simulator using approximate dynamic programming. Interf Art Adv (http://dx.doi.org/10.1287/inte.2014.0741)

Roberts K, Holcomb M (2012) Key factors and trends in transportation mode and carrier selection, pursuit. J Undergraduate Res Univ Tennessee 4(1):41–52

Siddiqee W, D'Esopo DA (1977) Computer-aided methodologies to develop blocking and train operations strategies for railroad networks. SRI International, Business Intelligence Program. Menlo Park, CA USA

Siddiqee W, D'Esopo DA (1975) User's manual for the network analysis computer programs. SRI International, Business Intelligence Program. Menlo Park, CA USA

Siddiqee W, D'Esopo DA, Tuan PL (1975) Blocking and train operations planning, SRI International, National Technical Information Service, 1975, USRA-R-106.1 Final Report, Accession No. 00129799. Menlo Park, CA USA

Stewart JC (1980) A decision support system for railroad freight operations management, M.S. Thesis, Technology and Policy, Massachusetts Institute of Technology, Cambridge, MA

Vaidyanathan B, Jha KC, Ahuja RK (2007) Multicommodity network flow approach to the railroad crew-scheduling problem. IBM J Res Dev 51(3/4):325–344

Van Dyke CD (1981) Microcomputers and the service planning model: designing a more useful tool for the rail industry, M.S. Thesis, Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA

Wilson PB, Hudson CJ (1970) Development, validation and application of the CN network model. Third international symposium on railway cybernetics, Tokyo

Chapter 9
Simulation of Yard and Terminal Operations

Roger W. Baugher

9.1 Introduction

In its simplest form, the operations of a railroad can be split into two disciplines: line-of-road operations and terminal operations. Some traffic, especially that moving in unit bulk, intermodal or automotive trains, will see few terminals between origin and destination, while traffic often termed "general merchandise" may visit several terminals en-route. Management devotes the largest amount of analytic effort to line-of-road operations, and tools exist for its analysis. Unfortunately, while terminal operations, including freight car classification activities, consume roughly 2/3 of railcar time—versus 1/3 for line of road—there are few tools to analyze terminal operations. This accounts, at least in part, for the limited capital investment railroads make in improving classification facilities. Most railroads and many consulting firms have staffs dedicated to using tools to analyze line-of-road operation and justify capital improvements; few similar efforts exist for terminal operations.

Why the discrepancy between analytic abilities for line of road and terminal? Fundamentally, line-of-road operation, while often difficult to optimize, is conceptually straightforward—establish a feasible meet and pass plan for a given set of trains on a fixed physical plant subject to constraints imposed by maintenance of way demands, operational control methods (signaling, e.g., CTC, ABS or dark territory), hours-of-service limitations, and other factors. The analytic scope is at the train level. Compare this to the terminal problem—establish a feasible train arrival, car inspection, car classification, train assembly, train departure, car repair, locomotive servicing and crew change plan for a given set of trains and a specified blocking

R.W. Baugher (✉)
Atlanta, GA, USA
e-mail: rwbaugher@aol.com

© Springer Science+Business Media New York 2015

B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3_9

plan on a fixed physical plant subject to constraints imposed by maintenance of way demands, operational control methods (e.g., hand-throw or remote control turnouts, remote control locomotive operation), work rule limitations, customer demand at industries serviced by that terminal, and other factors. The analytic scope is now on multiple levels—train, block, individual car, and possibly even individual resource level. Given the complexity, it is not surprising that development of tools for terminal analysis has lagged far behind the development of similar capabilities for line-of-road operations.

Consider the line-of-road and terminal operations visually. If a road operation is described as "single track Centralized Traffic Control with sidings," one need only know train schedules, running times and length and location of sidings to perform a basic analysis. Once a train has entered the line segment, it must follow a possibly large but well-defined sequence of moves before it exits the line segment. If, instead, a terminal facility is described as a "10 track flat classification yard," how much of its operations can be envisaged? As before, train schedules and track lengths are important, but the tracks' relative orientation may be the most critical. Each of the track layouts in Fig. 9.1, while matching the description, would have fundamentally different performance characteristics.

The analyst must also know full detail about the trains—the physical characteristics of each car and each car's destination—as well as the terminal's blocking plans—what function does this yard perform in support of the railroad's network, i.e., does the yard simply support local operations or does it have a role in switching cars moving through to other terminals? What resources, such as car inspectors and yard crews, are available by shift? Do work rules permit crew members to be used flexibly, and can switching be performed by locomotives remotely controlled by a crew member on the ground who can also line switches himself? No longer are we dealing with a well-defined sequence of moves as with the line-of-road operation, so process definition becomes far more complex, and analytic tools become more scarce.

9.2 Reasons to Simulate

Yard simulation, if it can be made cost effective, can serve many functions:

- Improve operations through training—with many yard personnel now hired off the street, it is important that basic skills and knowledge be learned before going on the job.
- Improve operations through improved processes—railroads seek to avoid capital investment by improving their procedures within existing facilities. Perhaps new methods of operation or addition of resources such as yard crews and car inspectors can be tested through simulation and found to provide sufficient benefits so that no capital expenditure is needed.
- Identify required capital investment—when possible, railroads will make modest investments in existing facilities. Simulation is needed to estimate the benefits derived from new or lengthened tracks, crossovers, improved car repair and

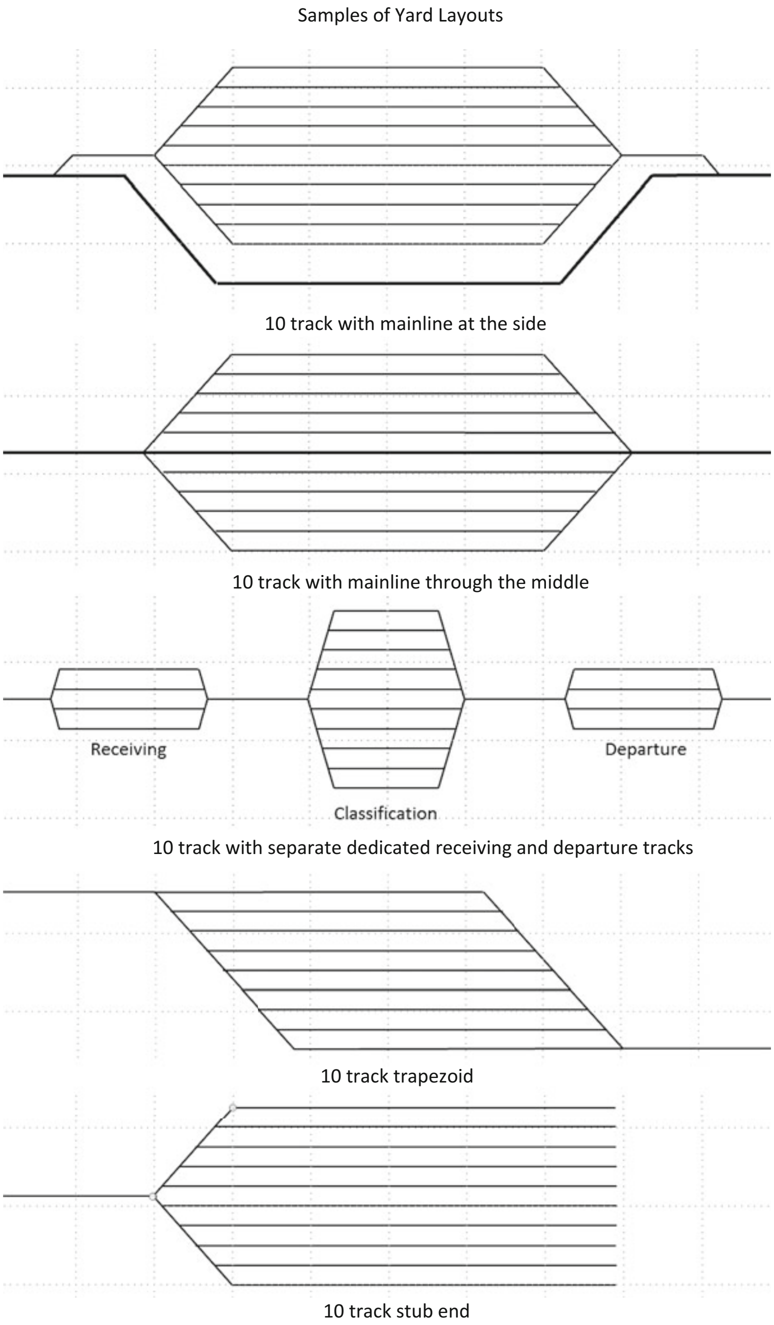

Samples of Yard Layouts
10 track with mainline at the side
10 track with mainline through the middle
Receiving
Classification
Departure
10 track with separate dedicated receiving and departure tracks
10 track trapezoid
10 track stub end

Fig. 9.1 Samples of yard layouts. 10 track with mainline at the side. 10 track with mainline through the middle. 10 track with separate dedicated receiving and departure tracks. 10 track trapezoid. 10 track stub end

locomotive servicing capabilities, and other modest investments. If large improvements in network performance are needed, larger investments will be required. In such cases, network models—not yard models—are generally used to identify the benefits that the network would experience after constructing the new facility. These models flow traffic over the railroad's network with and without the new yard, and identify the reduced car handlings, faster and less expensive switching, fewer miles, etc. derived from the investment. However, a yard simulation will still be needed to ensure that the new facility's performance is consistent with the performance parameters (e.g., blocks made, processing time, and costs) assumed in the network model.

- Evaluate train schedule feasibility—a railroad is a highly dynamic system with frequent changes in traffic patterns. This necessitates adjustments in train routes and train schedules, impacting both road and terminal operations. A frequent concern is that the existing operating plan—or an alternative under consideration—may call for a rate of train arrivals and departures at a terminal that exceeds the terminal's processing capacity. Yard simulation can then be used to determine whether the yard can process trains at a rate consistent with the desired arrival and departure schedules.
- Provide replay capability—well-designed yards will remain fluid under a variety of traffic and operating conditions, but problems can arise that overwhelm them, causing terminal congestion. Simulation can assist in two ways: design and evaluate recovery plans and identify the factors that caused the congestion in the first place.

9.3 The Problem

Let us start with a simple yard and understand the fundamental processes.

9.3.1 Train Arrival

Terminals often have little input on when a train arrives or the order in which they arrive, as these decisions are made by line-of-road dispatchers who give primary consideration to train priorities, schedule adherence, hours-of-service restrictions, and other line-haul factors. So, terminals seldom control the when and how of train arrival, but they are obligated to find a place to park inbound trains, clear of the main if possible. The terminal specifies the track into which the train will arrive, or a set of tracks if the train is longer than any single available track. In some cases, the terminal will be the destination of the train, and the entire train will be put away in the yard, but more commonly, the train will be picking up additional cars or setting out (i.e., dropping off) only a portion of its cars, so accommodations must be made for the through portion of the train as well as the cut (i.e., contiguous group of cars) to be picked-up from or set-out to the yard tracks. In general, picking-up cars is straightforward—separate the head end of the inbound train from the rest of the train at the point where the cars are to be added to the train, pull the head end forward, back to the cars to be picked up, couple and pull ahead, then recouple to the rear of the train. Setting-out cars can be more complicated. Depending on train blocking, yard configuration and other factors, the train may pull through a track to make a set-out at the rear of the train, or it may separate the head end from the rest of the train, pull the head end beyond the yard, and shove (i.e., push) a cut into a track.

9.3.2 Handling the Inbound Crew and Power

If the train is terminating, or if power is being changed at the terminal, the inbound power and its crew must be transported to a tie-up location—the engine house and/or locker room. This activity generally receives a high priority, because prompt repositioning frees resources for the next trip and avoids additional labor costs, especially final terminal pay. However, these considerations must be balanced against the likelihood that the movement of light power will interfere with classification and train make-up activities. Consequently, repositioning of power and crews may be delayed until higher priority activities clear the necessary route.

9.3.3 Inbound Car Inspection

When a car or cut of cars is separated from a train, the brake system will engage, setting the brakes on the cars. Before they can be switched, the brakes must be released on each car individually by a mechanical department employee who must walk the length of the cut of cars, pulling a lever that dumps the air pressure and releases the brakes. This same individual, or a pair of individuals walking either side of the cars, can also inspect the cars for mechanical damage, enabling the cars to be switched to a repair facility when they are classified. Each railroad, and often each terminal on a railroad, may have a different policy on the thoroughness of the inbound inspection, with some delaying a comprehensive car inspection until cars are being assembled into a train.

9.3.4 Switch (Classify) Cars

With the inbound power and crew clear and brakes released, it is time to sort the cars. If more than one track is ready to be switched, terminal management must decide at what time and in which order to switch the tracks. Several factors may influence the decision:

- Arrival order, e.g., First-In, First-Out.
- Traffic priority, e.g., intermodal/automotive versus general merchandise.

- Time to outbound train, e.g., which track has cars with the tightest connection?
- Track characteristics, e.g., is this track needed for a subsequent inbound train because it is the longest track?

Once the decision is made to switch a specific track, a yard engine will attach to the cut of cars and begin to handle them to different tracks based on their destinations. Physically, this is accomplished by shoving the cars, uncoupling one or more cars from the leading end of the cut, and allowing it to roll by gravity (a hump yard) or by momentum (a flat yard). While the process is conceptually simple, many complicating factors must be considered:

- If there are as many classification tracks as there are classifications to be made, then one track can be dedicated to each. However, block sizes and track lengths vary, so more than one track may be needed for large blocks.
- If there are more classifications to be made than there are classification tracks, cars destined to some tracks will need to be rehandled. Several blocks will be assigned to a reswitch or "sluff" track which must be pulled and switched again when time and space permit.
- Some tracks will be reserved for special functions. Cars requiring repairs will be routed to a bad order track, while cars missing movement instructions may be routed to a no-bill track.
- To ease train make-up, blocks destined to the same train should be assigned to adjacent or near-by tracks. This is especially important in larger yards where more than one job assembles trains, and their efficiency is tied to the ability to avoid conflicts.
- Yard layout may complicate the switching process, especially if some yard tracks are physically separated from others. This occurs, for example, when the main line runs down the middle of the yard, effectively creating two separate yards that must be coordinated. Since blocks are now spread across more than one set of tracks, only a portion of the cars can be switched directly to the appropriate track. Sluff tracks are again required, enabling cars to be held until they can be transferred to the appropriate portion of the yard. To minimize rehandling, the yardmaster will examine the consist of inbound trains, arriving them into the portion of the yard which best matches the destinations of the cars to the classifications made there.
- Time of day may affect which block is assigned to a track. If a block will move on a train scheduled to depart much later, the block may be switched to a reswitch or sluff track for later handling. Such reassignment is even more prevalent in yards that switch directionally—e.g., build eastbound trains half the day and westbound trains the other half. Such yards may even switch block-to-track assignments while a track is occupied with cars for the former block which have not yet been pulled for train assembly.

