Chapter 13
Freight Railroad Service Network Design

Mervat Chouman and Teodor Gabriel Crainic

1 Introduction

Rail transportation supports our social and economic life, providing economically-priced, environmentally-friendly, and timely transportation services for people and freight at the urban, regional, national, and international levels. A freight train, cargo train, or goods train is a group of freight cars (US) or goods wagons (International Union of Railways) hauled by one or more locomotives on a railway infrastructure network, transporting cargo all or some of the way between the shipper and the consignee. Railroads move large quantities of products, bulk materials (e.g., grains. minerals, petroleum and chemical products), intermodal containers and trailers loaded on flat cars, general freight, or specialized freight (e.g., automobiles and heavy machinery) in purpose-designed cars. Railroads are particularly efficient for long-haul movements in terms of per ton-km monetary, energy-consumption, and pollutant-emission costs. They are faster and more direct than ocean freight, which lead to setting up transcontinental land bridges, e.g., the North-American Landbridge linking the West and East coasts, and the Eurasian Landbridge between China and Western Europe. We focus on freight rail transportation in this chapter, while passenger rail transportation is discussed in Chap. 17.

Freight rail transport makes up an essential link in intermodal transportation and supply chains, supporting national and international trade. The efficiency of railroads, in terms of cost and reliable on-time delivery, thus directly impacts the

M. Chouman
Effat University, Jeddah, Saudi Arabia
e-mail: mchuman@effatuniversity.edu.sa

T. G. Crainic (✉)
CIRRELT and AOTI, Université du Québec à Montréal, Montréal, QC, Canada
e-mail: TeodorGabriel.Crainic@cirrelt.net

T. G. Crainic et al. (eds.), *Network Design with Applications to Transportation and Logistics*, https://doi.org/10.1007/978-3-030-64018-7_13

availability and cost of goods for the final customer, be it a private citizen, an institution, or a company. To achieve this efficiency, railroads operate mostly as *consolidation*-based carriers, similarly to less-than-truckload trucking (Chap. 14), liner shipping (Chap. 15), and city logistics (Chap. 16), for example.

The fundamental idea of consolidation is to take advantage of economies of scale and reduced handling in terminals, by grouping loads from different shippers, with possibly different origins and destinations, and loading them into the same vehicles for efficient long-haul transportation. Railroads generally implement a more complex double consolidation policy, however, as cars are grouped into blocks, which are then grouped into trains. Thus, loaded and empty cars, with different origins and destinations, being present simultaneously in the same terminal, are sorted and grouped into a block, which is then moved as *a single unit* by a series of trains until its destination, where it is broken down, the cars being either delivered to their final consignees or sorted for inclusion into new blocks. The performance and profitability of such a system depend on an offer of services meeting the cost and quality criteria of its potential customers, but also, for a large part, on efficient and coordinated terminal and long-haul transport operations.

Tactical, medium-term, planning for freight rail carriers aims to address this challenge at the network and system-wide level, through a transportation plan specifying the train services to operate over the contemplated schedule length (e.g., the week), together with their frequencies or schedules (timetables), the blocks that will make up each train, the blocks to be built in each terminal, and the routing of the cars, empty and loaded with the customers' freight, using these services, blocks, and terminal operations. As detailed in Sect. 2, tactical planning makes up a very complex problem, with many facets and decisions linked in a web of economic, resource utilization, and time-performance objectives, limitations, and trade-offs. Operations Research provides the *Service Network Design* (*SND*) methodology to build the railroad tactical plan making the most efficient use of the railroad's resources to achieve its economic and customer-service performance objectives. The chapter reflects this important relation between railroad planning and network design. It focuses on SND models for railroad tactical planning, both for particular activities, e.g., car blocking and train makeup, and for integrated planning processes.

The chapter is organized as follows. Section 2 briefly describes the rail transportation system, the associated tactical planning issues and the utilization of tactical-planning SND models, and concluding with the general notation used in the chapter. Section 3 is dedicated to SND formulations, which do not integrate the time dimension explicitly, for three problem settings: service selection and train makeup (Sect. 3.1), car classification and blocking (Sect. 3.2), and integrated planning (Sect. 3.3). Section 4 focuses on the case where the time characteristics of the problem components and decisions are explicitly addressed, and introduces the *Scheduled Service Network Design* (*SSND*) problem and model for the integrated planning of freight railroads. The SSND modeling framework is extended in Sect. 5 to account for existing schedules, the container-to-car loading rules of intermodal traffic, and resource management. Bibliographical notes are presented in Sect. 6 and we conclude with a number of research directions in Sect. 7.

2 Rail Transportation System and Planning

We initiate the section with a brief description of freight rail transportation, with its main objectives, system components, operations, and decision and planning challenges. Tactical planning issues and their complex interactions are discussed next, introducing the *Service Network Design* (*SND*) methodology generally proposed to address them and which is the object of this chapter. We conclude with a discussion on the various utilization modes of SND models, and the general notation used throughout the chapter.

2.1 Rail Transportation System

Railroads are complex transportation systems where several major components interact and compete for resources. The infrastructure of the system is made up of a large number of terminals and rail tracks linking them. Most of these terminals are stations where demand originates and terminates. A much smaller number are denoted *yards* and are specially equipped to handle large quantities of cars, sorting and grouping them for long-haul transportation, as well as to make up and disassemble trains. The term *classification* (*marshaling* is also found) yard is used to emphasize the major car-handling role of these facilities. Terminals are linked by a physical network of tracks. The backbone component of this network is made up of main lines connecting the yards of the system. The network is completed by a large number of secondary, branch lines connecting most stations to the backbone network. Even when stations are located on a main line, the movements of loaded and empty cars between stations and their respective designated yards are generally performed by local, so-called feeder trains.

Customer *demand* takes the form of a number of cars (the special case of intermodal transportation is discussed later in this section), of a type appropriate to the commodity that needs to be moved, to be shipped from an origin station to a destination one. The appropriate number of empty cars is delivered for loading by the railroad to the customer site, assuming it is connected to the rail network, or to a designated station, otherwise. The empty cars are generally delivered from a designated yard by a feeder train. Once loaded, the cars are moved back to the same yard or to a different one as appropriate for the long-haul movement on the main-line network toward the destination. Since the scheduling of feeder trains is usually not within the scope of the network-wide tactical planning process designing the long-haul service network, we assume in this chapter that demands are defined among origin and destination yards. Each demand is also characterized by a *volume* in terms of number of loaded cars of given physical and operational attributes, as well as by an *availability time* (and date) at the origin yard and a *due time* at the destination yard.

Trade is unbalanced among countries and regions and, consequently, so is the demand for particular car types in the case of railroads. Moving empty cars is costly and railroads aim to minimize such balancing flows. Yet, they cannot be entirely avoided and, thus, empty cars are often part of train composition. These movements must be accounted for when planning services and resources, to avoid underestimating traffic, resource utilization, and costs. Origin-destination "empty-car" volumes are thus often part of the demand definition.

Movements of freight on the rail network are performed by *train services*. A train is composed of one or more locomotives providing power and a series of cars (which may be loaded or empty; sometimes, locomotives are repositioned in the network and are part of a train without providing power). Each train has a particular origin yard where it is made up and a destination yard where it completes its journey, delivers all the cars currently hauled, and liberates the locomotives. The route may encompass a number of intermediary stops where the train delivers or picks up cars, eventually grouped into blocks as described below (locomotives and crews may also be changed, added and dropped off, at intermediary yards). Other than its route, the train service or, simply, the *service*, is also characterized by time-related information. In its simplest version, this information takes the form of a *frequency of service*, i.e., the number of times the "same" train is run during the length of time the railroad uses to define its recurring operations (e.g., 1 week), also called *schedule length*. A more precise definition is given by a service schedule indicating the departure time from the origin yard, arrival and departure times at each intermediary yard, and the arrival time at destination. This information may be strict, as for most European, Canadian, and a few U.S. Class 1 railroads, or relative (e.g., most U.S. Class 1 railroads), indicating time intervals for their departures which may be modified to account for particular events, e.g., the need to pass a direct train for an important customer. (Note that, there are still railroads around the world with schedules of an "indicative" nature, the train leaving when full and ready.) The railroad may operate a single type of service, e.g., dedicated intermodal shuttle trains between main yards. Alternatively, services of different types, e.g., general cargo, bulk, intermodal, may be defined and operated, often on the same infrastructure. Priority with respect to the other service types (often linked to the speed and capacity allowed on each section of track) is often used to define the service type.

