# 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 |

| | | | | | | | Activity Flags | | | | Blocks Carried | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Location | Arrival | Depart | Day offset | Max Cars | Max Length | Max 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 $\pm 6$ hours due to expeditious handlings, and multiple departures per day for each block, compared to times of $\pm 24$ 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.

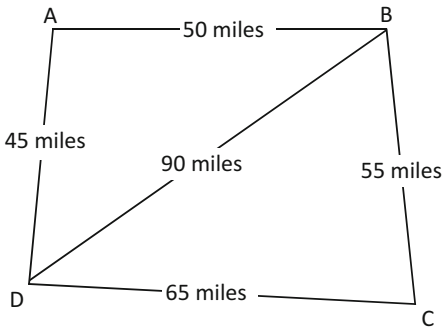



**Fig. 1.1** Railroad network

**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

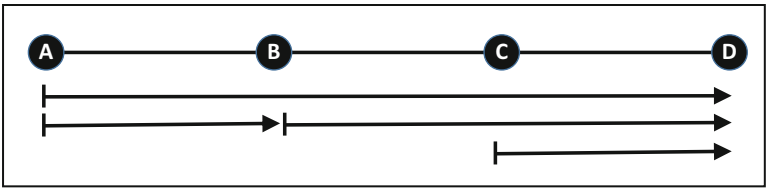
for the necessary empty repositioning movements. As a consequence, the scheduling strategies discussed in this section generally would not apply to most unit train operations.

- *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 traf-

fic 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