9.3.5 Train Assembly

Conceptually, train assembly is a straightforward process—pull cuts of cars from appropriate tracks and assemble them in the proper blocking order for the departing train. Consequently, it may be surprising that train assembly is often the largest bottleneck in a yard, especially in large facilities such as hump yards. Many factors must be considered for the process is to be efficient:

- Yard layout will dictate where and how blocks will be combined while building the train. If there is adequate headroom, the yard engine can pull from one track, and then shove to a coupling on one or more other tracks, a process known as "doubling." Where the headroom is limited, one track may have to be pulled at a time, with the train assembled on one or more dedicated departure tracks.
- In large yards where more than one yard engine is building trains, movements must be carefully choreographed to avoid conflicts between the engines. As noted earlier, assigning all of a train's blocks to adjacent or near-by tracks will cut down on cross-yard moves.
- While it is desirable to get all cars on the next outbound train, length and tonnage restrictions often necessitate that some cars be held for the following train. In such a case, getting the most critical cars onto the train is essential, and tracks may now be switched again to "cherry-pick" specific cars (i.e., pulling a cut of cars until the desired car is reached, then setting it aside for immediate processing).
- When assembling trains, one must be cognizant of "train-makeup" rules, which restrict where cars can be placed in a train. Some rules are tied to crew safety—cars carrying hazardous materials may be no closer than *X* cars from the locomotives and cannot be placed next to cars carrying explosives. Other rules are associated with managing in-train forces related to grades, curves, acceleration and braking, and take the form of limiting where empty, long or heavy cars can be placed in the train. Special yard engine moves are often required to assemble the trains consistent with these rules.
- Yards commonly have standards that specify how much before a train's departure the assembly should begin, with 3 h being a typical rule-of-thumb. However, if several trains are scheduled to depart within hours of each other, train assembly may have to begin earlier than desired and the process staggered among the trains to ensure that all are ready for departure on time.

9.3.6 Final Train Assembly

With all the cars for the outbound train assembled on one track (or more than one track if the train is too long for one track), the final train assembly processes begin. Mechanical forces must walk the train, lacing (i.e., connecting) the air hoses, testing the brake system, and making a comprehensive inspection to ensure that all cars are in safe working order. Cars found to be defective must either be repaired in place or switched out, possibly affecting timely train departure. Once road power is coupled to the train and road crews are aboard, the train is ready for departure.

9.3.7 Train Departure

Ideally, the assembled train is ready for departure at the time best suited for road movement, facilitated by communication between yard personnel and the road dispatcher. If cars are ready before a departure slot is available and if the yard needs the track space, cars may be advanced to a nearby siding by the yard engine, then picked up when the balance of the train can depart. Otherwise, the train will be held in the departure tracks until the dispatcher authorizes road movement. These delays can be costly, cutting into hours-of-service limits, incurring initial terminal payments to the crews and impacting crew and locomotive cycles.

9.4 Matching the Analytic Approach with Study Requirements

Simulation is a data-intensive, time-consuming process that should be avoided if simpler, more expedient approaches can provide the necessary results. Many types of studies—such as those analyzing traffic flows over an entire railroad network—do not require that individual processes in a yard be fully represented. Consider that, in its simplest form, a yard can be represented as a black box, with inputs, outputs and a specified processing time. In such a scenario, a cut-off will often be used, where any car arriving in the yard x-hours or more before a train's departure (the "cut-off" for that train) is assumed to make the outbound train. This approach mirrors current car scheduling systems, which do not account for capacity, scheduling conflicts or other important factors.

Similar approximations of yard performance are often used by line-of-road dispatching software to reflect interaction with yard operations. Since much track capacity will be consumed if a train is held on the main out of a yard because no track is available to receive it, such software assumes that a track will be made available for a subsequent train *x*-hours after the arrival of a train. This is obviously better than simply assuming the yard has infinite capacity to receive trains, but does not fully reflect actual yard performance.

Recognizing cut-offs' inherent disadvantages, researchers developed other approaches that overcame several of cut-offs' shortcomings:

- A cut-off assumes that all traffic arriving in the yard before the cut-off for an outbound train will make the train, while none of the traffic arriving after the cut-off will make it. This is clearly not realistic—on some days, cars arriving after the cut-off will make connections, while cars arriving before the cut-off will miss. This suggests that car connections are probabilistic, not deterministic.

- At any time, some cars in a yard become bad-ordered, lack movement instructions, are held until billing is resolved, are switched incorrectly, are scheduled to an annulled or size-restricted train, etc. These cars will remain in the yard until the physical or informational impediments to further travel are resolved. Thus, it is appropriate to introduce an upper limit on the percentage of cars that will make an outbound train irrespective of the cut-off time.

In the 1970s as part of a federally funded research effort, M. I. T. researchers developed an approach known as PMAKE analysis, which stands for the *P*robability that a car will *MAKE* a connection as a function of time. The function is piecewise linear as shown in Fig. 9.2 (the relationship can also be represented by a smooth logit curve). Unlike the cut-off, the probability increases with time to a maximum value PMAX, which represents the maximum percentage of traffic that will make its next connection. The slope of the line is specified by T50 and T90, where T50 is the necessary time for 50 % of cars to make their first available connection, and T90 is the additional time beyond T50 at which 90 % of cars connect.

This conceptually simple but powerful approach was integrated into a number of models. The Service Planning Model, released in the late 1970s and early 1980s, was one of the most successful network analysis tools of the period (Martland, Marcus, and Raymond, "Boston & Maine Achieves Control over Railroad Performance", INTERFACES 16:5, September-October 1986). Some analytic tools, like M. I. T.'s Intermediate Terminal Model, are founded on the PMAKE concept. To predict average yard dwell, time estimates for each yard process (i.e., inspection, switching, train assembly, etc.) are combined to form an overall PMAKE for the

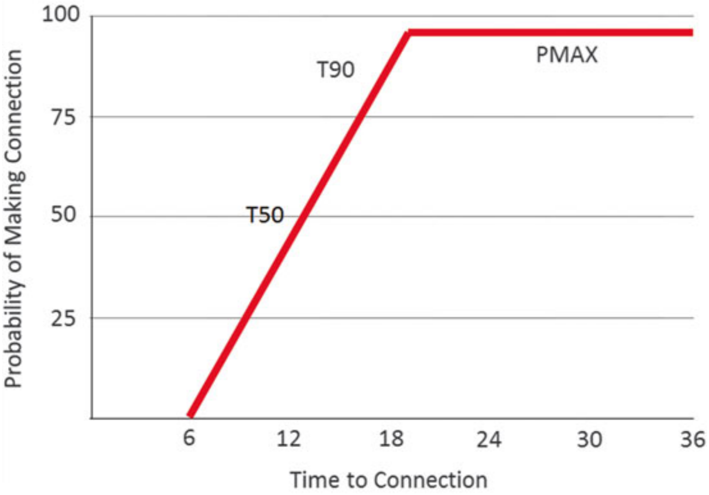


Fig. 9.2 PMAKE analysis

yard, with T50 being the sum of the mean times for each of the required processes, and T90 calculated using the standard deviations for the processing times.

One useful result of this approach is the ability to quickly estimate the reliability of trip times when a car moves through a sequence of yards. For instance, if all yards in the sequence have a PMAX of 90 %, then cars passing through 2 yards can be expected to achieve their trip plan standard 81 % of the time (0.9×0.9), while the reliability of a 3-yard sequence drops to 73 % (0.9×0.9×0.9).

The development of the PMAKE concept in the 1970s was driven, in part, by the failure of earlier network modeling software that had attempted a detailed simulation of line-of-road and yard operations in a comprehensive model. One such model was developed for the Association of American Railroads (AAR) by The Midwest Research Institute, and was installed at several railroads, including the Illinois Central (now part of Canadian National). Initially, yard processes—inbound and outbound inspection, picking up and setting out cars, assembling, classification and air pumping operations—were specifically modeled. Each used its own equation of the form $Y=A+BX$, where Y is the total time consumed for an operation, A is a constant time, B is a coefficient reflecting the time per unit of an activity, and X is the number of units of that activity. The functions also reflected the availability of yard resources, so the projected dwell time for a car in the yard was a function of both the number of jobs in the queue and the resource availability.

Unfortunately, this approach proved untenable, partly due to the computer limitations of the time, but mostly because it proved difficult to assemble data to allocate a resource's time to a specific activity. Instead, average times for these activities were substituted, and the simulation was completed.

On a historical note, several railroads, including the Canadian National and the former Chicago, Burlington and Quincy (now part of BNSF), built yard models internally, while some others, like the Baltimore and Ohio (now part of CSX) and Saint Louis— San Francisco (now part of BNSF) used commercial models produced by organizations like the Battelle Memorial Institute. At least one vendor, General Railway Signal, also developed a yard simulation model. It is believed that none of these models survive.

9.5 Building a Yard Simulation

9.5.1 Conceptual Design

Yard operations are typically modeled using discrete-event simulation. Initially, some exogenous event occurs—a train arrives into the simulation—which requires that a decision be made—route the train to Track 1. Once the simulation software executes this decision and the train comes to a stop in Track 1, a new event—train arrival—triggers the need for a new decision—uncouple the locomotives, route them to the engine servicing area and assign Car Inspector 1 to conduct an inbound inspection of the train. This event–decision–event–decision cycle is executed repeatedly to move cars from arrival through departure.

Yard simulation, then, can be viewed conceptually as the coordination of event simulation and decision generation and execution. The software could be envisioned as two discrete components, one we will call the "simulation engine"—capturing events, requesting decisions and executing those decisions—and the other we will call the "decision engine"—determining the next course of action. In practice, the two components are seldom separated and are often intertwined, but this view can provide useful insights for understanding the process.

9.5.1.1 Simulation Engine

From this perspective, the simulation engine can be either custom, dedicated software or a general purpose simulation package, especially if the software has been supplemented with special purpose rail libraries. Key capabilities of the simulation engine include:

- Ability to handle simultaneous operations over a large facility—many activities occur in parallel in a physically sprawling yard, requiring software that can manage large numbers of events and resources.
- Routing capabilities—yard movements involve often complex paths, with engines and cars moving forward, and then reversing repeatedly. Some tracks are clear for through movements, while others are blocked by standing cars. The simulation software must have an inherent ability to compute feasible, ideally optimal, routes, or the user will be required to manually prescribe large routing tables to perform this function.
- Managing conflicts—except for the simplest yards, there will be competing requests for resources, including track paths, yard engines and mechanical inspectors. The software must have the ability to resolve such conflicts. Resources can be provided to a requestor based on First-in—First-Out, relative priority or some other rule, and requestors not receiving the required resource must be delayed until the resource is available.
- Rich graphics—nothing sells a simulation better than snazzy graphics. The ability to reproduce physical movements—engines pulling and shoving cars, cars rolling into tracks—helps the developer ensure that the model is working properly and enables him to demonstrate to others that he has captured the problem correctly.
- Ability to compile a rich set of statistics at various levels—since yard simulation involves the movement of individual cars, statistics will be generated at this detailed level. However, statistics may be useful at a traffic type (i.e., intermodal, automotive, general merchandise), block, train, resource utilization (i.e., crews, yard jobs, bowl tracks, leads), and yard level.
- Ability to properly handle warm-up and shut-down periods—while it is possible to pre-populate a yard at the beginning of a simulation, it is common practice to start with an empty yard and arrive and depart trains and switch cars until a "steady-state" has been achieved, typically several days into the simulation. Similarly, during the shut-down period at the end of a simulation, no more arriving trains are processed, and the simulation ends when all cars have departed.

Obviously, neither period is typical and software should be capable of statistically excluding these periods from model results. Some simulation programs also enable a "warm start" where a simulation can be run until the steady state has been reached, and the model state recorded so that future simulations can start at that point, eliminating the need for simulating the warm-up period again.

- Probability distributions—simulations can employ fixed values, like a specific run time for arriving trains to enter the receiving yard track. Such values are often averages, meaning that there is a distribution about the mean. Simulations can be made more realistic if they rely on probabilistic rather than deterministic values to represent the start time and length of an activity, the number of resources available, and other factors driving the simulation.
- Interface with other software—much effort will be expended to define the track network, traffic data and operating plan, so the ability to import data from other systems is critical. The simulation will also generate volumes of output statistics, so interface with analytical tools is extremely useful. Integration with database engines and common management tools like spreadsheets is essential.
- Scenario management capability—most yard studies examine a set of physical plant and/or operating plan alternatives. Software must be able to manage multiple scenarios.
- Internal editors—data input, manipulation, editing and creation of alternatives is greatly simplified when editors are available within the simulation platform.

9.5.1.2 Decision Engine

More important than the choice of the simulation engine is the selection of a decision engine. Over the years, many different approaches have been used. One successful implementation, TRIM, developed by Canadian National and used by CN and CSX, relied entirely on human decision makers. The computer received and processed input data, pausing at each decision point for guidance from a staff of yard modelers, who then specified what actions to take. The computer then implemented the decisions and moved ahead to the next decision point. This approach produced results that could be easily understood since the process was completely transparent, but it had a number of shortcomings. These included long turn times (a study ran at three times real time, meaning one 8-hour yard shift required over 8-hour workdays to be analyzed), subjective results (only as good as the analysts running the model), large staffing requirements (eight or more personnel might be involved), and limited repeatability (too long to be repeated and too subjective to reach the same result). The software is no longer in use.

A better approach is replacing the external, human decision-making with internal, computer-based decision processes, which are often just sets of rules similar to those that the human operator might specify. Consider the decision as to which track should be used to handle an arriving train. When asked, the human operator may just specify the next available track—a simple rule that can also be coded into a computer to guide its decisions. This is the most common approach taken by yard simulation software, although the sophistication of the rules engine may vary widely among implementations.

Where are the majority of such decisions needed?

9.5.1.3 Inbound Process

- Train arrival sequence—If more than one train is approaching the yard, in what order should they be brought into the yard? Possible factors:
  - First-In/First-Out (FIFO)
  - Trains whose crews are about to exceed their Hours of Service (typically 12 h on duty).
  - Trains whose cars are most service-sensitive or whose connections are most in jeopardy.
- Train to arrival track planning—which track should receive the inbound train? Possible factors:
  - Inbound train direction.
  - Next available track.
  - Track versus train length.
    - Track that will hold entire train.
    - If no track will hold entire train, select tracks best suited for a "double" (splitting the train over two or more tracks).
    - Short trains can be added to tracks already occupied to maximize the length of the switch cut.
  - For split yards (e.g., yards with tracks on both sides of the main), arrive a train into the side of the yard better suited to handling the destinations in the train.
- Train to arrival track routing—if several routes exist, which should be chosen? Possible factors:
  - Shortest route.
  - Fastest route—e.g., route with fewest switches to be thrown.
  - Conflict avoidance—choose a route which does not interfere with other activities.