Railroads aim to maximize revenue, which often translates into achieving the best balance between the operational cost of operating resources and services, on the one hand, and the quality of the service according to the customer expectations in terms of tariffs, speed, flexibility, and reliability, on the other hand. Dedicated and direct non-stop services from origins to destinations (so-called "unit" trains) would achieve high customer satisfaction, reducing delivery time and the risk of delay (providing train congestion on rail tracks is avoided), and eliminating the risk of damage related to car handling at intermediate yards. This would also, however, imply high operational costs, particularly for the very large numbers of origin-destination demands with low and medium numbers of car to move. Railroads therefore operate direct trains only for particularly important customers or when

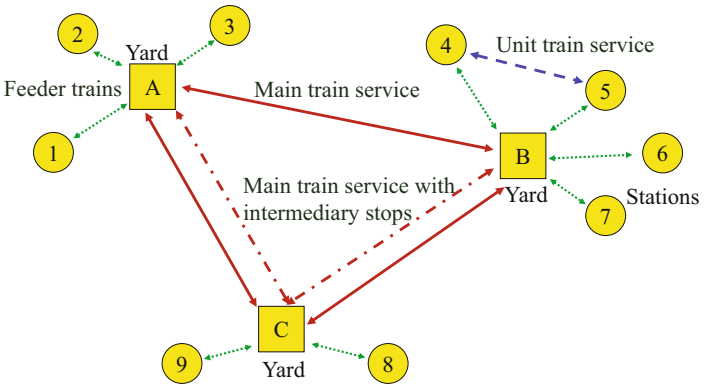

Fig. 13.1 Hub-and-spoke rail service network

the volume of demand between two stations is significant (i.e., the equivalent of at least a full train) and regular. Railroads rather aim for economies of scale for most of their operations through consolidation of freight from different demands, that is, cars with different origins and destinations, into blocks and blocks into trains.

Schematically, cars at their origin yard are sorted, *classified* is the term generally used, to be then grouped with other cars, with potentially different origins and destinations, into particular blocks. The *block* is then handled as a single unit from its origin yard, where it is formed, to its destination yard, where it is taken apart, its cars at their final destination being delivered to the respective consignees, the others being reclassified and blocked with other cars present at the yard for the next part of their trip. This classification and blocking operation contributes significantly to the economy-of-scale provided by rail transportation. Trains are thus made up of blocks and, when appropriate, it is blocks that are picked up and delivered at intermediate stops. Blocks may thus be *transferred* (*switched*) from one train to another.

Figure 13.1 illustrates this hub-and-spoke service organization for a network with three yards and nine stations. Dotted lines indicate feeder services moving cars and, eventually, blocks, between stations and main yards. The dash line illustrates a direct-train service between two stations, while the solid and dash-dotted lines represent the non-stop and one-intermediary-stop, respectively, long-haul train services moving on the main line network between yards.

Demand is moved along itineraries. Each *itinerary* for a particular demand specifies the sequence of blocks and trains, and thus the sequence of classification and transfer activities, between its origin and destination yards. The volume of freight of certain demands must be moved together, while for others, it can be split among several itineraries, as agreed between the railroad and the customer. From the railroad perspective, the possibility to split demand flows allows to better fill up blocks and trains, increasing the economies-of-scale, but requires additional care in monitoring the flows and making sure everything arrives in time to the final yard, for on-time delivery of the complete shipment. From an optimization perspective,

the model must include integer-valued flow variables when demand cannot be split, which increases the algorithmic challenge. To simplify the presentation, we assume in this chapter that flows may be split for all demands.

Operations are constrained by the physical characteristics of the infrastructure and the operational policies of the railroad and, thus, "capacity" is a multi-facet concept in rail transport. Consider, for example that, the car classification capacity of a yard, for a given time period, may be defined in terms of the maximum numbers of cars that may be handled, blocks that may be built (number and length of tracks on which the blocks are composed), trains that may be made up or serviced, and so on. Similarly, the capacity of the rail tracks limits operations with respect to the number of trains that may operate "simultaneously" on a given track segment (meeting or overtaking), as well as to the total weight, length or both a train may haul on the track. The length and weight of trains are thus limited and translate into lower and upper limits on the length and weight of blocks. Representing the operational characteristics and limits at a level appropriate for the network-wide nature of system and planning is one of the challenges of developing Operations Research-based methods for railroad freight transportation.

We conclude this section with a short discussion of an important and growing component of rail transportation, namely, intermodal traffic and operations. In its general sense, intermodality means that different transport modes are combined to seamlessly move containerized freight from a point of origin to a point of destination. A sequence involving a truck or rail (or a sequence of both) movement to a port, ocean navigation to another port, and a truck or rail sequence to destination is typical of intermodal transport and makes up the backbone of international trade. Rail plays a major role in this context as illustrated by the European Commission policy on intermodality, the new rail services being set up between China and Europe, and the intermodal-rail divisions of North American railroads linking the continental ports to the industrial and heavily populated regions of the continent.

Intermodal traffic is often handled separately from the general one, being moved on dedicated intermodal trains (attaching intermodal traffic to regular main-line trains may be viewed as a recourse operation to mitigate variations in forecast demand). Moreover, even when intermodal and regular cars and trains are handled in the same yards, the classification of intermodal cars is performed separately.

The most important difference, however, concerns the loading and unloading operations of intermodal traffic, which is actually taking place in particular zones of the railroad's yards. There is a large variety of container types, e.g., 20-, 40- and 53-feet long, and railroads use fleets of cars of various types, each with one or several platforms and slots on the platforms. Single- and double-stack platforms have one and two slots, respectively. The containers are delivered at yards (or maritime port facilities) and the railroad must determine the matching/loading of containers to available cars and types. This is an important but complex issue since not all combinations are legal or suitable, a very large number of loading alternatives exist, and decisions taken at any given yard impact the availability of cars at later periods at the yard and the other yards, as well as the performance of the railroad operations.

The double consolidation organization of freight railroads provides the sought-after economies of scale in operation costs and resource utilization, reduces car handling activities at yards, and fosters a timely service for markets (origin-destination pairs of cities or regions) with low traffic volumes. It also implies more complex operations in terminals, with potentially higher possibilities for delays and incidents. Which translate into more complex decision-making problem settings. The complexity is even larger for intermodal transportation which implies a third consolidation operation, of containers on multi-platform cars.

Network design-based models and methods are proposed to address these challenges and support decision making at various levels of planning. We now briefly recall these planning issues and the links to network design.

2.2 Tactical Planning and Network Design

The planning activities undertaken by railroads may be broadly classified into three levels, similarly to most other consolidation-based transportation systems. Strategic planning involves long-term decisions on system design, operation strategies, and acquisition of major resources (e.g., buy or rent locomotives or cars and enhance track or yard capabilities). Tactical planning is dedicated to building an efficient service and resource-utilization network and schedule. Short term planning, monitoring, and adjustment of operations make up the so-called operational planning (e.g., running the trains, crew and locomotive scheduling, repositioning crews, locomotives and cars for the next operations, and maintenance of infrastructure and rolling stock). We focus on tactical planning in this chapter, as it involves arguably the strongest connection to network design. We discuss at the end of this section the utilization of the related network design methodology in varied contexts, including the other levels of planning. Section 6 points to general references addressing railroad challenging planning activities and problems at the three levels.

Tactical planning is performed over a medium-term planning horizon, e.g., 6 months, called *season* in the following. Planning generally takes place some time before the beginning of the season. It aims to select and schedule services, together with the demand itineraries used to move the freight from origins to destinations using the resulting service network. Determining strategies for managing important resources supporting the selected services, as well as activity profiles for terminals, in terms of car, block, and train-handling policy for example, is also increasingly part of tactical planning. The goal is to satisfy the forecast regular demand in the most efficient way possible with respect to costs (profits) and resource-utilization, while satisfying the service-quality levels set by the carrier to answer customer requirements. Notice that, even though some part of demand, e.g., long-term contracts with customers, may be known at planning time, most is forecast using history, customer-relation representatives knowledge, and customer input, among other data sources. The service network and plan is determined for a rather short schedule length and it is repeatedly applied over the season.