9.5.1.4 Switching Process

- Hump sequencing—which track should be humped/flat switched next? Possible factors:
  - First-In/First-Out.
  - Tracks whose cars are most service-sensitive or whose connections are most in jeopardy.
  - Tracks which, when cleared, are best suited to handle inbound trains.

  - Blocks being made—some yards block differently at different times of day, with some creating one set of blocks for half a day, and another set for the other half.

- Block to track assignment—which block should be assigned to which track? How should cars that need to be reswitched/rehumped be handled? Possible factors:

  - Blocks being made.
  - Track versus block length.
  - Arrangement of pull-down leads to departure tracks.
  - Placing all blocks for an outbound train on adjacent tracks.
  - Determination of which track or tracks best serve as sluff tracks. (These tracks will be used to hold cars awaiting billing, empty car distribution orders, etc., and those which need to be rehumped/reswitched.)
  - Splitting a large block over two or more tracks—this creates an opportunity to split a large block into several smaller blocks, reducing handlings at downstream yards.
  - Changing block assignments during the day in order to create more blocks—in this scenario, one track may hold more than one block, with the cars at the pull-down end of the bowl tracks in a block departing on trains earlier than cars at the hump/flat switch end of the bowl tracks.

9.5.1.5 Train Assembly Process

Pull down and train assembly sequencing—which tracks should be pulled and in what order? Possible factors:

- Block order on trains to be built (often referred to as the train's block standing order).
- Desired standing order of individual cars on assembled train (e.g., keep hazardous loads five cars from the locomotives).
- Train make-up restrictions that may require blocks of loads and blocks of empties to be placed at specific locations within the assembled train.
- Need to perform more blocking—at times, tracks may be flat-switched at the pull-down end to create more blocks.
- Presence of high-priority cars that require special handling. For instance, an outbound train with weight or length restrictions may not be able to handle all cars scheduled to it, so additional switching may be required to ensure that high-priority cars make the train, no matter their position in the bowl tracks.

9.5.1.6 Departure Process

Choosing an outbound route, if more than one exists—which route is best? Possible factors:

- Direction of outbound train.
- Conflicts with other activities.

Let us examine one of these decision points in greater detail and see how a set of rules can evolve from simple to complex.

As noted above, many factors can affect the assignment of blocks to tracks in the classification bowl. This produces a wide variety of possible decision rules:

- The simplest rule is to make no assignment at all. The simulation can simply monitor the number of blocks needed and the length of each; as long as the required track length does not exceed the actual capacity, the simulation produces feasible results. RULE = No Rule.
- For many yards, block-to-track assignment is quite fixed, so the decision rule is simply a set of fixed assignments. RULE = Block A → Track 1, Block B → Track 2, …
- Consider the yard layout depicted in Fig. 9.3 where all tracks are the same length. Here, there is no advantage to use of any specific track. RULE = Use next available track for the next block created.
- The yard layout in Fig. 9.4 is similar, but the switch lead now lines up with Track 1, making it the easiest, and therefore the most desirable, to use. RULE = Use Track 1, then use next available track.
- The yard layout in Fig. 9.5 is perhaps the most common in practice. Tracks are now of different lengths, so block-to-track assignment should consider their relative sizes. One rule is appropriate when all blocks can be held in at least one of the tracks: RULE = Longest block to longest track, next longest block to next longest track,… However, the situation becomes more complicated if a block must be held on more than one track. (Note that this also creates the opportunity to break up that large block into two or more smaller blocks that might enable fewer reclassifications downstream.) RULE = For blocks exceeding the length of any bowl track, select a pair or set of tracks whose combined length most closely

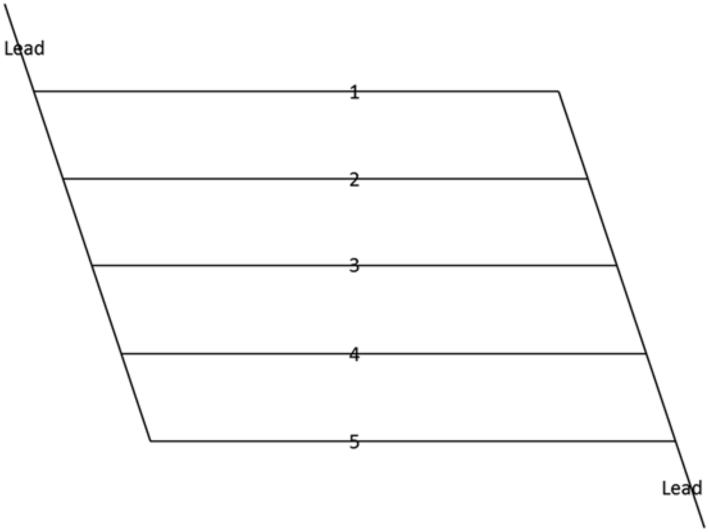


Fig. 9.3 All bowl tracks same length

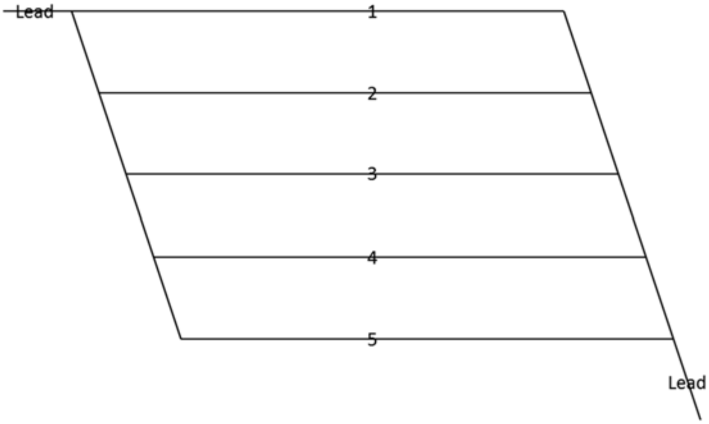


Fig. 9.4 All bowl tracks same length, lead aligns with track 1

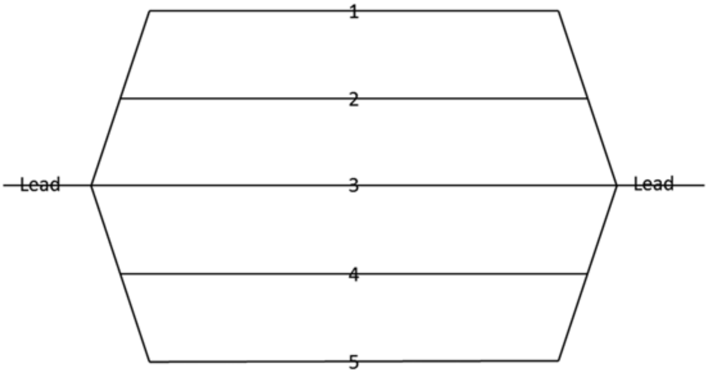


Fig. 9.5 Bowl tracks of varying lengths

matches the block length; for remaining blocks, longest block to longest available track, next longest block to next longest available track,… Note that more sophisticated rules can be devised to better handle this situation.

- Large classification yards with many bowl tracks will have multiple pull-down tracks like those depicted in Fig. 9.6. In many yards, the pull-down capacity is the constraining factor for yard throughput, as it directly affects the ability to efficiently assemble trains. Block-to-track assignment has to consider the factors noted previously (e.g., block length vs. track length), but must also consider that tracks holding blocks for the same train should be reached from the same pull-down track. This enables more than one pull-down job to work the yard at the same time, permitting more than one train to be assembled simultaneously. RULE = For each outbound train, select a "side" of the bowl; Build all blocks for that train on the same "side"; For trains with one or more blocks exceeding the length of any bowl track, select a pair or set of tracks on the chosen "side" whose

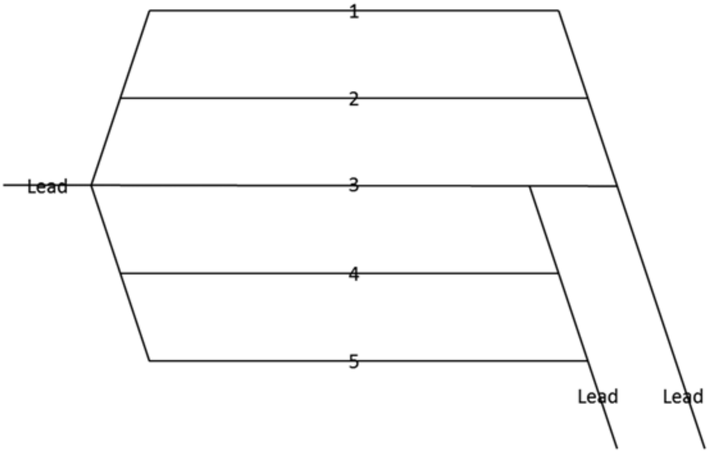


Fig. 9.6 Bowl tracks of varying lengths, multiple pull down leads

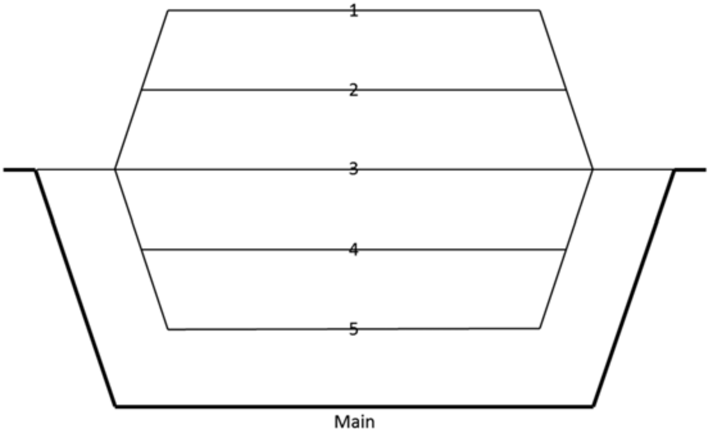


Fig. 9.7 Main line around yard, no dedicated receiving or departure tracks

combined length most closely matches the block length; for remaining blocks destined to the same train, longest block to longest available track on the same "side" of the bowl, next longest block to next longest available track on the same "side" of the bowl,... Again, note that far more sophisticated rules can be devised to better handle this situation.

- The layouts depicted so far have focused strictly on the orientation of the classification bowl tracks, assuming the presence of dedicated receiving and departure tracks to either side of the bowl. However, many yards are situated like the layout in Fig. 9.7, where there is no distinction between receiving, bowl and departure tracks. Here, any track can be put to any use. One must now establish a comprehensive set of integrated rules to manage all decisions, from train arrival through train assembly.

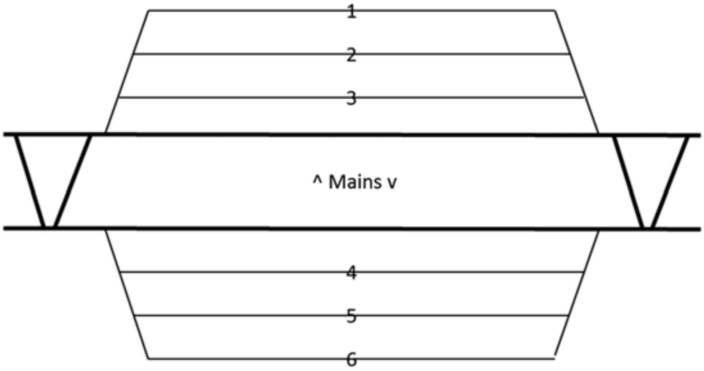


Fig. 9.8 Main line splits bowl tracks

- Further exacerbating the situation are yards like that depicted in Fig. 9.8, where the mainline splits the yard into two. This significantly complicates yard operation because:
  - Not all blocks will be created on both halves of the yard, so cars for blocks not handled on a side must be rehandled on the other side. This necessitates the creation of sluff tracks on both sides to hold cars for transfer to the other side for reswitching.
  - To minimize rehandling, inbound trains can be routed into the side of the yard whose blocking plan best matches the cars' destinations on the inbound train.

  In this scenario, the rule for block-to-track assignment must be paired with the rule selecting the receiving track for an inbound train.

9.5.2 Data for Simulation

Yard simulations require physical plant and operational data. The former can be derived from track charts, yard diagrams, geographical information systems, and even satellite images from Google Maps and other sources. Capturing operational data once required teams of Industrial Engineers riding jobs and pouring over written records, but new technologies ease the burden. Yard management systems record the arrival and departure of trains and monitor car location throughout the yard, equipment identification systems record the passing of cars, electronic control systems for controlling switches monitor track occupancy and switch position, and cameras record much of the activity. Yard locomotives are often equipped with GPS devices, although the frequency of location logging may be inadequate for simulation purposes.

9.5.3 Other Issues to Be Resolved

Even with the software chosen, there are major decisions to be made. Among these are:

- Physical limits of study area—consider only the yard itself, or include adjacent mainlines or other terminal facilities in the vicinity, like nearby intermodal or automotive facilities.
- Level of detail—determined by the nature of the study.
  - Training—full detail of all phases.
  - Operational analysis—sufficient detail to understand the area of interest.
  - Physical plant analysis—sufficient detail to understand the area of interest.
  - Network modeling—can approach black box.
  - Line-of-road dispatch analysis—can approach black box.
- Methods of calibration—comparison of model statistics with actual yard performance ensures that the model has captured reality well. In general, the yard's performance in the model will be superior to its actual performance, because yard operation in the real world includes derailments, maintenance activities, inaccurate accounting, engine and car failures, staffing problems and many other disruptions to ideal operations not captured in the model. Aggregate measures have been used for calibration purposes, including average dwell time in the yard and Right Train/Right Day (cars departed on the correct train on the correct day).
- Required level of interaction with line-of-road operations—while the focus of a study may be within a yard, movements on the adjacent main line will impact yard operations. The ability to fluidly arrive and depart trains requires that yard and line-of-road operations be coordinated. Hence, a yard simulation will have to model this interaction, at least at a basic level. To be fully robust, the model may have to include train dispatching logic to handle the routing of trains on the adjacent main lines and to and from the yard.
- Deterministic or probabilistic—if various alternatives are to be compared, the former might be the right choice; the latter is superior when a scenario has been chosen for detailed analysis and introduction of random factors enables the robustness of the operation to be measured.
- How to model each process—even if all processes are to be modeled in the simulation, the analyst must still decide how thoroughly the details of each process will be depicted. Consider:
  - When humping a cut of cars, will each car be individually modeled with a time processed, or will the entire cut of cars be recorded with its start time and end time only?
  - Must block-to-track assignment be modeled, or is it adequate to determine how many tracks will be needed to hold the blocks of cars and that number compared with the bowl track capacity as a whole?
  - Must activities on each track in the yard be monitored, or is it adequate to only monitor key tracks such as leads, ladders, main lines?

9.6 Recent Past to Current State of the Art

Before computers, yard analysis was a manual effort. Due to their complex and labor-intensive nature, such analyses were only performed when large investments were contemplated, such as the construction of a new yard. Many railroads invested heavily in new hump yards between the 1960s and the early 1980s, by which time computers had become available to assist with hump yard design, so yard models were built internally on many roads. Once the hump yard construction boom waned, models were seldom used, and few, if any, survive from that period.

At a minimum, these programs automated the data processing component of the simulations. However, they varied in the amount of internal decision-making capability each possessed.

On one extreme is the program TRIM mentioned earlier that relied entirely on interacting with human decision makers at every decision point. At the other extreme are programs whose decisions are made with comprehensive rule-based decision systems or even optimization.

At present, the author knows of several yard simulation efforts in the U.S.

Norfolk Southern Railway YardSim—beginning in 2008, Norfolk Southern embarked on the development of its YardSim software. Its research indicated that no platform or software package then existed on which to build, so it constructed its simulation program internally, relying on open source software when possible. The program includes four components:

- Rail Yard Simulator—simulates the events and activities in the yard; its components include:
  - Simulation engine.
  - 3D animation engine.
  - Process engine.
  - Statistics engine.
  - Library of business rules, heuristics, and algorithms.
- Rail Yard Editor—receives yard layout data from CAD systems and converts the data to the form required by the Rail Yard Simulator.
- Rail Yard Modeler—enables user to develop the operating policy and scheduling strategy; uses a role-based approach, where each role is modeled as an object with its responsibilities prescribed.
- Yard Scenario Manager—enables users to specify simulation inputs and configure simulation parameters; its "what-if" capability permits manipulation of:
  - Yard layout.
  - Yard resource availability.
  - Yard operating policy and scheduling strategy.
  - Traffic volume and train plan.

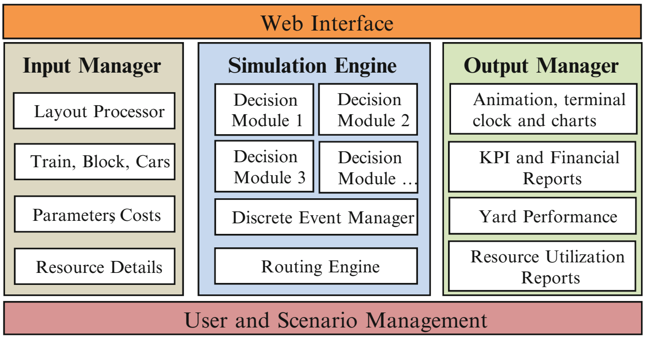
To supplement the rules-based system, NS devised heuristics to improve decisions in three areas:

- Hump sequencing—which cut of cars to hump next?
- Block-to-track assignment—which tracks are best suited for a specific block or set of blocks?
- Pull-down planning—what are the proper activities and what is the proper sequence to pull bowl tracks?

Mixed integer programming solutions were also developed in an effort to optimize the decisions in these areas.

To date, the model has examined the operations of a couple of hump yards on the NS system, and played an important role validating the design of the $100+ million expansion of Bellevue Yard in Ohio.

Innovative Scheduling (*Optym*) *YardSim*—working with CSX, Innovative Scheduling (now Optym) has constructed a web-based simulation system. The design emphasizes ease of data input (yard layout, operational data, parameters, resources by shift), realistic animation, and various output reports. A user can modify the yard layout by adding or removing tracks and analyze the impact in an integrated environment. A robust discrete-event simulation model has been adopted to ensure the portability of the model from one yard to another. The routing engine facilitating movements of locomotives and cars is independent of specific yard layout. Output reports are designed to provide easy visibility to core performance indicators as well as detailed performance reports by car, train, or specific resource as illustrated below.



Nine major hump yards, including a double hump yard, have been modeled using YardSim. The system has been validated for each yard using multiple sets of historical data, with values of primary key-performance indicators from the simulated model falling within 5 % of observed values. The system consists of 30+ decision modules and 50+ configurable parameters. This system is being used for what-if analyses of capital investment and operational case studies, and as such, is designed for use by Network Planning, Terminal Improvement Team, Service Design, Front-line managers, and Finance personnel.

AnyLogic—AnyLogic is a general purpose simulation language with the ability to perform discrete event, agent-based and system dynamics modeling. One of the features that distinguishes it from other software packages is its set of special-purpose libraries, one of which is the Rail Library, a suite of tools specifically designed to simulate railroad operations. The software inherently understands key elements of railroading—tracks, trains, cars, coupling, uncoupling, acceleration, deceleration, etc. It automatically develops routes for train movement, aligning switches as necessary. Internally, the software generates Java code, but the user's need to write code is minimized through various constructs within the system. Logic can be developed graphically by dragging logic blocks from a template to a canvas, setting the appropriate parameters and connecting the logic blocks to describe a process. Components that model state-transitions and process charts (i.e., flow charts) convert graphical representations of logic to code.

The author constructed a simple working model for the switching activities at a small flat yard, and has worked with a consulting firm that developed and commercially deployed both a comprehensive yard model and a basic line-of-road dispatching model using AnyLogic. With the application of other special-purpose libraries—the Road and Traffic Library and the Pedestrian Library—AnyLogic can perform detailed simulations of large, multi-purpose rail yards, modeling, for instance, the train, crane, truck and crew movements within a major rail intermodal hub. Given these capabilities, AnyLogic has been acquired by four North American railroads to support their simulation efforts.

With its ability to construct agent-based models, AnyLogic may even enable development of new yard modeling paradigms. Instead of a strict discrete-event approach, what if resources were treated as agents that interacted with events seamlessly? For instance, a mechanical inspector could be an agent looking for work to be done, instead of waiting for the discrete-event engine to call him. Such an integration of simulation approaches may provide new insights.

The success of yard simulation at one road has inspired the development of more advanced software for yard simulation and management, and coordinates with line of road operation. The road is working on a mixed integer programming solution to the "yard problem," finding that set of feasible decisions that maximizes an objective function reflecting efficiency, service quality, and cost.

9.7 Future Directions

One of the most exciting possibilities for yard simulation is developing real-time or near-real-time capabilities. Imagine providing yard management and service planners with a tool to play out a current scenario or to replay a recent scenario to see how different decisions would have improved the outcome. Such a tool may simply play out a set of manually specified decisions, might rely on optimizing software to provide superior solutions, or might use a mix of human- and computer-derived decision rules.

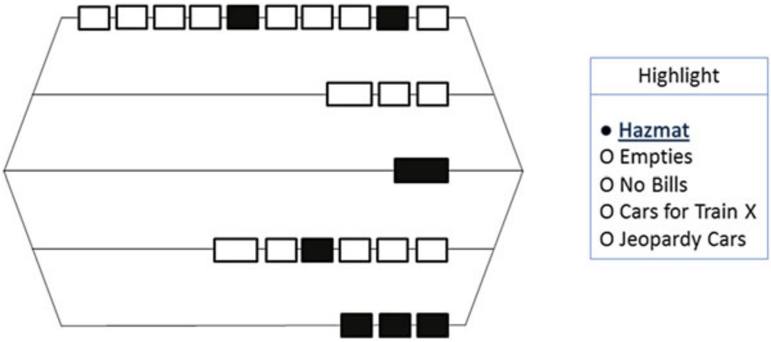


Fig. 9.9 Visualization of yard inventory and status

Today's yard managers rely on a written or verbal turnover report at shift change to inform the incoming staff of the current status of the yard. What if that report took the form of a visual depiction of the yard's status? The system would convey the necessary turnover information, while providing a platform for additional capabilities. The underlying database would contain real-time information on cars, blocks, tracks, trains, yard jobs, mechanical forces, etc., which could be queried to provide insights into the yard's health. Results of the queries could be displayed graphically, highlighting, for instance, all cars in the yard containing hazardous materials or destined to a specific outbound train. Figure 9.9 shows what such a capability might look like.

With that platform, one could easily create the ability to "see into the near future." Simulation software would take in the current yard status, schedule known events (e.g., train arrivals, train departures) for a specified planning horizon and use an established set of rules and known resource levels to predict what the yard would look like in 1, 2, 3 or more hours into the shift. If the simulations forecast that yard conditions will worsen, alternative decisions could be tested. If a scenario can be turned quickly, it would be possible to introduce probabilistic distributions for train arrival times, process durations, and other factors so that the robustness of the solution can be ascertained.

A modest extension of this platform is the ability to replay the activities and decisions of a previous period. Many yards perform poorly at least part of the time, so yard management and service planners need to be able to examine how the yards got into trouble and identify changes that can ameliorate or prevent future problems. Once the user chooses a specific start time for the simulation and builds the initial conditions in the yard, the simulation plays out known events like train arrivals while enabling the user to specify the decision rules, resource levels, process rates, and other variables that affect yard performance. If a set of variables enables the yard's collapse to be avoided, they can be used to specify best practice. Testing a wide variety of conditions—seasonal peaks, service disruptions, etc.—enables one to develop a set of procedures to respond quickly to those conditions before yard performance worsens.

Ideally, the planning horizon can be extended beyond a few hours to one or more days out. Unfortunately, this is not currently feasible. Car scheduling systems which provide forecasts of train consists are not capacity-constrained, so trains of unreasonable lengths can be scheduled into the future. Yards with a large amount of traffic received from another railroad or released by local industry do not know what traffic will be received or released a day from now. Until superior forecasting abilities become available, the planning horizon for real-time yard simulation will be limited to one or two shifts.

Chapter 10
Operations Research in Rail Pricing and Revenue Management

Michael F. Gorman

10.1 Introduction

10.1.1 U.S. Freight Rail Pricing History

Market-based pricing decisions in U.S. freight rail are relatively new. Prior to 1980, U.S. freight rail prices were closely regulated by the federal government, leaving little flexibility for railroads to manage prices, and to a large degree handicapping the railroads financially. Since 1980, railroads have undergone an operating renaissance, rationalizing both physical plant and operating plan efficiencies. At the same time, while still obligated to "common carrier" commitments to continue service, railroads have been freed to price services according to market forces.

Interestingly, despite the Staggers Act's removal of rate ceilings, real rail rates have fallen over the last 40 years due to rail mergers, rationalization of rail capacity, and more efficient operations. (For more, see AAR 2013.) For many years after deregulation, railroads were massively over capacity with multiple parallel rail lines. Essentially, all incremental traffic that could be moved cost effectively was helpful to profitability. However, reducing parallel routes, falling real prices, and increasing traffic have created a rail network that is tightly constrained and congested. For this reason, careful pricing of existing and potential services is essential to modern rail performance.

M.F. Gorman (✉)
University of Dayton, School of Business,
Department of Operations Management, Dayton, OH, USA
e-mail: mgorman1@udayton.edu

© Springer Science+Business Media New York 2015

B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3_10

10.1.2 Revenue Management for Rail: Importance

Freight rail revenue management is an analytical approach to pricing services that provides the maximum return from a fixed network of assets. In a sense, pricing action can be used to shape demand to better match the physical capacity of the rail network. In order to successfully implement a revenue management program, a railroad must understand both the behavior of its customers, and have a good definition of its capacity to provide service.

Because of the relatively high physical plant costs of rail (yards and rail lines), the economies of scale of individual trains (locomotives, crews) and relatively low marginal costs per load on an individual train (fuel and car), rail capacity is generally scarce even though the marginal cost per unit is relatively low. It is incumbent on railroads not only to allocate this capacity carefully so that service commitments can be met, but also that the available capacity be priced in a way that maximizes the return on investment in rail assets. In this environment, prices can be used to attract or deter incremental business, as deemed appropriate.

Because of the high fixed cost structure of rail, high volumes and high capacity utilization are mandatory for earning a reasonable financial return on the network. In this situation, the pricing decision must consider multiple capacity constraints and network effects of the pricing decisions of each service. Railroads revenue management decisions must sell the right combination of products, to the right customers, at the right time to maximize the return on the rail network. In some cases, incremental traffic can be very profitable to service as incremental traffic on existing trains, costing only additional fuel and equipment. The marginal costs of these loads are quite low, given the train is going to run in any case. These loads on average will lower the average the fixed costs of crew, line capacity consumption, and to some degree locomotives. In other cases, incremental traffic can be very expensive to service when the result is incremental trains which create demands for expensive locomotive, crew, line and yard assets, increasing average costs of service and reducing profitability.

10.1.3 Revenue Management for Rail: Challenges

Rail revenue management is similar to other industries with high fixed and low marginal costs, such as airlines, hotels, cruises, and the like. Given some relatively fixed capacity, appropriate pricing can maximize the return on that capacity. Unlike those industries, however, rail network capacity is considerably more flexible, both due to variable train sizes as well as the North American practice of highly flexible train services. Capacity is measured in a number of dimensions, such as car, locomotive, track, train, and yard. Further, that capacity is shared among competing business units. So, the unit of capacity is more difficult to define than, say, in the hotel industry where a room is a clearly defined unit of capacity.

As a result, pricing in the rail industry is very complex. First, the network is shared by numerous distinct traffic types: Intermodal, mixed merchandise, coal, grain, and automotive. Each has its own traffic patterns and seasonality, business shipping patterns and volumes. The vast majority of rail business moves on shipper-specific, medium-to-long-term contracts over which prices are predetermined. The structure of these contracts varies by traffic type; for example, intermodal tends to be shortest term, with a fair portion moving on spot market prices, and automotive and coal tends to be longest term. For this reason, for these lines of business, intermodal is perhaps the most conducive to revenue management techniques, followed by mixed merchandise. Coal, grain, and automotive have little short-term pricing flexibility, and capacities are often fully utilized in "unit" train operations.

To complicate matters further, many freight rail movements occur over multiple railroads, which then split the revenue according to predetermined agreements, limiting pricing flexibility and increasing revenue management complexity by requiring multiple railroads to agree on appropriate pricing.

Unlike airlines and hotels, railroad capacities are somewhat flexible, and fungible between product lines. In the hotel industry, an empty hotel room in one city cannot be substituted for a shortage in another; however, railroads can reallocate capacity between product lines. For example, an additional train could be created to provide service to excess demand in one product line and corridor, at the expense of capacity in another area of the rail network. Conversely, if demand is low, a train can be annulled, saving the capacity for other lines of business. In its simplest and most extreme form, railroads run trains only when they reach critical scale, known as a "tonnage" operation, and customer demand (not asset supply) determines the train operations. As such, a railroad's flexibility in service provision, traditionally viewed as an advantage to profitability and efficient operations, is a challenge to the use of traditional revenue management techniques which are generally based on the assumption of a fixed capacity.