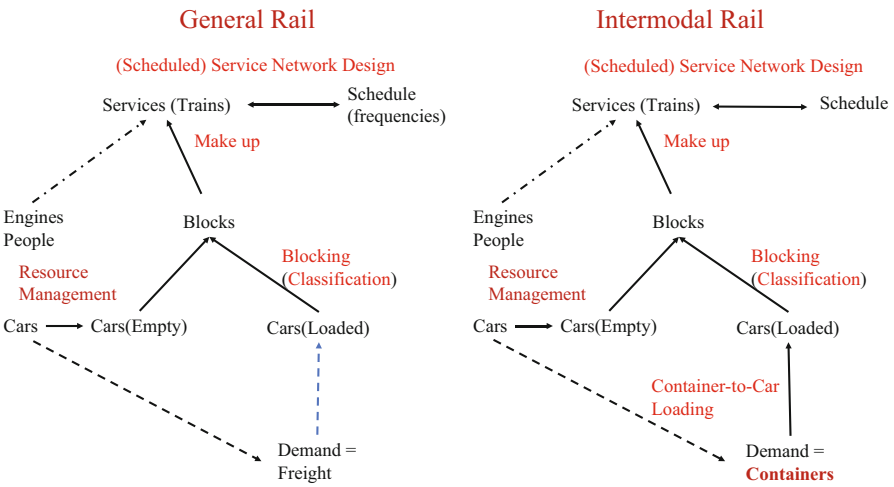

Fig. 13.2 Main activities and tactical planning decisions

A number of main decisions/issues make up the tactical planning process and are addressed through SND models and methods. These problems and their relations are briefly defined in this section and schematically illustrated in Fig. 13.2 for the general and the intermodal cases.

Service selection is concerned with choosing among a set of possible services the ones to operate the next season to service demand efficiently, profitably, and on time. The set of possibilities could represent a complete yard-to-yard network with intermediate stops for all service types, or the last-season network enriched with additional potential services to address changes in demand profile and railroad policies. The resulting service network specifies the movements through space and time of trains and cars, demand itineraries corresponding to paths in this network. The problem is defined as *static* when one assumes that neither demand nor the other problem characteristics vary during the schedule length considered. The time dimension of the service network is then implicitly considered through the definition of services and the inter-service operations at terminals, and one generally addresses the issue of *service frequency* assuming a more or less uniform distribution of departures over the schedule length. *Time-dependent* problem settings and formulations address the cases when the moments demands become available and are due at destinations are explicitly considered, which implies an explicit representation of demand and activities in time. Time-dependent formulations thus usually target the planning of *schedules* to support decisions related to *when* services and freight (demand itineraries) arrive at and leave from yards.

The *blocking* and *classification* problem addresses the issue of how cars are grouped in yards yielding the blocks to be moved by trains. It encompasses several strongly interrelated decisions: (1) select the blocks to build at each yard; (2) specify for each block its origin and destination terminals, the path through the

infrastructure network, the sequence of intermediate yards on the path where it will be transferred from one train to another (when relevant), the sequence of train services performing the transportation; and (3) define how cars, empty and loaded, are classified and assigned to blocks at their origin and at the intermediary (when brought in by blocks being dismantled at destination) yards on their journeys. The decisions concerning the empty cars are generally linked to car-fleet management concerns about providing the appropriate cars to yards given the particular associated demand. Loaded-car assignment to blocks, on the other hand, is related to the freight-routing objective of delivering shipments efficiently and on time.

The problem includes an additional dimension for intermodal services, namely, the assignment and consolidation of containers of various types and dimensions to multi-platform cars of different types and dimensions. The first challenge is how to reflect the differentiation of the many types of containers, cars, and loading rules, and how to represent the container-to-car assignment and loading in a way appropriate for the level of aggregation proper of tactical planning models and solution methods. Second, the containers-to-car consolidation adds a third combinatorial dimension and design decision to the blocking problem, yielding different SND formulations harder to address than for the regular-traffic case. This difference in illustrated in Fig. 13.2 by including the demand loading component in the intermodal-rail box on the right of the figure.

Train makeup yields the list of blocks, and cars of particular origin-destination demands, each train service hauls out of its origin yard, it drops and picks up at intermediary stops, and delivers at its destination yard.

Freight routing determines the *itinerary*, or itineraries when splitting of demand is allowed, used to transport the cars of each particular demand from its origin to is destination through the selected service network.

Resources, e.g., locomotives and cars, are required to operate services. *Resource management* addresses the issue of, on the one hand, assigning the appropriate resources to services to support the planned activities while, on the other hand, determining the general rules dictating the economically and operationally-efficient resource movements over the schedule length. Resource management is generally considered an operational-level managerial activity and its impact on tactical-level decisions was often limited to the somewhat simple case of accounting for the need to reposition empty cars for the next cycle of operations. The situation is evolving, however, and more comprehensive problem settings are detailed later in the chapter.

These problems may be, and have often been, addressed individually echoing a tactical-planning process decomposed into a series of sequential decisions. Increasingly, however, the strong interconnections among decisions, in particular in their cost and service quality consequences, lead to integrated approaches addressing several of the problems identified above jointly. Such approaches do not make problems easier, however, as planning must be performed network-wide aiming for the best trade off among the not necessarily convergent operational and economic characteristics of the individual problems and decisions. Thus, for example, one could increase the level of service by increasing the frequency of services, but this could result in higher levels of congestion in yards and on the tracks, resulting in increased delays and, thus, lower quality service (increased costs as well, of course).

Service network design (SND) models are generally proposed to address freight railroad planning problems. SND models take the form of *fixed cost, capacitated, multicommodity network design* formulations. Minimization of the total operating costs is the primary optimization criterion in most cases, reflecting the traditional objectives of railroads and expectations of customers to "get at destination fast but at the lowest possible cost". As customer expectations for high-quality service and environmental concerns rise, however, service performance measures are increasingly included in tactical planning and SND formulations. Service performance measures are generally modeled through delays incurred by freight and resources or the amplitude of violation of predefined performance targets (e.g., delivery within a given time length). Constraints may then be imposed on the values of the service-performance measures, or one may add them as costs and penalties to the objective function of the SND optimization formulation. The resulting generalized cost function then captures the trade offs between operating costs and service quality. The sections that follow present the main classes of railroad service network design models proposed, in increasing degree of problem and decision integration.

We conclude this general presentation with a short discussion on the utilization of SND methodology developed for tactical planning. To start, notice that, although network design models may be built to address strategic-planning issues, SND formulations may be used to evaluate the impact of strategic scenarios, relative, for example, to economic (e.g., fuel prices or changing production and consumption levels of certain goods) and regulatory (trade restrictions or speed and weight limits when carrying hazardous goods) variations on operations, resources, and system performance. One may thus use SND as a simulation tool in the context of cost-benefit analyzes, with appropriate approximation of railroad and demand characteristics. Clearly, generalized service network design models may be built to answer strategic-level decisions such as the number of each resource type to buy or rent, and the capital-intensive enhancement of infrastructure.

The service network design formulations may also be used to review, weekly for example, the tactical plan built for the season. One would then re-optimize and adjust the plan and operations to current conditions. What may be adjusted depends strongly on the railroad application context. Canceling or adding services on a short notice is not easily performed by railroads and SND models may assist in selecting the best alternative and determining the network-wide impacts. Updating the actual demands or resources, or both, assigned to blocks and trains is taking place quite often within railroad management and, again, SND models are appropriate. Obviously, the scope of the SND model has to be more focused when in plan-adjustment mode, parts of the system which should not be modified being fixed.

2.3 Notation

This section is dedicated to the definitions and general notation used throughout the chapter. It is summed up in Table 13.1 (together with notation proper to particular problem settings and defined in the next sections).

Let $\mathscr{G}^{\mathrm{PH}} = (\mathscr{N}^{\mathrm{PH}}, \mathscr{A}^{\mathrm{PH}})$ represent the physical network on which the railroad operates, where $\mathscr{N}^{\mathrm{PH}}$ stands for the set of terminals (yards and, possibly, main stations), connected by the physical track arcs of set $\mathscr{A}^{\mathrm{PH}} = \{(\eta_i, \eta_j), \eta_i, \eta_j \in \mathscr{N}^{\mathrm{PH}}\}$.