Because of the complexity of pricing contracts, definition of a single unit of capacity, flexible operations and shared revenue between railroads, pricing decisions are extremely difficult. As a result, operations research models have not addressed revenue management in rail extensively; rather, such models have focused on asset capacity management given a level of demand, rather than transforming demand to better fit the capacity of the rail network.

10.1.4 Revenue Management for Rail: Recent Opportunities

More recently, as rail networks grow and become more complex due to mergers, ad hoc operations become progressively more challenging to manage. Downstream repercussions in the network from various on-the-fly service decisions can be unintended negative results. Further, customer service is more variable in a tonnage operation, making aggressive service-based pricing more difficult. As a result, more railroads are tending toward running a (more) regular schedule with fewer extras and annulments, improving the feasibility of leveraging revenue management methods.

Railroads can establish revenue management strategies that balance the various needs of the different lines of business across the asset classes. Because the problem is relatively complex and ill-defined, most efforts in this area have been limited to one or a few dimensions on the problem, and few have been successfully implemented. Further, there are relatively few examples of advanced methods in revenue management in practice in U.S. freight rail. The remainder of the chapter describes progress and opportunities for revenue management in freight rail.

10.2 Analytical Techniques in Freight Revenue Management

Multiple analytical techniques have leveraged railroads' efforts to improve revenue management in the last 30 years since deregulation. In most cases, for simplicity, the pricing methods are applied to a single or a few dimensions of rail capacity. The review here focuses on US freight rail; for further reading in passenger and non-US applications, the interested reader is referred to Armstrong and Meissner (2010).

There are two major components of effective revenue management: Understanding customer behavior (level and price sensitivity of demand), and defining rail capacity (in a number of dimensions). The remainder of the chapter follows this organization; revenue management methods will be covered according to the following dimensions of the problem.

- Characterizing customer behavior: Estimating product demand.
  - Forecasting demand levels—prediction of traffic volumes and changes in demand due to exogenous factors.
  - Predicting customer price sensitivity—changes in volumes as a result of changes in price as a function of competitive factors.
- Characterization of rail capacity.
  - Train capacity—given train plan, selling available train capacity.
  - Equipment capacity—railcars and containers.

10.3 Characterizing Customer Behavior: Estimating Product Demand

A critical component of revenue management is understanding the underlying customer demand behavior. There are two primary measures of customer demand: Estimated response to price changes of the railroad, known as price elasticity, and any other factor that affects customer demand levels, such as seasonality, competitor prices, changes to industrial production levels or international exchange rates given some price level, known as demand forecasting. Demand forecasts based on exogenous factors assume no pricing action on the part of the railroad. As price decisions are endogenous to the railroad's decision options, such changes are not subject to statistical methods and not included in firm-level demand forecasting. However, customer reactions to such price changes are not controlled by the railroad, and important to understand. Typically, but not always, statistical methods are used to estimate these customer behaviors.

There is a third component of customer behavior that has been studied as well, customer service elasticity. These lines of study look at the gain or loss of freight as the frequency, speed, and reliability of service improves or falls. With respect to traditional demand curve estimation, service level is a product attribute that can shift the demand curve just as a competitor's price change can. Because rail prices tend to be somewhat difficult to modify in the short run, an alternative is to modify service in order to attract or deter traffic. In so doing, costs and revenues are affected through service elasticities. The reader interested in a detailed review of service elasticity is referred to Small and Winston (1998).

10.3.1 Forecasting Demand Levels

Having anticipated customer demand as a function of exogenous factors and historical patterns is critical to revenue management. The demand forecast creates a baseline traffic level from which to plan capacity allocation and necessary pricing actions to affect change in anticipated traffic levels. As important as this demand forecast is, very little applied published research has been conducted on the subject in a practical, railroad-specific setting. Researchers have focused more on aggregate models of demand and modal split of freight flows at an aggregate level, which is not useful at a microeconomic level. Statistical methods, inventory-based methods, and utility maximization methods have been proposed and tested for projecting the total rail freight flows for a month or quarter based on the truck-rail modal split for broad geographic regions. However, such aggregate projections do not help the decision on how to price specific products, origins and destinations or days of the week. Thus, research to date on demand forecasting is of little assistance for revenue management. The interested reader is directed to Clark et al. (2005) for a survey of these methods.

Demand forecasting to support revenue management decisions is decidedly more practical in nature. A typical forecast will predict the number of shipments at varying degrees of specificity in geographical, commodity, and time period definition. For example, a forecast might project the number of shipments of a specific commodity or commodity group from an origin region to a destination region on a certain day, or even time of day. The level of specificity of a forecast is an important decision. The more specifically defined the demand, the more series there are to estimate, and generally the less accurately they are estimated. Often, the level of specificity of the forecast relates directly to dimension of capacity of interest (e.g., car, yard, train). For example, to forecast equipment needs, the type of commodity must be specified in order to assess the type of car needed. However, for predicting locomotive needs, a railroad need only predict the total tonnage between yards, which is far less specific a demand forecast. In any case, to support precise revenue management decisions, a large number of demand series must be estimated.

In practice, demand forecasting is done statistically using traditional time series and regression-based forecasting methods. Of course, such statistical models at a detailed level of disaggregation are fraught with error. Railroads, unlike airlines and hotels, do not have reservations and thus have little forward knowledge of customer orders, so forecasts are more uncertain in freight rail. ("Car orders", or requests of shippers for an empty car to enable a future shipment is one forward indicator in advance of a shipment, but it does not require shipment on a certain date.) This deficiency of forward information has been another impingement on the progress of revenue management in freight rail. As a result, many pricing initiatives are considerably more aggregate and less precise in their efforts to define capacity and demand. For example, pricing may be applied at the train level to line capacity based on an average day, or price for a particular service may be based on quarterly demand.

10.3.2 Predicting Customer Price Sensitivity

The demand forecast becomes the baseline for estimating asset requirements to provide services. As the service requirements based on demand forecast are established, areas of capacity shortages and excess in the rail network can be identified. Pricing actions can incent or discourage additional shipments. But it is critical to estimate how much customer shipments will be affected by a change in price, which requires an estimation of customer price sensitivity, known as demand price elasticity.

Early efforts focused on statistical methods to estimate demand. For example, Winston (1983) describes demand and elasticity estimation for transportation demand. There are many other such works. While not explicitly related to revenue management, understanding transportation demand is a critical input to appropriate pricing. Most of these manuscripts use statistical methods such as regression to characterize demand.

Similar to demand forecasting, a significant challenge for estimating demand sensitivity is determining the level of fidelity to conduct a study. On one hand, customer-specific response to a specific origin–destination pair (and possibly day of week, time of day, and level of service) is of utmost interest, but such detailed data are sparse and unreliable. For this reason, many studies utilize more aggregate demand for which sufficient data density exists. For example, an elasticity study might look at monthly flows between regions. However, pricing action is often at a more specific product level.

Recently Gorman (2005) proposed a practical alternative based on an optimization-based alternative to statistical methods, based on optimization methodology, and observed profit margins in a market. The reasoning goes as follows: Railroads earn higher margins on products where demand is more inelastic, thus, the current margins are an indication of customers' price elasticity. This "implied elasticity of demand" has the advantage of having relatively low data requirements: only current price and current marginal cost. Railroads typically cannot attain a marginal cost, so an average variable cost is substituted. Gorman (2005) compares calculated implied elasticities with the intuition of market managers, and finds the estimated elasticities largely agree. The approach thus provides plausible elasticity estimates which are consistent with existing market prices. The method is dependent on some level of rational pricing; to the extent that current prices are severely suboptimal, the resulting elasticity estimates are biased. Further, the method also does not predict future levels of demand due to market forces; it is meant only to predict market response to changes in prices.

10.4 Research in Revenue Management Models

Given an estimate of shipper demand and its sensitivity, revenue management models attempt to maximize the return of some dimension of capacity. As noted, because of the uncertainty of demand and the flexibility of supply, relatively little work has been done in the area of revenue management in freight rail. What work has been done is primarily with respect to a single dimension of capacity to allocate, such as train space or container allocation. Some work has focused on service sensitivity of customers.

10.4.1 Train and Block-Based Capacity Approaches

The first freight rail-based researches to consider revenue management explicitly were Campbell and Morlok (1994) and Campbell (1996). Not surprisingly, this research converted the relatively successful approach of revenue management in the U.S. passenger airline industry and adjusted it to freight rail.

This research assumed a fixed train network with known capacities. The train network was assumed feasible with respect to other dimensions of capacity such as yard and line. Customer demand was assumed known and deterministic, and prices predetermined. As such, the revenue management model was based on deciding which set of customers who share these trains (i.e., intermodal, general carload) to provide service in order to maximize profits. The challenge of this model was to trade off various services, given the capacity each customer consumed on a sequence of trains. As in the airlines, customers of different origin–destination pairs share capacity of intermediate trains; thus, the decision variable is in determining "block" capacity (allocation of train capacity by origin and destination of the block network) to be allocated amongst various origin–destination pairs. The model aims to maximize the expected profits of a set of blocks, given total block sizes do not exceed train capacity or customer demand levels. As in air revenue management, the complication arises in the complexity of defining a shared train network amongst blocks of different origins and destinations. In this model, capacity and block routings are fixed, and customer demand is accepted or rejected based on capacities. The work was never directly implemented by a major railroad.

However, in a simpler and more applied setting, CSX transportation experimented with a simple form of train-centric revenue management in the early 1990s. Several routes or trains were identified where the average tonnage and length of the trains were well below what could be handled with a standard locomotive consist. However, these trains were critical enough to the network that they were operated more than 5 days a week. The idea was to see if business could be attracted to these routes or trains by offering reduced rates to new customers who would move their freight only on these lanes.

To test this, the Operations Research group within the Service Design department developed flow maps of all the routes that regularly had at least 20 % available capacity using a standard locomotive consist. These maps were updated on a monthly basis and presented to the Sales and Marketing departments as well as the Finance department. In fact, there was a "Yield Management" team composed of members from Sales and Marketing, Service Design, and Finance that met each month to review progress and the most recent flow maps.

Recognizing that the key fixed costs of operation, crew and locomotive costs, would be incurred by these trains whether or not additional, incremental traffic was generated, Sales and Marketing teams were given reduced rates for these lanes that they could use to develop new business. New business was targeted because there was a goal to not diminish the revenue received from existing customers on these trains. No long-term contracts could be entered into with these new customers, since the space availability on the existing train service could not be guaranteed into the future. The team recognized that if the incremental business caused the need to add new train service, then the economic assumption that the fixed costs were being covered by existing business would no longer be valid.

This process was initially successful in generating some new business on these lanes. However, a problem was soon identified. Reports that were run each quarter to identify the profitability of customers used a process where the operation costs, including crew and locomotive costs, were allocated to *all* customers on a tonnage basis whose freight was handled on a given train. For the trains that were included in this initial Yield Management test, the impact of this cost allocation process was to make the existing customers look even more profitable since they were being allocated a smaller amount of the fixed operation costs than before. It also had the impact that the profitability of the new customers who were brought in under the Yield Management pricing program looked abysmal since their rates were not designed to cover any of the fixed costs of the trains, and yet those costs were being included in this profitability analysis. The Yield Management team tried to have the algorithm by which costs were allocated changed, but without success. Given that salespeople's compensation and bonuses were impacted by these "profitability reports," sales that were based on the Yield Management pricing model soon dried up and the program was dismantled. Yield Management became linked to price cutting and unprofitable customers, and had a poor reputation within CSX for years to come.

More recently, a project with Amtrak used yield management techniques for the combined passenger and vehicle in Amtrak's "Auto Train" service product (Sibdari et al. 2008). This project is somewhat unique in freight rail both because of the joint passenger and vehicle decision, and because Amtrak's schedule and capacity is more fixed than the typical US freight railroad. They describe a discrete-time revenue management model for the single-leg Auto Train and evaluate three different heuristic solution methodologies. Reportedly, this approach is in use in the Amtrak revenue management department.

10.4.2 Service-Based Pricing Strategies

Kraft (1998, 2002) and Kraft et al. (2000) suggested another approach for rail revenue management. Railroads do not have fare classes as do the railroads, and prices are difficult to adjust; thus the approach has short comings because the rail industry is inherently different and more flexible than airlines in its capacity allocation. This line of research develops a multi-commodity network flow approach, where each shipment is a separate commodity. The model allocates potential demand over a number of different service options given a train network, maximizing expected revenues. Rather than allocating block capacity among customers, customers are assigned to different blocks based on their expected willingness to accept different service times. Critical to the approach is the assessment of the probability that a shipper will accept the service level of various routing options. As a result, demand is shaped by adjusting service levels in a way that is consistent with the train service network.

Other service-based models evaluate revenue implications for railroads. Strasser (1996) evaluates the development of a service-differentiated intermodal rail network and pricing impacts. This research suggests that service differentiation helps to enable revenue management strategies by allowing differentiated pricing by service type. Other thesis work (Nozick 1992; Kwon 1994) has considered service implications of network design and revenue implications. These projects are tangentially related to revenue management, but like the other service-centric projects described above, none are implemented in the U.S. rail industry.

10.4.3 Container-Centric Yield Management

In intermodal networks, container capacity availability and balance is a potential source for revenue management-based approaches. As discussed in the railcar management chapter of this book, empty repositioning is an integral part of service provision for railroads; cars and containers must be moved to where they are needed. Over the span of a month or a quarter, "strategic" container flows management must be balanced either through customer orders, or costly equipment repositioning. In the shorter term (e.g., 1–3 days), inventory of available containers in a geographic region must be "tactically" allocated profitably among tendered loads. Strategic and tactical container-centric revenue management approaches are described below.

From the railroad's perspective, some applied work has been published with container-centric revenue management objectives. Gorman (2001, 2002) discusses the use of pricing to help obviate such imbalances for BNSF intermodal. Instead of repositioning empties in an optimal way given an imbalance, pricing action can be taken to help balance the network in a profit-maximizing way. By raising prices in high demand lanes and lowering them in low volume lanes, imbalances can be reduced via pricing action. Gorman proposes a stochastic non-linear pricing optimization over a medium-term (e.g., quarterly) horizon. The work shows BNSF railway experienced an improvement of balance and therefore a reduction in repositioning costs while increasing expected revenues.

Since the early 2000s, the intermodal marketing company or IMC, which acts as a retail arm and third-party transportation management coordinator, has been taking ownership of its own containers, and thus has started to think about container-capacity based revenue management. Adelman (2007) evaluates strategic network pricing decisions in intermodal by modeling a dynamic fleet management problem on a closed logistics queueing network. Adelman's model leads to internal cost parameters similar to shadow prices based on network costs for improved dispatching decisions. The improved container allocation better balances supply and demand in the network. This work was not put to use in a practical setting.