Yards $\eta \in \mathscr{N}^{\mathrm{PH}}$ are characterized by several capacity measures, defined for a given time period (which can be the schedule length or shorter for multi-period, time-dependent formulations), namely, the *classification capacity* u^{C}_{η}, for the number of cars that can be sorted and assigned to blocks, the *blocking capacity* u^{B}_{η}, for the number of blocks which can be built, the *block-transferring capacity* u^{T}_{η}, for the number of block which may be transferred from one service to another, and u^{M}_{η}, for the number of trains which can be made up at the yard during the period.

Train services run on this network to answer demand. Following the general Service Network Design (Chap. 12) and rail-planning literature, SND models select these services out of a set of *potential services* Σ, given an estimated *regular demand* for transportation represented by set $\mathscr{K}$ of origin-destination (OD) commodities, each commodity $k \in \mathscr{K}$ standing for the request to move a quantity d^k of freight from its origin terminal $O(k)$ to its destination terminal $D(k)$.

Each service $\sigma \in \Sigma$ is characterized by a path in the physical network between its origin and destination yards, $O(\sigma)$ and $D(\sigma)$, respectively. *Single-leg* services operate non-stop between their respective origins and destinations, while *multi-leg* services stop at one or several yards on their routes to drop and pick up blocks and cars. Let $\mathscr{N}^{\mathrm{PH}}(\sigma) = \{O(\sigma) = \eta_0, \eta_1, \eta_2, .., \eta_{n-1}, D(\sigma) = \eta_{n(\sigma)}\}$ be the sequence of yards visited by service $\sigma \in \Sigma$, and $\mathscr{L}^{\mathrm{PH}}(\sigma) = \{l_i(\sigma) = (\eta_{i-1}, \eta_i) \mid i = 1, \ldots, n(\sigma)\}$ be the sequence of service legs of the service, with $n(\sigma) = 1$ for single-leg services. Let $\mathscr{L}^{\mathrm{PH}} = \bigcup_{\sigma \in \Sigma} \mathscr{L}^{\mathrm{PH}}(\sigma)$. Several "cost" and "capacity" measures may be associated to services depending on the particular problem addressed. In almost all cases, however, one finds the fixed cost to select the service, f_σ, the leg unit transportation costs $c^k_{l_i(\sigma)}$, and the leg-service capacity $u_{l_i(\sigma)}$.

Each of the SND tactical planning models described in the following sections is defined on a network $\mathscr{G} = (\mathscr{N}, \mathscr{A})$ built out of the physical network and the set of potential services, enriched to address problem-specific characteristics, the modeling of time, in particular. Static SND problem settings, discussed in Sect. 3, do not include an explicit representation of time and, thus, $\mathscr{G}$ has $\mathscr{N} = \mathscr{N}^{\mathrm{PH}}$ and $\mathscr{A} = \mathscr{L}^{\mathrm{PH}}$ in those cases.

Scheduled service network design, *SSND*, targets time-dependent problem settings, where time and service schedules are explicitly considered. SSND formulations are built on *time-space* networks $\mathscr{G} = (\mathscr{N}, \mathscr{A})$, using a discrete or continuous representation of the schedule length $\mathbf{T}$. Let t stand for a time instant within the schedule length, i.e., $0 \leq t \leq \mathbf{T}$. A discrete representation of the schedule length is

Table 13.1 SND and SSND main notation

$\mathscr{G}^{\mathrm{PH}} = (\mathscr{N}^{\mathrm{PH}}, \mathscr{A}^{\mathrm{PH}})$	Physical network
$\mathscr{N}^{\mathrm{PH}} = \{\eta\}$	Set of terminals (yards and main stations)
$\mathscr{A}^{\mathrm{PH}} = \{(\eta_i, \eta_j)\}$	Set of rail-track arcs
$\mathscr{G} = (\mathscr{N}, \mathscr{A})$	Potential (service) network for the SND formulations
$\mathscr{N}, \mathscr{A}$	Set of terminals and arcs in $\mathscr{G}$
$\mathscr{H} = \{(\eta_t, \eta_{t+1})\}$	Set of holding arcs
u_a	Capacity of arc $a \in \mathscr{A}$
$\Sigma = \{\sigma\}$	Set of potential services
$O(\sigma), D(\sigma)$	Origin and destination terminals of service $\sigma \in \Sigma$
$\mathscr{L}^{\mathrm{PH}}(\sigma) = \{l_i(\sigma)\}$	Set of legs of service $\sigma \in \Sigma$
f_σ	Fixed selection cost for service $\sigma \in \Sigma$
$c^k_{l_i(\sigma)}$	Unit transportation cost for commodity $k \in \mathscr{K}$ on leg $l_i(\sigma) \in \mathscr{L}^{\mathrm{PH}}(\sigma)$ of service $\sigma \in \Sigma$
$u_{l_i(\sigma)}$	Capacity of service $\sigma \in \Sigma$ on its leg $l_i(\sigma) \in \mathscr{L}^{\mathrm{PH}}(\sigma)$
$o(l_i(\sigma))$	Scheduled departure time of service $\sigma \in \Sigma$ from the origin of its leg $l_i(\sigma) \in \mathscr{L}^{\mathrm{PH}}(\sigma)$
$d(l_i(\sigma))$	Scheduled arrival time of service $\sigma \in \Sigma$ at the destination of its leg $l_i(\sigma) \in \mathscr{L}^{\mathrm{PH}}(\sigma)$
$\mathscr{K} = \{k\}$	Set of origin-destination demands
d^k	Quantity of demand $k \in \mathscr{K}$
$O(k)$	Origin terminal of demand $k \in \mathscr{K}$
$D(k)$	Destination terminal of demand $k \in \mathscr{K}$
$o(k)$	Availability time of demand $k \in \mathscr{K}$ at its origin terminal
$d(k)$	Due date of demand $k \in \mathscr{K}$ at its destination terminal
$\mathscr{P}^k$	Set of itineraries in the service network for demand $k \in \mathscr{K}$
c^k	Unit service-quality cost, per unit of time, for demand $k \in \mathscr{K}$
u^{C}_η	Classification capacity, in number of cars, at yard $\eta \in \mathscr{N}$
u^{B}_η	Blocking capacity, in number of blocks, at yard $\eta \in \mathscr{N}$
u^{T}_η	Block-transfer capacity, in number of blocks, at yard $\eta \in \mathscr{N}$
u^{M}_η	Service make-up capacity, in number of services starting at yard $\eta \in \mathscr{N}$
τ^{C}_η	Expected delay to transfer a car at yard $\eta \in \mathscr{N}$
c^{T}_η	Unit car-transfer cost at yard $\eta \in \mathscr{N}$
c^{C}_η	Unit car-classification cost at yard $\eta \in \mathscr{N}$
f_b	Fixed building, transferring, and dismantling cost of block $b \in \mathscr{B}$
c^k_b	Unit transport cost for commodity $k \in \mathscr{K}$ on block $b \in \mathscr{B}$
$\mathscr{B} = \{b\}$	Set of blocks
$O(b), D(b)$	Origin and destination yards of block $b \in \mathscr{B}$
$\mathscr{N}(b) \subseteq \mathscr{N}$	Sequence of yards making up the route of block $b \in \mathscr{B}$
$\mathscr{L}(b)$	Sequence of service legs making up the route of block $b \in \mathscr{B}$
u_b	Capacity of block $b \in \mathscr{B}$
$\mathscr{B}(l_i(\sigma))$	Set of blocks assigned to service leg $l_i(\sigma) \in \mathscr{L}^{\mathrm{PH}}(\sigma), \forall \sigma \in \Sigma$
$\Theta = \{\theta\}$	Set of resource cycles
u^{R}	Quantity of resources available in the railroad system
$\mathscr{L}(\theta)$	Set of service legs of the resource cycle $\theta \in \Theta$
f_θ	Fixed cost of selecting and operating resources on cycle $\theta \in \Theta$
$\mathbf{T}$; $\mathscr{T}$	Schedule length; Set of discrete time periods

obtained by defining a sequence of time instances $t = 0, \ldots, \mathbf{T}$, grouped in set $\mathscr{T}$. The time period t in this representation corresponds to the length of time between instances t and $t+1$, $t = 0, \ldots, \mathbf{T}-1$, grouped in set $\mathscr{T}$. $\mathscr{N} = \{\eta_t, \eta \in \mathscr{N}^{\mathrm{PH}}, t = 0, \ldots, \mathbf{T}-1\}$ includes copies of all the yards in the physical network at all the time periods defined. This definition may be specific to each yard. To simplify the presentation, however, and without loss of generality, we assume the same time definition for all yards in this chapter, and time periods of equal length.

Each SSND potential service $\sigma \in \Sigma$ is defined on the time-space network $\mathscr{G}$. It is characterized by a *schedule* indicating arrival and departure times at the yards where it originates, stops, and terminates. The sets of yards and service legs identifying the service then become $\mathscr{N}(\sigma)$ and $\mathscr{L}(\sigma)$ (with $\mathscr{L} = \bigcup_{\sigma \in \Sigma} \mathscr{L}(\sigma)$), respectively, with departure time from origin, $o(l_i(\sigma))$, and arrival time at destination, $d(l_i(\sigma))$, for each service leg $l_i(\sigma) \in \mathscr{L}(\sigma)$. Note that, the schedule may also be described in terms of arrival and departure times at the yards in $\mathscr{N}(\sigma)$. Similarly, each demand $k \in \mathscr{K}$ is characterized by an *availability time* $o(k)$ at origin $O(k)$ and a due date $d(k)$ at destination $D(k)$.

The network is completed by the set of arcs $\mathscr{A}$, which includes *moving* and *holding* arcs. The former correspond to the legs of the potential services $\mathscr{L}$, an arc being defined for each service leg, while the latter are arcs connecting two time-consecutive representations of each terminal in $\mathscr{N}$. The attributes of the moving arcs $a \in \mathscr{A}$, the capacity u_a and the unit commodity-transportation cost c_a^k, inherit the values of the corresponding service legs for both SND and SSND cases. Holding arcs $\mathscr{H} = \{(\eta_t, \eta_{t+1}), \eta_t, \eta_{t+1} \in \mathscr{N}\}$ represent the possibility for equipment and freight to wait at a terminal for a time period.

Several measures are used in the industry and the literature for the amplitude of demand and the capacity of the system components, e.g., number of cars, number of containers, tonnage, and length. Moreover, more than one may be used simultaneously to constrain decisions and operations. Thus, the characteristics of a track segment may limit both the total tonnage a train may haul on the track, its total length (and this, independently of the locomotive power assigned to the train). To simplify the presentation, but with no loss of generality, we use a single and same unit to measure demand and service capacity, the latter being specific for each of the legs of the service.

3 Static SND

The section is dedicated to static service network design formulations. The general hypothesis of this class of planning problems and SND formulations is that neither demand nor the other problem characteristics vary during the schedule length and, thus, they do not integrate the time dimension explicitly. Time may still be accounted, however, through the selection of services. Instead of a simple yes or no decision, the formulations may select a service and its *frequency* of operation over

the schedule length. In the literature, one generally assumes that frequencies are uniformly distributed over the schedule length. We follow this trend in this chapter.

Models addressing particular components of tactical planning are presented in the first two subsections. Models integrating several tactical decisions are discussed in the third. We complete this part with a general discussion on the issue of generating the sets of potential services and blocks.

3.1 Service Selection and Train Makeup

The problem of selecting services and determining the cars their will haul arises when there is no blocking performed at yards (e.g., in most European railroads), or blocking is performed once the main service network is decided (see, e.g., the intermodal case described in Sect. 5). Given the set of possible services, the problem aims to (1) select the services to run and their frequencies over the planning period, and (2) assign cars to trains and determine the associated freight routing to accommodate all demand at minimum cost.

As in all problem settings considered in this chapter, freight routing may involve single-train itineraries from origin to destination and itineraries with service-to-service transfers at intermediary yards. Car transfers require time and resources. They generate costs and may cause delays related to many factors, including but not limited to, the number of cars to be transferred, the number of trains involved in transfers, and the capacity of the yard. Such delays not only increase the costs of the system, but may also decrease the service quality to customers.

The *Service Selection and Train Make-up Network Design (SMND)* problem thus aims to address these issues and decisions by minimizing the total operating cost, including penalties $\Phi(\cdot, \cdot)$ representing the interplay among delays and service quality standards. The model is built on a static network with $\mathscr{N} = \mathscr{N}^{\text{PH}}$ representing the physical yards of the system, and $\mathscr{A} = \mathscr{L}^{\text{PH}} = \bigcup_{\sigma \in \Sigma} \mathscr{L}^{\text{PH}}(\sigma)$ standing for the legs of the potential service set. Time is implicitly considered through (1) the possibility to define services with different travel times between the same pairs of yards, (2) the frequencies of the selected services, and (3) the cost-penalty associated to the delays.

Let us define the decision variables

- $y_\sigma \in \mathbb{Z}_+$: Frequency of service $\sigma \in \Sigma$;
- $x_a^k \geq 0$: Flow of commodity $k \in \mathscr{K}$ traveling on arc $a \in \mathscr{A}$, with $x_{l_i(\sigma)}^k = x_a^k$, for $a = l_i(\sigma), l_i(\sigma) \in \mathscr{L}^{\text{PH}}(\sigma), \sigma \in \Sigma$;
- $z_\eta^k = 1$ if commodity $k \in \mathscr{K}$ is transferred at yard $\eta \in \mathscr{N}^{\text{PH}}(\sigma), \sigma \in \Sigma$, and 0, otherwise.

Let $\mathscr{A}_\eta^+ = \{a = (\eta, j) \in \mathscr{A}, j \in \mathscr{N}\}$ and $\mathscr{A}_\eta^- = \{a = (j, \eta) \in \mathscr{A}, j \in \mathscr{N}\}$ be the sets of outward and inward arcs (service legs) of node $\eta \in \mathscr{N}$. The SMND is then formulated as

$$\text{Minimize} \quad \sum_{\sigma\in\Sigma} f_\sigma y_\sigma + \sum_{k\in\mathscr{K}}\sum_{\sigma\in\Sigma}\sum_{l_i(\sigma)\in\mathscr{L}^{\text{PH}}(\sigma)} c^k_{l_i(\sigma)} x^k_{l_i(\sigma)} + \sum_{k\in\mathscr{K}}\sum_{\sigma\in\Sigma}\sum_{\eta\in\mathscr{N}^{\text{PH}}(\sigma)}\sum_{a\in\mathscr{A}^+_\eta} \Phi(x^k_a, z^k_\eta) \tag{13.1}$$

Subject to

$$\sum_{a\in\mathscr{A}^+_\eta} x^k_a - \sum_{a\in\mathscr{A}^-_\eta} x^k_a = \begin{cases} d^k, & \text{if } \eta = O(k), \\ -d^k, & \text{if } \eta = D(k), \\ 0, & \text{otherwise,} \end{cases} \quad \forall \eta \in \mathscr{N}, k \in \mathscr{K}, \tag{13.2}$$

$$\sum_{k\in\mathscr{K}} x^k_{l_i(\sigma)} \le u_{l_i(\sigma)} y_\sigma, \ \forall l_i(\sigma) \in \mathscr{L}^{\text{PH}}(\sigma), \sigma \in \Sigma, \tag{13.3}$$

$$x^k_{l_i(\sigma)} - x^k_{l_i(\sigma)n} \le d^k z^k_{\eta_i},$$

$$k \in \mathscr{K}, i = 1, \ldots, n(\sigma) - 1, \ l_i(\sigma) = (\eta_{i-1}, \eta_i) \in \mathscr{L}^{\text{PH}}(\sigma), \sigma \in \Sigma, \tag{13.4}$$

$$y_\sigma \in \mathbb{Z}_+, \ \forall \sigma \in \Sigma, \tag{13.5}$$

$$x^k_{l_i(\sigma)} = x^k_a \ge 0, \ \forall k \in \mathscr{K}, a \in \mathscr{A}, \tag{13.6}$$

$$z^k_\eta \in \{0, 1\}, \ \forall k \in \mathscr{K}, \eta \in \mathscr{N}^{\text{PH}}(\sigma), \sigma \in \Sigma. \tag{13.7}$$

Constraints (13.2) represent the usual flow conservation and demand satisfaction requirements. Linking constraints (13.3) ensure that the total load on any service leg cannot exceed the capacity of the service on that leg, provided the service is selected. Constraints (13.4) make sure that the transfer (and classification, eventually) costs are paid whenever such an operation is performed, by setting the transfer variable z^k_η to 1 whenever the flow of commodity k is transferred from service σ to a different at node η, except at the origin and destination of the demand.