In the short term horizon (e.g., 1 day to 1 week), container capacity in a geography in a geographic region is largely fixed because container repositioning takes considerable time. Gorman (2010) considers the decision facing Hub Group, an intermodal marketing company, when a shipper tenders an order. Given limited container capacity in the near term, Hub can accept the tendered order and its revenue, or reject the order in order to preserve the capability to accept a higher-revenue order that may be tendered subsequently. Further, the decision to accept an order should consider the anticipated future container supply and demand conditions at the destination of the shipment, which affects the future profitability of container capacity. This accept/reject decision does not allow pricing decisions, which are fixed in the short run, but manages container capacity in a way that maximizes expected revenue over the short run. Gorman suggests a simple probability-based heuristic based on expected revenues and the probability of running out of container capacity. Hub Group experienced both an increase in margin, a decrease in low-margin moves, and an increase in container velocity from container capacity-based load acceptance.

10.5 Future Directions and Opportunities for Revenue Management and Freight Rail

Recent research by Crevier et al. (2012) proposes joint capacity management and pricing decisions, attempting to bridge the gap between operations and pricing. This ambitious research expands the decisions beyond pricing given a capacity, and combines the two decisions. Working with Canadian National for practical input and data, they develop a largely theoretical approach that establishes both optimal pricing as train service provision, rail car handling, as well as capacities for handing railcars at classification yards. The ambitious project has not been implemented, but points in the direction of more integrated and holistic pricing and operations decisions.

References

Adelman D (2007) Dynamic bid prices in revenue management. Oper Res 55(4):647–661

American Association of Railroads (AAR) (2013) https://www.aar.org/keyissues/Documents/Background-Papers/A-Short-History-of-US-Freight.pdf

Armstrong A, Meissner J (2010) Railway revenue management: overview and models. Working Paper (available at http://www.meiss.com), Lancaster University Management School

Campbell K (1996) Booking and revenue management for rail intermodal services. Ph.D. thesis, University of Pennsylvania

Campbell K, Morlok EK (1994) Rail freight service flexibility and yield management. Proc Transp Res Forum 2:529–548

Clark C, Proulx B, Thoma P (2005) A survey of the freight transportation demand literature and a comparison of elasticity estimates. IWR Report 05-NETS-R-01. http://www.corpsnets.us/docs/other/05-nets-r-01.pdf

Crevier B, Cordeau JF, Savard G (2012) Integrated operations planning and revenue management for rail freight transportation. Transport Res Part B Meth 46(1):100–119

Gorman MF (2010) Hub group implements a suite of OR tools to improve operations. Interfaces 40(5):368–384

Gorman MF (2005) Estimation of an implied price elasticity of demand through current pricing practices. Appl Econ 37(9):1027–1035

Gorman MF (2002) Pricing and market mix optimization in freight transportation. J Trans Res Forum 56(1):135–148

Gorman MF (2001) Intermodal pricing model creates a network pricing perspective at BNSF. Interfaces 31(4):37–49

Kraft E (2002) Scheduling railway freight delivery appointments using a bid price approach. Transport Res Part A 36:145–165

Kraft ER, Srikar BN, Phillips RL (2000) Revenue management in railroad applications. Transport Q 54(1):157–176

Kraft ER (1998) A reservations-based railway network operations management system. Ph.D. thesis, University of Pennsylvania

Kwon OK (1994) Managing heterogeneous traffic on rail freight networks incorporating the logistics needs of market segments. Ph.D. thesis, Massachusetts Institute of Technology

Nozick LK (1992) A model of intermodal rail-truck service for operations management, investment planning, and costing. Ph.D. thesis, University of Pennsylvania

Sibdari S, Lin KY, Chellappan S (2008) Multiproduct revenue management: an empirical study of Auto train at Amtrak. J Rev Pric Manag 7(2):172–184

Small K, Winston C (1998) The demand for transportation: models and applications. Irvine Economics Paper, No.98-99-06

Strasser S (1996) The effect of yield management on railroads. Transport Q 50:831–844

Winston C (1983) The demand for freight transportation: models and applications. Transport Res Part A Gen 17(6):419–427

Chapter 11
Intermodal Rail

Bruce W. Patty

Intermodal shipments have become an increasingly important aspect of the North American railroad industry, both from a profitability perspective as well as one of growth and opportunity. In this chapter, we will present background information on this market since many readers who have worked in a traditional rail arena may not have experience or knowledge of the intermodal segment. We then provide examples of areas where OR models have been used to some degree of success. Following that, we go into three example problems in more detail. We conclude with a discussion of areas of opportunity for expansion of the use of Operations Research in the Intermodal Rail market.

11.1 Introduction and Background Information

11.1.1 Definition of Intermodal

BusinessDictionary.com defines Intermodal as: *Movement of containerized* (*unitized*) *cargo over air*, *land*, *or sea through the use of different transport modes* (*aircraft*, *truck*, *rail*, *boats*, *ships*, *barges*, etc.) *capable of handling containers*. For the purposes of this chapter we will focus on rail and the modes that are most commonly combined with rail, truck, and ship.

The term "international" when used in the Intermodal industry refers to shipments and equipment that move on ship, truck, or rail. These shipments often take many weeks to be delivered, and spend a high percentage of their time onboard a ship.

B.W. Patty (✉)
Veritec Solutions, San Rafael, CA, USA
e-mail: bpatty@veritecsolutions.com

© Springer Science+Business Media New York 2015

B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3_11

The term “domestic” when used in the Intermodal industry refers to shipments and equipment that stay within North America and move via rail and truck. Larger equipment sizes are used than in International for various reasons, including the wider roads in the US as well as not needing to be able to be easily loaded and transported on ships. In recent years, growth in domestic intermodal has exceeded that in the International market for several reasons including the conversion of freight from moving strictly over-the-road in trailers to using domestic containers. Much of this has been driven by increases in fuel prices, congestion on roads, and shortages of truck drivers.

11.1.2 Brief History of Intermodal

From Wikipedia—Origins of Intermodal Transportation

- Intermodal transportation goes back to the eighteenth century and predates the railways. Some of the earliest containers were those used for shipping coal on the Bridgewater Canal in England in the 1780s. Coal containers (called “loose boxes” or “tubs”) were soon deployed on the early canals and railways and were used for road/rail transfers (road at the time meaning horse drawn vehicles).
- Wooden coal containers used on railways go back to the 1830s on the Liverpool and Manchester Railway. In 1841 Isambard Kingdom Brunel introduced iron containers to move coal from the vale of Neath to Swansea Docks. By the outbreak of the First World War, the Great Eastern Railway was using wooden containers to trans-ship passenger luggage between trains and sailings via the port of Harwich.
- The early 1900s saw the first adoption of covered containers, primarily for the movement of furniture and intermodal freight between road and rail. A lack of standards limited the value of this service and this in turn drove standardization. In the U.S. such containers, known as “lift vans”, were in use from as early as 1911.
- In the United Kingdom containers were first standardized by the Railway Clearing House (RCH) in the 1920s, allowing both railway owned and privately owned vehicles to be carried on standard container flats. By modern standards these containers were small, being 1.5 or 3.0 m long (5 or 10 ft), normally wooden and with a curved roof and insufficient strength for stacking. From 1928 the London, Midland and Scottish Railway offered “door to door” intermodal road–rail services using these containers. This standard failed to become popular outside the United Kingdom.
- Pallets made their first major appearance during World War II, when the United States military assembled freight on pallets, allowing fast transfer between warehouses, trucks, trains, ships, and aircraft. Because no freight handling was required, fewer personnel were needed and loading times were decreased.
- Truck trailers were first carried by railway before World War II, an arrangement often called “piggyback,” by the small Class I railroad, the Chicago Great

Western in 1936. The Canadian Pacific Railway was a pioneer in piggyback transport, becoming the first major North American railway to introduce the service in 1952. In the United Kingdom, the big four railway companies offered services using standard RCH containers that could be craned on and off the back of trucks. Moving companies such as Pickfords offered private services in the same way.

- Highway semi-trailers in piggyback service at Albuquerque, New Mexico.
- In the 1950s, a new standardized steel Intermodal container based on specifications from the United States Department of Defense began to revolutionize freight transportation. The International Organization for Standardization (ISO) then issued standards based upon the U.S. Department of Defense standards between 1968 and 1970.
- The White Pass and Yukon Route railway acquired the world's first container ship, the *Clifford J. Rogers*, built in 1955, and introduced containers to its railway in 1956. In the United Kingdom, the modernization plan and in turn the Beeching Report strongly pushed containerization. The British Railways freightliner service was launched carrying 8-foot (2.4 m) high pre-ISO containers. The older wooden containers and the pre-ISO containers were rapidly replaced by 10-foot (3.0 m) and 20-foot (6.1 m) ISO standard containers, and later by 40-foot (12 m) containers and larger.
- In the U.S., starting in the 1960s, the use of containers increased steadily. Rail intermodal traffic tripled between 1980 and 2002, according to the Association of American Railroads (AAR), from 3.1 million trailers and containers to 9.3 million. Large investments were made in intermodal freight projects. An example was the USD $740,000,000 Port of Oakland intermodal rail facility begun in the late 1980s.
- Since 1984, a mechanism for intermodal shipping known as double-stack rail transport has become increasingly common. Rising to the rate of nearly 70 % of the United States' intermodal shipments, it transports more than one million containers per year. The double-stack rail cars design significantly reduces damage in transit and provides greater cargo security by cradling the lower containers so their doors cannot be opened. A succession of large, new, domestic container

sizes was introduced to increase shipping productivity. In Europe, the more restricted loading gauge has limited the adoption of double-stack cars. However, in 2007 the Betuweroute was completed, a railway from Rotterdam to the German industrial heartland, which may accommodate double-stacked containers in the future. Other countries, like New Zealand, have numerous low tunnels and bridges that limit expansion for economic reasons.

11.1.3 Equipment Variations

There are numerous specialty types of equipment used in the intermodal market, but over time, the types of equipment have evolved into four major sizes of containers with chassis designed to fit each. These are 20′ containers, 40′ containers, 48′ containers, and 53′ containers. The 20′ and 40′ containers are primarily used for International freight. Ships are configured to handle them both in their holds as well as on deck. 20′ containers are primarily used for high-density shipments that reach weight maximums before they reach cubic capacity limits. In industry parlance, they "weigh out before they cube out."

Before the development of the 53′ container in the late 1990s, 48′ containers were used for both international and domestic freight. However, since domestic trailers were typically 53′ in length, domestic shippers wanted access to longer containers that could be loaded like trailers before they would convert from using trailers to using containers. Initially, in order to provide a container that could withstand the rigors of rail and yet be light enough to carry significant weights while on roads, aluminum 53′ containers were used. These containers however were prone to leaks and had high maintenance costs. In the early 2000s, Pacer Stacktrain pioneered the use of 53′ steel containers manufactured in China using high-strength lightweight Swedish steel. These containers weighed only slightly more than the 53′ aluminum containers, yet had much lower maintenance and repair costs and were not as prone to leaks. While some container providers were slow to transition from aluminum to steel, steel containers currently dominate the marketplace. Containers can last up to 12–14 years depending upon the number of loads they move and the forces they encounter, as well as the level of maintenance they receive. For steel containers, the floors are often the first element to fail or require significant investment in maintenance.

For each of the container sizes, chassis have been designed. Chassis provide the wheels and support structure to allow a container to be taken off of a ship or a train, and then be pulled by a truck to its final destination. While some chassis can be adjusted via a sliding trombone structure to support multiple sizes, the majority of chassis are fixed-frame, that is, designed to fit just one size of container. While containers are designed to be transported around the world in the international market and around North America in the domestic market, chassis are designed to move on roads within about a 500-mile radius of where the container is mounted. (Note: In the domestic intermodal marketplace, containers are designed to move around

North America on trains, then be mounted on chassis at rail terminals in order to be transported from the rail terminal to the destination by trucks.) The largest expenses for maintaining chassis are related to the tires and brakes.

11.1.4 Role of Railroads and IMCs

Railroads play a key role in intermodal. The low-cost economics of moving heavy shipments across thousands of miles has driven the conversion of over-the-road freight transport to rail. Plus, as railroads have dramatically improved their reliability over the past 15–20 years, customers who were concerned about having their shipments arrive in a timely manner have been much more comfortable with using intermodal transport. For railroads, the increasing volume of intermodal freight has helped them to remain profitable while much of their traditional freight that was related to manufacturing in North America has been lost as factories have moved to other lower cost locations like China. For example, there used to be high volumes of iron ore and steel that were transported via rail from mines to steel plants and from steel plants to automobile factories. With the closure of many US steel plants, these freight volumes have greatly diminished.

Many railroads also play another role in the intermodal marketplace. Not only do they transport the containers, they also provide the equipment themselves. Union Pacific, Norfolk Southern, and CSX Transportation all maintain fleets of domestic containers which they make available to shippers who are using their railroads for all or part of the move from origin to destination. Container programs like EMP, CSXU and UMAX are owned and operated by one or more railroads. BNSF used to also provide a fleet of containers for use by its customers but made the decision in the mid-2000s to focus on transporting containers owned by others.

Intermodal Marketing Companies, or IMCs, serve as an intermediary between the railroads and the beneficial cargo owner (BCO) or shipper. While large BCOs can negotiate rates directly with the railroads, smaller companies do not have the knowhow or the leverage to get competitive rates. The IMCs can negotiate rates directly with the railroads, then pass those rates onto their customers. Some of the IMCs, like JB Hunt, the Hub Group, Schneider and Pacer, also have their own fleets of containers that they can make available to their customers. They also can use the railroads' container programs when appropriate.

11.1.5 Chassis Pools, Both Domestic and International

Historically, each container owner also had a fleet of chassis. These chassis would be marked with their logo, they would be owned or leased by the container owner, and the container owner would maintain the chassis. The container owners would also be responsible for repositioning chassis from one location to another in situations where container flows shifted. Some container owners viewed being in control of their chassis fleet as a competitive advantage since they could control and in many cases eliminate chassis shortages at loading points like steamship terminals or rail terminals. Others viewed it as a "necessary evil" that was not part of their corporate culture.

Having each container owner have its own fleet of chassis created additional challenges for the railroads. They would be responsible for mounting a container from Company A on a chassis from Company A, even if those chassis were not conveniently located to where they were needed. So, their truck driver might have to haul a 40′ Hapag Lloyd chassis away from trackside to the storage location and pickup a 40′ APL chassis and bring it back to trackside. This caused both inefficiencies for the railroads as well as the need for more chassis to be located at rail facilities than would be needed if the same 40′ chassis could be used for both moves. In the mid-2000s, chassis pools under the auspices of OCEMA, an Ocean Carrier group, were started at most major rail locations in the US. Chassis owners contributed their chassis to these pools and were charged by a pool manager based on a combination of factors including chassis usage and maintenance and repair costs.

On the domestic side, chassis pools were also started in the mid-2000s. BNSF signed an exclusive agreement with TRAC/Interpool, a leasing company, to provide domestic chassis at all of their rail terminals. All customers, except for JB Hunt who had a separate agreement with BNSF, were required to use TRAC pool chassis for their shipments. Later, CSX Intermodal also entered into a similar agreement with TRAC/Interpool. Some customers, like Pacer who had agreements that allowed them to keep chassis on terminals, have continued to use their own chassis, but that is unlikely to continue past their existing agreements.

11.2 Examples of Decisions to Be Made Where OR Models Can Be Used

11.2.1 Pricing

Intermodal companies must frequently make decisions regarding what rates to charge their customers. For the railroads to determine their rates to charge IMCs, they must take into account such factors as:

- Underlying costs to transport the container(s) from origin A to destination B.
- Competitive factors.
- Network factors.

While the underlying costs may seem to be easy to calculate, one always has to determine whether an additional container being moved on a train should only be allocated the incremental cost of that additional move or should bear some share of the overall costs of operating the train that would have been incurred even if that additional container had not been added to the train. Competitive factors include not only the rates being offered by other railroads, but the rates being offered by trucking companies. Network factors include the impact of having one more container being moved from A to B among all of the other concurrent movements, whether or not containers were available at location A to be used for this move, and whether or not location B needs to have more containers arrive in order to make that location more in balance. For example, if a railroad has 20 surplus containers in Jacksonville and a shortage of containers in Chicago, then a move from Jacksonville to Chicago would have beneficial network impacts while a move from Chicago to Jacksonville would have a detrimental network impact. Given the complexity of making these decisions, especially for rates that could be in effect for many months into the future, there is an opportunity for OR models to play a valuable role.

11.2.2 Container Fleet Sizing

Most containers are manufactured in China and then transported to the US if they are to be used in the domestic intermodal marketplace. Production lead times for containers are about 90 days assuming that raw materials like steel and flooring are readily available. Intermodal companies normally experience their peak periods in late fall and early winter when stores are stocking up for the upcoming holiday sales season. So, in order to ensure that containers are in place for this critical season, negotiations need to be completed in early spring so that raw materials can be ordered, transportation from China to the US can be reserved, and equipment can be manufactured in time. However, if a company sizes its fleet to meet all of the peak season demand, then they will have surplus equipment during the remainder of the year, if nothing else is changed. In order to properly size these fleets, models that take into account revenue, profits, cost of shortages, storage costs for surplus equipment, lease expiration dates, rates for new equipment leases, rates for old equipment leases, and forecasted demand need to be used in order to make the right fleet sizing decisions. Evaluating the combinations of all these factors requires complex models and provides an opportunity for the use of Operations Research approaches.

11.2.3 Demand Forecasting

Demand forecasts are required in order to make decisions that will both *impact* the future as well as those that *are dependent* on the future state. Most companies use information from either the same month, last year, or the most recent month to project what will happen in the future. This often ignores changes in economic factors, competitive factors, seasonal factors, etc. that should be taken into account. Many companies develop very complex operational models to route traffic, size fleets, and manage facilities yet spend relatively little effort on determining the demand forecast that will be fed into those models. For example, companies may develop an optimization model to determine their plan for repositioning containers around their network. Since repositioning moves may take up to 7–10 days to get the containers from origin to destination, the value or quality of the repositioning decision is dependent on understanding the demand levels 7–10 days from when the optimization model is run. Yet, these same companies use very simplistic approaches to develop their estimates of demand in the future. The inaccuracies of these forecasts can often negate the value of the optimization model's recommendations.

11.2.4 Assignment of Equipment to Customers

In times of equipment shortages, wholesale equipment (container) providers often need to choose which customer gets a container and which does not. There are several factors that should be taken into account when making this decision, including:

- Profitability of the movement for which the container would be used.
- Network impact of the movement for which the container would be used.
- Importance of keeping the customer happy (are they a large customer? Are there service commitments included in the contract? Have they been denied equipment recently?)
- Risk of loss of the customer or movement (Will their freight be lost by them finding another equipment provider or will it be retained if a container is provided on the next day? Do other equipment providers have available equipment?)

Generally, when a shipper or IMC requests a container from the wholesale equipment provider, the IMC does not specify where the container will be shipped. So, to accurately assess the above factors, the wholesale equipment provider must either guess the most likely use of the container, or develop a set of likely uses with a probability estimate for each, and then compute weighted estimate amounts for the above factors. Given all of this complexity, one can see where OR models could easily add value. More information regarding how one IMC, Hub Group, has approached this problem is provided later on in this chapter.

11.2.5 Chassis Fleet Sizing and Positioning

While much of the problem of Chassis Fleet Sizing is analogous to that of Container Fleet Sizing, the chassis problem also has the added complexity of needing to determine how to distribute or position the chassis around the intermodal network to support the container fleet. Containers are designed to flow around the network so even though new containers may all arrive from the manufacturer at one location, say Los Angeles, they can be added to the network at that point and then move around the network as they are used for outbound shipments. This approach does not work for chassis. If the container fleet is growing, then additional chassis will be needed. Typically, companies must forecast chassis demand for each metropolitan area for the future container fleet, then compare that forecast to current inventories. This determines where additional chassis may be needed. If the chassis network has not been “balanced” recently, there may be some locations that already have surpluses of chassis. When negotiating the acquisition of new chassis from leasing companies, this shortage and surplus information is used to determine where to take delivery of the new chassis. Often, the leasing companies will factor a delivery cost into the lease cost and in other situations, the lessee may need to arrange for transportation from one or more centralized delivery locations. This is another area where OR models can add value to improving the decision making process. Some major chassis pool operators are experimenting with the use of a network optimization approach to help make decisions regarding chassis repositioning at minimal cost.

11.3 Detailed Examples of Actual Model Implementations

11.3.1 Empty Container Repositioning

11.3.1.1 Background on Problem

For both domestic and international intermodal container movements, flows across the country are not balanced. That is, the flow of loaded containers into a location is rarely the same as the flow of containers out of the location. There are several reasons for this including the fact that the US imports more products than it exports, and that many factories use as input products that can be transported inside intermodal containers but output products that cannot be transported inside intermodal containers or vice versa. For example, the component assemblies used to make automobiles can often be transported inside an intermodal container, but the finished automobile cannot. Because of these imbalances, surpluses of empty containers can build up at various locations, known as “surplus” locations, if not repositioned to locations where there are more loads originating than terminating, known as “deficit” locations.

The decision to reposition these containers from surplus locations to deficit locations is an ongoing one. If the number of surplus containers is allowed to get too large, then the presence of these containers can congest container yards as well as constrict the availability of chassis to support inbound loaded containers. While some locations have available acreage where surplus empty containers can be stacked and stored, this acreage is limited and there are lift charges to be paid to third-party contractors.

11.3.1.2 Typical Decision Making Approach

While the number of containers that move into and out of a given location can vary significantly from day to day, most locations can be categorized relatively consistently as either balanced, surplus, or deficit. Because of this, most companies that have a fleet of containers to manage develop guidelines or rules of thumb that provide direction as to when and where to move the surplus containers. They may have a guideline that says that when the number of surplus containers exceeds a given pre-defined number at a given location, then move them to a pre-specified deficit location. For example, at a location like Kearny NJ, Pacer Stacktrain might move all the surplus containers to Chicago once the number of empty containers there got above 50.

The decision making approach can differ from one company to the next depending upon the cost to them for repositioning empties. Union Pacific's approach to repositioning containers from Chicago to Los Angeles could be much different than Pacer Stacktrain's approach since UP would not have to pay a railroad to move the containers. Also, UP would have much better visibility into the availability of flatcar platforms and can schedule their empty container repositioning to take advantage of capacity that would go unused otherwise when empty flatcars needed to be repositioned. An IMC that manages a fleet of containers would not have the ability to take advantage of that kind of information.

To determine these guidelines, container fleet operators take into account the locations that are typically surplus and the locations that are typically deficit, as well as the availability of rail service and the transportation cost between these locations. Using this information, they come up with guidelines that drive empty containers from surplus locations to the nearest, or lowest cost, deficit location. For example, empty containers in Jacksonville may always be sent to Atlanta, empty containers in Phoenix may always be sent to Los Angeles, and empty containers in Laredo may always be sent to Dallas. Because of the time that it can take to move containers from one location to another, container fleet operators are basing these decisions on an expectation of the future demand for these containers at the deficit locations.

11.3.1.3 Optimization Approach

Pacer Stacktrain decided a few years ago to see if Operations Research optimization models could be applied to this repositioning problem in order to reduce costs and improve decision making. They engaged with Innovative Scheduling, now Optym, a consulting firm based in Gainesville, Florida that had considerable experience with these kinds of problems, especially as related to freight railroads, to develop a proof-of-concept model. People from Pacer's Logistics and Equipment team conducted several meetings with Innovative Scheduling to describe the problem and the decision making context. The approach described below is a high-level description of the result of those efforts.

The group decided to develop a model that could be used in two different planning scenarios. In both scenarios, the demand for empty containers is driven by a set of loaded container movements. That is, the model would solve for the movement of empty containers that would position them to be available at the right time and place to support a projection of loaded movements. In Scenario 1, the movement of loaded containers would be the same as the actual loaded movements for some historical time period, normally a month. However, rather than force the use of the actual empty container movements that took place historically, the optimization model would solve for these movements. The empty container movements that the model solved for and thus recommended would be compared to those that actually took place to see what improvements were possible and what could be learned. In Scenario 2, additional potential loaded movements would be added to the demand projection. The model would try to select the optimal mix of loaded and empty movements that maximized net revenue, or minimized net costs, with the goal being to see if the model could be more efficient about managing the fleet and come up with a more effective strategy that allowed either more volume to be carried or to select a more profitable mix of loaded movements to be carried.

11.3.1.4 Network Construction

The problem described earlier was modeled using a multi-commodity time–space network structure. The commodities represent the sizes of containers in the fleet. At the time the model was being developed, Pacer had two sizes of containers, 48′ and 53′. Because it can take many days to move containers from one location to another, the problem must be modeled on a time–space network, where each node actually has a geographical component (Dallas, Atlanta, Chicago, etc.) and a time component (Day 1, Day 2, …, Day n). An arc $A(l)$ is constructed between two nodes $N(i,j)$ and $N(k,j+t)$ if there is rail service between the two cities I and k and it takes t days to move between them.

Inventory control arcs are also constructed between node $N(i,j)$ and $N(i, j+1)$ to be used by containers that will remain at node i from day j to day $j+1$. These arcs have upper bounds on them to enforce capacity limitations at each location and are primarily used to address the empty container storage or parking limitations at a location. If containers are allowed to stay at a location and not be moved, then more containers can accumulate than can actually be stored there.

When loaded containers arrive at their destination, they are delivered to the customer who then unload them. These containers then become available to be used as empties some number of days later. This process was handled in the model by creating an arc that connected the node on the arrival date to the same node on the empty availability date.

Node balance constraints are created to ensure that the flows into a node on a given day are equal to the flows out of the node.

For each potential container movement, either empty or loaded, paths are generated from origin to destination using the arcs of the network. Circuity logic is put into place to prevent the generation of paths that would either take too long, expressed as an allowable percentage increase compared to the most direct routing, or contain cycles where the container moved through the same location more than once during its movement. Each path is represented by a binary variable, whose value was 0 if the path was not used and 1 if it was.

11.3.1.5 Objective Function Components

Since some decision variables represent paths through the arcs of the network, their objective function component was comprised of the costs involved with traversing that path. For example, if a path involved the movement of an empty container on two trains that interchanged with a time gap at the interchange location, the cost could represent the rail costs of the two train moves plus a handling and storage cost at the interchange location. Other objective function components were related to the use of the storage arcs described earlier. For paths that represented the movement of loads, the revenue from the loaded movement was also included in the objective function component for those paths. The model was constructed as a minimization problem, so the revenue was actually subtracted from the costs of a path so that profitable routes actually had a net negative objective function component. Paths that represented the movement of empties had no such revenue offset.

Note: this treatment of revenue was especially important when using the model for Scenario 2. In this Scenario, it was important to realize that an optimal solution might actually have higher total empty repositioning costs that were more than offset by higher revenue amounts created by either selecting different loads for movement or by creating a solution that allowed for more loaded movements, or a combination of the two.

11.3.1.6 Constraints

There were several types of constraints involved in this model. As mentioned earlier, some of these constraints were node balance constraints that were created to ensure that containers did not stay at a given location, at least not without creating a storage cost component.

Since multiple paths, or variables, could be created for each loaded or empty movement, constraints were generated to ensure that no more than one of these variables could be non-zero in the solution. That is, only one path could be selected for each movement. For loaded movements that had to be moved, for example in Scenario 1, the constraints were constructed so that exactly one of the paths for each loaded movement was selected. For empty movements or loaded movements that were optional, as in Scenario 2, these constraints were constructed so that no more than one of the paths could be selected.

Other constraints were constructed to ensure that capacity limitations were enforced for storing containers, either loaded or empty. While it was initially thought that this could be handled by upper bounds on the variables related to storage from 1 day to the next, many situations were found where no feasible solutions could be found using the information provided by Pacer Operations personnel. Basically, the limits that they provided on the number of containers that could be stored had actually been exceeded historically. In order to allow the model to actually solve the problem and for this issue to be identified, modelers decided to set this up as a constraint with an additional variable with a high objective function component that, in a sense, represented "buying" additional capacity. If a solution had this additional variable set to a non-zero amount, further investigation was performed to determine if capacity was actually higher than had been estimated.

11.3.1.7 Solution Approach

The resulting problem formulation can be categorized as a set partitioning or set covering problem with additional variables and additional constraints. The set partitioning or set covering aspect of the model addresses the need to select no more than one path from a set of paths for each container movement. To solve this large, integer programming problem with hundreds of thousands of binary variables, CPLEX's MIP solver was used with good results.

Since Pacer did not have the ability to provide a "snapshot" of data that would provide information regarding the location of all the containers at any given time in the past, a process was needed to be run to "populate" the network. To do this, a 3-month set of demands was provided to the model. The model was run for all 3 months, but only the solutions for the middle of those months were compared to historical results. The first month of demand was used to "load up" or "warm up" the system, and the results related to the last month were not of value since there was no reason to reposition containers near the end of the month since there was no loaded demand in the future that needed empties to be positioned at the right locations.

Also, since the snapshot information on locations was not available, a "supersource" node was added to the network. The fleet of containers was made available at that node at time 0, with arcs connecting that node to all other origin nodes. Arcs with zero costs and no capacity limitation were used to connect the "supersource" so that containers could be absorbed into the network.