The objective function (13.1) represents the total cost of the system computed as the sum of selecting, i.e., making up, operating, and dismantling, services at determined frequencies, and moving demand shipments on the selected services, plus a monetary evaluation of customer-service satisfaction. The later is captured through a penalty term $\Phi(x^k_a, z^k_\eta)$, which is application specific and may take various forms.

To illustrate, consider that transfers not only require time and resources, generating costs and delays, but also increase the possibility of missed connections and late arrival at destination of certain demand flows. Railroads thus aim to reduce the number of transfers and may also pay particular attention to commercially sensitive customers. Let c^{T}_η be the unit car-transfer cost at yard $\eta \in \mathscr{N}$, and τ^{C}_η the expected delay to transfer a car at the same yard. (τ^{C}_η may be defined to account for congestion in the yard and for the type of rail car or commodity involved, but, for simplicity of presentation, we use a linear term here.) Let also c^k be the service-quality cost per unit of time for demand $k \in \mathscr{K}$. We may then define for each arc (service leg)

$$\Phi(x_a^k, z_\eta^k) = \Phi(x_{l_i(\sigma)}^k, z_\eta^k) = (c_\eta^{\mathrm{T}} + c^k)\tau_\eta^{\mathrm{C}} x_{l_i(\sigma)}^k z_\eta^k, \tag{13.8}$$

which captures the yard and demand-specific costs of transferring cargo between services and potential loss of service quality. This modeling approach yields nonlinear objective functions, however, increasing the computational challenges.

3.2 Car Classification and Blocking

As described in Sect. 2, car classification and grouping into blocks is central to aiming for the goal of efficiency and revenue maximization railroads. Recall that, cars with possibly different origins and destinations are classified and grouped into blocks; blocks are moved by trains and are handled as single unit from their origins to their destinations, where they are broken up. On the other hand, cars at their origin yard may follow itineraries where they are classified and assigned to a block which brings them to their destination yard, from where they are to be delivered to their consignees. Alternative itineraries may also be used where the initial block brings cars to an intermediate yard, where they are reclassified into new blocks, and continue their trip towards the final destination or another intermediate yard and reclassification. More than one reclassification may make up the itinerary.

The objective is to minimize cost. Decisions are highly constrained by the operating policies of the railroad (e.g., what blocks may be put on particular services), as well as by the resource and physical limitations of the yards in terms of, e.g., yard type and layout, numbers and characteristics of the yard equipment such as yard locomotives, personnel, and number and length of the classification tracks to which sorted cars are directed and where blocks are built. This translates, for each yard $\eta \in \mathcal{N}$, into unit car classification cost, c_η^{C}, as well as limits on the total number of cars which may be classified and blocked during a certain period of time, u_η^{C}, the total number of blocks one may build, u_η^{B}, or transfer, u_η^{T}, during the same time, the number of trains one may make up, u_η^{M}, etc.

Two approaches have been proposed to address this challenging car classification and blocking problem: (1) Develop the block plan first, then devise the set of train services (and schedule, possibly) to accommodate the blocks; (2) Select first the set of services and, second, build the block plan on the resulting service network. Both problems are challenging and SND formulations have been proposed to address them. Notice that, although it is the former which is mainly found in the literature, there is no methodological difference between the two in a static setting, except for the network on which the SND model is built, physical or service, respectively. We present the blocking problem in the block-first context in this subsection, together with a general formulation. The second case, often encountered when intermodal services are planned, is further detailed in Sect. 5.

The problem setting considers the physical (first case above) or the designed service network (second case). It is defined on a network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, with $\mathcal{N} =$

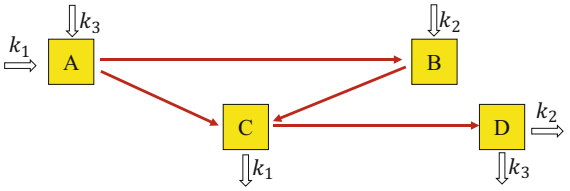

Fig. 13.3 Blocking SND network

$\mathscr{N}^{\text{PH}}$, $\mathscr{A} = \mathscr{B}$, the set of *potential blocks* linking the yards in $\mathscr{N}$, and the OD demand represented by set $\mathscr{K}$.

Similarly to the service definition, let $\mathscr{N}(b) \subseteq \mathscr{N}$ be the sequence of yards where the block $b \in \mathscr{B}$ is formed (origin $O(b)$), is dismantled (destination $D(b)$), and transferred from one service to a different one. Then, let $\mathscr{L}(b)$ be the sequence of service legs, $l_i(\sigma) \in \mathscr{L}^{\text{PH}}$, transporting the block from its origin to its destination, through the transfer yards, when relevant. Let $\delta_\sigma^b = 1$, if at least a service leg of service $\sigma \in \Sigma$ moves block $b \in \mathscr{B}$, and 0, otherwise, and let $\mathscr{B}(l_i(\sigma))$ be the set of blocks assigned to service leg $l_i(\sigma) \in \mathscr{L}^{\text{PH}}, \forall \sigma \in \Sigma$. Finally, let f_b be the block (fixed) building, transferring, and dismantling cost, c_b^k the unit cost of transporting commodity $k \in \mathscr{K}$ from the origin to the destination of the block, and u_b the block *capacity*, measured in the same units used for services.

To illustrate, consider the simple four-yard network displayed in Fig. 13.3, with four directed rail tracks, three OD commodities, and eight potential blocks $b_1 = (A, B), b_2 = (A, C), b_3 = (A, B, C), b_4 = (A, C, D), b_5 = (A, B, C, D), b_6 = (B, C), b_7 = (B, C, D), b_8 = (C, D)$, the intermediate yard labels identifying the block route not transfers. The possible commodity itineraries are: three for k_1: b_2, b_3, and (b_1, b_6) with reclassification at B; two for k_2: b_7, (b_6, b_8) with reclassification at C; and six for k_3: two direct, blocks b_4 and b_5, or with reclassification at yards B or C, or both, via the block paths (b_1, b_6, b_8), (b_1, b_7), (b_2, b_8), and (b_3, b_8), respectively.

The goal is to select the blocks to build from within $\mathscr{B}$, and to assign OD demand commodities to them, at minimum total cost, computed as the sum of the fixed cost of building, transferring, and dismantling the blocks, the cost of car classification, and the car transportation cost. Notice that, even though $\mathscr{A} = \mathscr{B}$, we write the formulation in terms of $\mathscr{B}$ to emphasize the classification and blocking scope of the model. Define the decision variables

- $y_b = 1$, if block $b \in \mathscr{B}$ is built, and 0, otherwise;
- x_b^k, continuous flow variable representing the volume of commodity $k \in \mathscr{K}$ assigned to block $b \in \mathscr{B}$.

The car classification and blocking service network design formulation takes then the following form:

$$\text{Minimize} \quad \sum_{b\in\mathscr{B}} f_b y_b + \sum_{k\in\mathscr{K}}\sum_{b\in\mathscr{B}} c_b^k x_b^k + \sum_{k\in\mathscr{K}}\sum_{\eta\in\mathscr{N}}\sum_{b\in\mathscr{B}_\eta^+} c_\eta^{\mathrm{C}} x_b^k \tag{13.9}$$

Subject to

$$\sum_{b\in\mathscr{B}_\eta^+} x_b^k - \sum_{b\in\mathscr{B}_\eta^-} x_b^k = \begin{cases} d^k, \text{ if } \eta = O(k), \\ -d^k, \text{ if } \eta = D(k), \\ 0, \text{ otherwise}, \end{cases} \quad \forall\, \eta \in \mathscr{N}, k \in \mathscr{K}, \tag{13.10}$$

$$\sum_{k\in\mathscr{K}} x_b^k \le u_b y_b, \ \forall\, b \in \mathscr{B}, \tag{13.11}$$

$$\sum_{k\in\mathscr{K}}\sum_{b\in\mathscr{B}_\eta^+} x_b^k \le u_\eta^{\mathrm{C}}, \ \forall\, \eta \in \mathscr{N}, \tag{13.12}$$

$$\sum_{b\in\mathscr{B}_\eta^+} y_b \le u_\eta^{\mathrm{B}}, \ \forall\, \eta \in \mathscr{N}, \tag{13.13}$$

$$y_b \in \{0,1\}, \ \forall\, b \in \mathscr{B}, \tag{13.14}$$

$$x_b^k \ge 0, \quad \forall\, b \in \mathscr{B}, k \in \mathscr{K}, \tag{13.15}$$

where $\mathscr{B}_\eta^+ = \{b = (\eta, j) \in \mathscr{B}, j \in \mathscr{N}\}$ and $\mathscr{B}_\eta^- = \{b = (j, \eta) \in \mathscr{B}, j \in \mathscr{N}\}$ are the sets of outward and inward arcs of node $\eta \in \mathscr{N}$.

The objective function (13.9) represents the total cost measured as the total fixed cost of building, transferring, and dismantling blocks, total cost of moving cars on blocks, and total car classification cost at yards. Constraints (13.10) and (13.11) are the classical flow conservation and block linking and capacity constraints, respectively. Constraints (13.12) and (13.13) enforce the yard capacity limits in terms of the number of cars that can be classified and the number of blocks that can be built during the planning period. Decision-variable ranges are defined by constraints (13.14) and (13.15).

3.3 Integrated Planning SND

Sections 3.1 and 3.2 addressed the issues of selecting and making-up trains, and classifying cars and building blocks separately. Yet, the solution to one problem is affecting the planning and solution of the other, no matter which problem is considered first. To emphasize the strong relations among these issues, consider, on the one hand, that the availability and frequency of a service determine the possibility of building and transporting blocks using that service while, on the other hand, the usefulness of a train service depends on the amount of traffic, in terms of blocks and cars, the train may service. Integrated-planning SND formulations address these issues simultaneously to select the service and the block networks and,

thus, define the classification strategy, as well as to determine the demand itineraries, establishing how freight is to be routed through the service and block network.

We start with the arc and path formulations of the SND model for this problem in Sects. 3.3.1 and 3.3.2, respectively. Section 3.3.3 presents a path formulation when one extends the SND model to account for more advanced features such as non-additive costs/tariffs and congestion phenomena. We conclude the section with a short discussion of service and block generation issues in Sect. 3.4.

3.3.1 Arc-Based Integrated SND

The integrated SND model is built on the network G of potential services. Then, $\mathscr{N} = \mathscr{N}^{\text{PH}}$ and $\mathscr{A} = \mathscr{L}^{\text{PH}}$. The components and notation of Sects. 3.1 and 3.2 apply directly to the integrated context, in particular the yard, service, and block definitions and characteristics. We recall the decision-variable definitions allowing the formulation to address simultaneously the selection of services with their frequencies, the selection of blocks to build at each yard, and the itineraries of demand determining the routing of the flows within the service and block networks and, thus, the classification strategy at each yard:

- $y_\sigma \in \mathbb{Z}_+$: Frequency of service $\sigma \in \Sigma$;
- $y_b = 1$, if block $b \in \mathscr{B}$ is built, and 0, otherwise;
- $x_b^k \geq 0$, continuous flow variable representing the volume of commodity $k \in \mathscr{K}$ assigned to block $b \in \mathscr{B}$; as the cars grouped within a block are the same over all the route of the block, that is, on all the service legs of the services carrying it, $x_b^k = x_{l_i(\sigma)}^k$, $l_i(\sigma) \in \mathscr{L}^{\text{PH}}(\sigma)$, $\sigma \in \Sigma$ (and equal to x_a^k as $a = l_i(\sigma)$).

The integrated service and block selection with classification model is formulated as mixed integer SND:

$$\text{Minimize} \quad \sum_{\sigma\in\Sigma} f_\sigma y_\sigma + \sum_{b\in\mathscr{B}} f_b y_b + \sum_{k\in\mathscr{K}} \sum_{b\in\mathscr{B}} c_b^k x_b^k + \sum_{k\in\mathscr{K}} \sum_{\eta\in\mathscr{N}} \sum_{b\in\mathscr{B}_\eta^+} c_\eta^{\text{C}} x_b^k \tag{13.16}$$

Subject to constraints (13.5), (13.10)–(13.15), and

$$y_b \leq y_\sigma, \quad \forall b \in \mathscr{B}(l_i(\sigma)), l_i(\sigma) \in \mathscr{L}^{\text{PH}}(\sigma), \sigma \in \Sigma, \tag{13.17}$$

$$\sum_{k\in\mathscr{K}} \sum_{b\in\mathscr{B}(l_i(\sigma))} x_b^k \leq u_{l_i(\sigma)} y_\sigma, \quad \forall l_i(\sigma) \in \mathscr{L}^{\text{PH}}(\sigma), \sigma \in \Sigma, \tag{13.18}$$

where the objective function (13.16) computes the total system cost of selecting and operating services, building and hauling blocks and cars, and classifying cars. (Note that the costs related to car handling in yards captured by $\Phi(x_a^k, z_\eta^k)$ in the service-selection case, Sect. 3.1, are included in the classification and blocking

costs.) Constraints (13.17) link the building of blocks to the selection of the services which move them, while constraints (13.18) enforce the service capacity limits in terms of cars hauled on each service leg given the blocks that can be moved on that leg.

3.3.2 Path-Based Integrated SND

It is well-known that one may write network design models in arc and path forms (see, e.g., Chap. 2), each with its own pros and cons. For example, path formulations generally involve a huge number of variables, but are amenable to decomposition and the utilization of solution techniques based on column generation. For service network design, services are paths in the physical network. The same is true for blocks in the freight railroad SND case. Hence the "arc" or "path" qualification in this context refers generally to the representation of the freight flows on the service network, that is, to the modeling of the demand itineraries.

We define, for the path-version of the model of Sect. 3.3.1, the set $\mathscr{P}^k$ of itineraries, *paths through the service network* $G = (\mathscr{N}^{\text{PH}}, \mathscr{L}^{\text{PH}})$, with potential set of blocks $\mathscr{B}$, which may be used to transport all or some part of the volume of demand $k \in \mathscr{K}$. Indicators detail the definition of each itinerary, linking the arc and path flow variables on blocks and at classification yards. Let $\delta_\eta^p = 1$ when the cars following the itinerary $p \in \mathscr{P}^k$ (re-)classify at yard $\eta \in \mathscr{N}$, and 0, otherwise. Similarly, let $\delta_{l_i(\sigma)}^p$ and δ_b^p to equal 1 when the itinerary p includes the service leg $l_i(\sigma) \in \mathscr{L}^{\text{PH}}$ and the block $b \in \mathscr{B}$, respectively, and 0, otherwise.

Let us assume that the unit itinerary (path) cost may be computed as the sum of the unit transportation and classification costs associated to the services and yard classification activities making it up. This is a wide-spread hypothesis in the literature and practice and it does correspond to many actual problem settings. The unit cost of itinerary $p \in \mathscr{P}^k$ then becomes $c_p^k = \sum_{\sigma \in \Sigma} \sum_{l_i(\sigma) \in \mathscr{L}^{\text{PH}}(\sigma)} c_{l_i(\sigma)}^k \delta_{l_i(\sigma)}^p + \sum_{\eta \in \mathscr{N}} c_\eta^{\text{C}} \delta_\eta^p$.