11.3.1.8 Results

The model results were validated by comparing the solutions for empty repositioning recommended by the model to those that actually took place, and investigating the differences. Much of the time spent in this process was involved in reviewing the storage capacities that were input into the model since often the model needed to increase the initial capacity values in order to get feasible results. Once this review was completed, the Pacer logistics team found that the model was identifying some opportunities that were worthy of inclusion into their normal repositioning plans. However, to move forward and use the model on a more regular basis, the need for a detailed demand forecast became clear. Work was initiated to develop such a forecast when the recession of 2008 started. The recession generated dramatic reductions in loaded container demand and dramatic surpluses of containers network-wide. Since the model added value primarily in conditions where the container fleet was stressed, i.e., there were insufficient containers in the fleet to carry all the demand, work on the model was put to the side.

11.3.2 Chassis Pool Sizing

11.3.2.1 Overview

One of the key responsibilities of the Equipment Planning group at Pacer Stacktrain was to determine how many chassis of each size (20′, 40′, 48′, and 53′) needed to be positioned at each of the 50–60 locations across North America where Pacer containers would arrive on trains. At the time, Pacer had the largest domestic container fleet in North America with over 27,000 containers and also had contracts with its rail partners which allowed Pacer to provide its own chassis at rail terminals across North America. Initially, Pacer developed a spreadsheet-based model to estimate the number of chassis of each size that would be needed at each equipment supply point (EQSP). This analytic model used traditional inventory planning inputs like turn-times (estimated number of days that an arriving container would use a chassis), forecasted number of containers of each size arriving on a train each day, and the number of days each week that trains arrived or departed. Using this approach, the model determined the average number of chassis needed to support the inbound volume of containers. This model did a good job at estimating the number of chassis that would be needed in "steady state" conditions. However, more often than was desirable, the number of chassis actually needed far exceeded the projection. The Equipment group needed to identify what was causing the model to be so far off.

11.3.2.2 Approach

Since the model was developing accurate projections at about 90 % of the EQSPs, it was believed that the fundamentals of the model must be working properly. Given that, an initial guess was that one or more of the inputs to the model was off. The most likely possibilities were that inbound freight had surged, turn-times had significantly increased, or the number of trains operated each week had dramatically dropped. However, when updated measurements for these values were analyzed, it was found that actual numbers were quite close to those used in the model. With that first hypothesis proven wrong, other possibilities needed to be considered.

The group decided to step back from the problem and see if they could identify any business conditions that consistently were present at EQSPs where the actual number of chassis needed exceeded the projections. They set up conference calls with both the Equipment team and the Operations team to discuss what was happening at the terminals that were "in trouble." After several calls it became evident that they needed to conduct some historic analyses PRIOR to the calls, or they would get bogged down with anecdotal discussions about what happened on one particular day when some unusual situation took place; this made it virtually impossible to move the discussion to the underlying fundamentals. After using these analyses to discredit some theories that were driven by these one-time occurrences, they realized that EQSPs where they were running short of chassis tended to be locations where empty containers would build up until they were repositioned out on trains. That is, inbound loaded container volume exceeded outbound loads, and empties were building up at the terminal.

They then went back and looked at the model to see how it handled this situation, and found out that turn-times were being measured from when the container and chassis left the terminal after arriving on an inbound train to when the container and chassis "ingated" the terminal after being released by the customer. The time between when the container ingated the terminal and when the container was taken off the chassis and placed on the outbound train was not included in this measurement, often because those events were not transmitted to Pacer by the rail carrier. However, this time was not included for both loaded and empty containers. Why was its omission only causing problems at terminals where empties accumulated?

To answer this question, they arranged another round of conference calls with the Operations team, and found out that a key difference in the way that loaded containers and empty containers were handled by the railroads was that, if there was limited space on the trains, the loaded containers would get priority. So, empty containers would be left behind. While this worked fine in terms of meeting delivery promises for the loaded containers, it caused situations where empty containers would stay mounted on chassis for days. And since these days were not being captured in the measurement of turn-times, the model was not accounting for this in the chassis projection. In short, they discovered that under certain and occasional conditions, the modeling assumptions did not reflect operational practice.

The Equipment group ended up modifying the model that estimated chassis requirements by using historic chassis usage trends that did include chassis on terminal, and then looking at averages, maximums and variances from the norm to develop demand projections. With this change, they were able to dramatically improve the accuracy of the model. The change in the modeling approach was one of the key reasons that Pacer was able to meet chassis needs with an industry low chassis-to-container ratio of 85 %.

11.3.3 Container Selection Process

Note: this section is based on an article by Michael Gorman published in Interfaces in 2010 (Gorman, M.F., "Hub Group Implements a Suite of OR Tools to Improve Its Operations", Interfaces, 40 (5), 2010).

11.3.3.1 Background

As of 2010, Hub Group was the largest IMC in North America with annual revenue approaching $2 billion. Historically, Hub had used containers and trailers that were provided by US railroads to move their customer's shipments. One of the features of this approach was that they were allowed to evaluate each shipment individually, based on the profitability of that shipment alone and not in the context of impact on their network. One of the shortcomings of this approach, however, is that when peak season arrived, Hub was not guaranteed access to enough capacity to make sure that they could handle demand since they were competing for these rail-owned assets with other IMCs. This led Hub to the decision to acquire its own fleet of containers; a fleet which had grown to about 16,000 containers by 2010.

Since Hub now had its own fleet of containers, they were responsible for managing their own network flows in order to keep utilization high and avoid expensive repositioning moves. At the same time, Hub still used rail-owned assets, so they needed to match the right assets to the right customer move. That is, they had to determine when to use a Hub-owned asset and when to use a rail-owned asset to best serve their customers at lowest cost. If they used a rail-owned asset for a customer's move, then afterward they could either reuse that container for another move or just return it to the railroad that supplied the container. However, if they used a Hub-owned asset, they needed to continue to manage the use of that container at the destination location of that given move.

In order to most effectively manage this process, Hub turned to the use of Operations Research models.

11.3.3.2 Approach

As each order for a container was handled, a decision needed to be made as to whether or not the order should be accepted, and if so, whether a Hub-owned container should be dispatched or if a rail-owned container should be used. In 2007, Hub and a consulting team led by Mike Gorman of the University of Dayton implemented a suite of five models to support these real-time dispatching decisions that needed to be made.

- Supply and demand forecasting.
- Capacity valuation.
- Fleet inventory targeting.
- Load accept optimization.
- Load routing optimization.

In this section, we will briefly discuss the roles and relationships of each of these models as well as the solution approach used within each model.

11.3.3.3 Supply and Demand Forecasting

The first step in the process is developing a supply forecast and a demand forecast. This provides a context within which to make decisions regarding the impact to network conditions of the dispatch decisions. The supply forecast is made up of two components: controlled supply and street supply. Controlled supply represents the containers that are under Hub's control. These are either containers currently in use by Hub's customers, Hub-owned containers currently empty and available at Hub's container yards, or Hub-owned containers being repositioned. Historical unloading time distributions are used to predict when containers currently in use will become available. Street supply represents the rail-owned containers that will be available for use on a specific date at a specific ramp location. This is forecast based on Hub's experience with rail providers and current conditions. The combination of street supply and controlled supply provides a daily supply forecast by location.

Demand forecasts are generated for about 150 major origin–destination pairs in the Hub network. These forecasts are developed using statistical approaches driven by customer ordering behavior. Traditional time-series methodologies were used. The long range demand forecast was also used to inform the long range supply forecast since loaded movements provide empty supply at the destination.

11.3.3.4 Capacity Valuation

The role of the capacity valuation (CV) model is to estimate the marginal profit potential of having an additional container at a location on a given date. For a given origin–destination pair, the capacity valuation at the origin is the opportunity cost of accepting a load and at a destination, the value is the potential profit that having an additional container available at destination would generate. To calculate the CV, Gorman created a two-step heuristic methodology that involved the calculation of the profitability of loads originating from a location and the probability that an additional container would actually be used. The complexity of this task was increased by the fact that the profitability not only varies by lane but that the profitability can vary by customer within a lane due to different contractual arrangements, i.e., rates, that Hub may have with different customers. The details of the approach used can be found in Gorman's paper published in Interfaces.

11.3.3.5 Fleet Inventory Targeting

The goal of this module is to address the question: "What inventory of Hub containers is desirable at each location to allow a profit-maximizing mix of Hub and rail container fleet assignments?" The idea here is that having too little Hub fleet inventory might result in missed shipments or lower profit margins due to having to use more expensive rail-owned containers. Having too many Hub fleet containers can result in low utilization or increased repositioning costs. One factor that makes this question difficult to solve is that the answer can vary from day to day since demand can vary from day to day. Gorman developed a nonlinear optimization model to estimate the optimal Hub container inventory levels by day by location. To solve this, the model optimizes the assignment of containers to orders subject to various operational constraints including that no more containers can be assigned than are available and that the next day's inventory is linked to the solution for the current day. These target inventory levels could be provided to dispatchers who can use them to inform their dispatching decisions. However, it would appear that the fluctuations from day to day may have been too volatile, so heuristics were developed that address this situation.

11.3.3.6 Load Accept Optimization (LAO)

Historically, order requests were handled on a first come—first served basis. If an order had a positive one-way profitability, it was accepted. This approach is valid in a situation with unlimited capacity and no need to be concerned about what happens to the container at destination, which was the case when Hub was only using rail-owned containers and had unlimited access to that fleet. When they transitioned to both having their own container fleet and competing with other users for a constrained rail-owned fleet, this approach was no longer valid. They now needed a process to determine whether an order request was more valuable than a potential future load that might not be able to be fulfilled if there was no capacity.

To address this, a heuristic was developed using the supply forecast, the demand forecast, and the capacity valuation. A "load profitability threshold" was developed for each location and date using the CV of that location as an origin and the weighted average of the CV's for the destinations from that location. If the tendered load value meets or exceeds this threshold, it is accepted. If not, it is rejected. Dispatchers have access to this recommendation and can override it based on their experience, the specifics of the customer, and other knowledge they have that is extraneous to the model.

11.3.3.7 Load Routing Optimization

While Load Acceptance focuses on whether or not to accept an order, Load Routing Optimization (LRO) focuses on whether a Hub-owned container should be dispatched or a rail-owned container. In general, the customer is indifferent to the assignment, but there can be significant operational and profitability impact to Hub. Historically, dispatchers assigned the container with the lowest costs on a first come—first served basis. Once these were depleted, the higher cost containers were used. However, this did not take into account such factors as the impact at destination of the decision.

To solve this, Gorman formulated a transportation problem that "combines forecasts, capacity valuations, and inventory targets to balance near-term costs with future potential profit." Such factors as the one-way costs of the load that vary based on the type of container assigned, the target inventories by container type, and the expected capacity valuations were incorporated in the objective function. Constraints enforced that each order that had been accepted in LAO was to be assigned a container and that equipment capacity was not exceeded. The results of this model can be provided to dispatchers to provide guidance in the form of a list of recommended assignments. Dual values from the transportation problem also helped provide insights regarding the cost of not following the recommendation.

11.3.3.8 Results

As was the case with the equipment repositioning model that was prototyped for Pacer Stacktrain, this work for Hub was developed during a timeframe when capacity was tight and it was important to determine how best to use the limited resources. When the recession hit the US in 2008, intermodal volumes dropped significantly and equipment became plentiful, in fact, too plentiful. Companies who owned fleets resorted to storing a significant amount of equipment, selling older containers, or returning those to leasing companies whose leases had expired. So, the focus on this effort was lost and many of the potential benefits are yet to be realized.

11.4 Opportunities

There are several other areas of decision making in the Intermodal market place where Operations Research models can be of significant value. In this section, we describe a few of these and suggest how they could be helpful.

11.4.1 Forecasting

As in most industries, the ability to make insightful decisions regarding actions to take in the future is dependent on the ability to predict conditions in the future. In the intermodal arena, a key element that must be predicted is origin–destination volume on a day-by-day basis. While decision-makers recognize the need for this information, very little effort has been put into applying Operations Research tools to generate these forecasts, at least based on the presentations made at INFORMS conferences by the Operations Research groups of the various railroads. Most companies either assume that tomorrow will be like yesterday, or that next Thursday will be like last Thursday, or some combination of the above. There are some attempts to use information about what is already moving in the network to provide projections of supply. That is, some companies are using information about loaded movements and when they are expected to arrive at their destination, then incorporating historic information about how long a customer normally takes to unload the container and make it available. However, there is little science being applied to the problem of predicting how much demand there will be originating at a given location, and even less effort being applied to predicting the origin–destination volumes.

11.4.2 Tactical Equipment Matching

Those intermodal companies who provide containers in a wholesale marketplace can dramatically improve equipment utilization by "matching" available empties to demand points without requiring the container to be returned empty to their container yard, known as a "CY." If the company can arrange for an empty container to be moved directly from one customer (customer A) who has recently used the container and where it is now sitting empty to another customer (customer B) who has the need for an empty container, significant economies can be achieved. Among the benefits are:

1. Reduced dray costs—normally an empty is drayed from customer A back to the CY. Then, customer B sends their trucker into the CY to pull a container and deliver it to their facility. If the "matching" can take place, then the trucker for customer B can go directly to customer A's facility and pull the container.
2. Reduced equipment costs—by reducing the total number of days that containers sit idle in CY's, the total number of containers that are needed to satisfy the same demand can be reduced. In addition, the total number of chassis that are needed can be reduced.
3. Reduced number of trucks and drivers that are needed—reducing the number of truck moves required will allow there to be fewer trucks and drivers needed to support the same volume. This does not imply that drivers would be furloughed or laid off, but rather that, as intermodal volume increases, fewer additional trucks and drivers would be needed.

To perform this equipment matching most effectively, information on the status of each container is needed so that dispatchers, or dispatching systems, know when containers are empty and available to be retrieved. Many intermodal companies, especially those that have moved into intermodal from the trucking industry (Swift, Schneider, JB Hunt) as opposed to moving into intermodal from the rail industry (Pacer, Union Pacific, CSX), have installed devices on the containers that use GPS technology along with cellular data networks that report the location of the container and whether or not it is empty. Those companies who have moved into intermodal from trucking have been accustomed to having these data available to them and have incorporated them into their dispatching approaches. Those companies who have moved into intermodal from the rail industry have not been accustomed to using this information and would have to revise their approaches and software in order to take advantage. For those companies who have incorporated this information, it has dramatically improved their capability to perform tactical equipment matching. In addition, traditional OR models, like network optimization algorithms can be used to determine which available, or projected available, empty container should be assigned to which demand point for an empty.

Index

© Springer Science+Business Media New York 2015
B.W. Patty (ed.), *Handbook of Operations Research Applications at Railroads*,
International Series in Operations Research & Management Science 222,
DOI 10.1007/978-1-4899-7571-3