With respect to decision variables, the service and block selection variables defined previously are also part of this model. Flow variables, however, are defined as h_p^k, standing for the quantity of commodity $k \in \mathscr{K}$ assigned to its itinerary $p \in \mathscr{P}^k$. The path formulation of the integrated service design & block selection with classification model may be written as:

$$\text{Minimize} \quad \sum_{\sigma \in \Sigma} f_\sigma y_\sigma + \sum_{b \in \mathscr{B}} f_b y_b + \sum_{k \in \mathscr{K}} \sum_{p \in \mathscr{P}^k} c_p^k h_p^k \tag{13.19}$$

Subject to constraints (13.5), (13.11)–(13.15), (13.17)–(13.18), and

$$\sum_{p \in \mathscr{P}^k} h_p^k = d^k, \qquad \forall k \in \mathscr{K}, \tag{13.20}$$

$$h_p^k \geq 0, \qquad \forall p \in \mathscr{P}^k,\ k \in \mathscr{K}, \tag{13.21}$$

$$x_b^k = \sum_{p \in \mathscr{P}^k} \delta_b^p h_p^k, \qquad \forall b \in \mathscr{B}, k \in \mathscr{K}. \tag{13.22}$$

Constraints (13.20) ensure all demand is moved to its final destination (and enforces the flow conservation at the nodes of the network), while constraints (13.21) define the domain of the path flow variables. Finally, relations (13.22) are definitional constraints linking the arc and path flow variables on blocks.

3.3.3 Advanced Path-Based Integrated SND

Most planning models in the literature and this chapter, including the previous path-based one, assume strict capacity restrictions, no waiting due to congestion, additive path characteristics with the composing arcs and nodes, and linear cost (and time) functions in the decision-variable values. While reasonable in many cases, and making solving somewhat easier, such hypotheses limit the scope of the planning models. We discuss these limitations in the following, together with a modeling framework addressing them. Although presented for the static SND case, the discussion and modeling framework are general, including for the time-dependent case.

Capacity constraints are ubiquitous in practice and OR models. They obviously apply at operation time. One cannot load more containers on a car than it can physically accommodate. At the tactical planning however, one is generally less concerned with how the capacity of each individual car, train or yard is filled up, and much more interested in identifying the service network and flow distribution for an optimal usage of those resources and capacities. Thus, the assignment of some quantity of freight to a particular service resulting in exceeding its capacity may indicate either that the frequency of the service should be increased, or that some less important (in terms of priority or delay costs) traffic should pass to another service. The formulation of strict capacity constraints would prevent, however, the detection and handling of such a situation by the solution method. Moreover, it is also known that assigning more flow to a service or a yard does not result in stopping the system activities. It rather translates, in practice, either in delays for the respective freight, which will wait for the next departure, or in additional resources being brought on line. Increased costs and, possibly, delays, occur in both cases.

Treating such limits as *utilization targets* rather than strict constraints, and including in the objective function penalties for the over utilization of the capacity, addresses these issues. Consider, to illustrate, the service-leg capacity constraints (13.18). Let α_σ be the unit penalty cost of overloading service $\sigma \in \Sigma$. A rather simple utilization-target penalty may be written for each service leg of the service as

$$\Psi_{l_i(\sigma)}(y_\sigma)=\alpha_\sigma\left(\max\left\{0,\sum_{k\in\mathcal{K}}\sum_{b\in\mathcal{B}(l_i(\sigma))}x_b^k-u_{l_i(\sigma)}y_\sigma\right\}\right)^n,\ \forall l_i(\sigma)\in\mathscr{L}^{\mathrm{PH}}(\sigma),\sigma\in\Sigma, \tag{13.23}$$

where n represents a certain degree of unwillingness of letting the tactical plan overloading the resource too much.

A similar approach may be used to model resource availability limitations with respect, e.g., to locomotives or railroad cars of particular types. Global limits for the network or targeted by yard may be handled in this way. *Service-quality targets* for demand, specifying, for example, the total duration of the origin to destination activity chain, may also be addressed in this way. Such targets may have been publicized or promised to specific customers only. Delay (time) measures are associated in such problem settings to the yard and long-haul movement activities and, thus, to services and itineraries. A capacity-like constraints may then be imposed on the itinerary duration with respect to the service target. But, again, such a constraint would provide the opportunity to trade off a penalty on some ODs against a more significant reduction in costs in other regions, generated by a more cost-or time-efficient deployment of resources. An itinerary-specific penalty may then be computed as the difference between the itinerary duration and the service target, weighted by a demand-specific penalty cost, which may represent the penalty the railroad must pay when delivering late or an estimation of the potential market-share loss.

Penalties defined according to (13.23) represent a rather strict translation of the capacity and target constraints, which does not account for the well-known fact that getting close to the capacity limits is not suitable in several cases. Consider, for example, yard classification capacities. Trains bring cars in batches, each according to its more or less followed schedule. These cars are then handled by a limited number of resources, with varying characteristics and performance measures, proper to the yard type. Queuing phenomena and congestion are a direct consequence of such situations, which may be observed for various yard activities (e.g., classification, container loading/unloading, and interservice block transfer), as well as for long-haul movements when several freight and, possibly, passenger trains share a single or double-track with restricted capacity (due, e.g., to too few or too short sidings). Models based on queuing theory have been proposed in the literature to account for these phenomena. Queuing models or networks of queues were proposed and used mostly to simulate operations. Such models are very detailed, however, and generally yield non-continuously differentiable functions, which is very hard to handle, particularly for large-scale formulations. Consequently, continuous non-linear functions were proposed to approximate such congestion behavior within network-wide SND formulations addressing tactical-planning issues.

Let the decision-variable vectors $\mathbf{y}$ and $\mathbf{h}$ indicate a given level of service in Σ and flow distribution in $\mathscr{P}$, respectively. Let then

$F_\sigma(\mathbf{y}, \mathbf{h})$: Total (fixed) cost of operating service $\sigma \in \Sigma$;
$F_b(\mathbf{y}, \mathbf{h})$: Total (fixed) cost of building, hauling, and dismantling block $b \in \mathscr{B}$;
$C_p^k(\mathbf{y}, \mathbf{h})$: Total unit cost for itinerary $p \in \mathscr{P}^k$;
$\Psi(\mathbf{y}, \mathbf{h})$): Penalty terms capturing various relations and restrictions, such as the limited service capacity.

The objective function of the path-based integrated SND formulation may then be written in a general form

$$\text{Minimize} \sum_{\sigma \in \Sigma} F_\sigma(\mathbf{y}, \mathbf{h}) y_\sigma + \sum_{b \in \mathscr{B}} F_b(\mathbf{y}, \mathbf{h}) y_b + \sum_{k \in \mathscr{K}} \sum_{p \in \mathscr{P}^k} C_p^k(\mathbf{y}, \mathbf{h}) h_p^k + \Psi(\mathbf{y}, \mathbf{h}), \tag{13.24}$$

where costs depend on the complete status of the network as given by the $\mathbf{y}$ and $\mathbf{h}$ vectors at the corresponding iteration of the solution method. It may be very difficult, in practice, to develop and calibrate such general functions for railroads of realistic dimensions and complexity. The impact of distant activities on a given service, block, or yard may be hard to evaluate and might not be very important. Consequently, most models adopting this approach consider nearby interactions only, within each individual yard, for example.

Notice that service-quality targets and time-related measures open a number of possibilities for more flexible modeling and accounting for the cost of time, even in a so-called static formulation. Thus, one may model frequency (or connection) delays encountered when one must transfer between two services with different frequencies and, thus, with differences in their presence at the same yard. One may also define time-related costs for services and demands, and use them to weight the total time required to go from origin to destination through the various system activities. The delay cost for demand usually represents the penalties in case of late delivery. It may also be used to model priorities, the time sensitiveness for certain commodities, and customer-service classes, a higher cost pushing the corresponding demand flows more rapidly through the system. For services, these costs may represent depreciation values and inventory costs as well as, according to the railroad's accounting practices manpower or energy-consumption costs.

The objective function thus computes a generalized cost, in the sense that it may include a broad range of productivity measures related to terminal and transportation operations, in terms of time, cost, and quality and reliability of the service offered. This enhances modeling refinement and flexibility, providing the opportunity for enhanced trade-off analyses among cost and service-quality objectives, as well as among the impact and value of activities and resource utilization. The gains come, however, at the price, the SND formulations taking the form of nonlinear integer multicommodity network design problems.

3.4 Service & Block Generation and SND Models

The static and time-dependent models described in this chapter proceed as most network design models do, by selecting from a set of candidate, potential, arcs. More precisely, for the railroad case, from potential sets of services Σ and blocks $\mathscr{B}$. How these sets are generated is a valid question, which is relevant for most SND applications (see, e.g., Chap. 12), and more so for railroad transport with its several levels of consolidation and combinatorial complexity. We briefly discuss the topic in the context of static SND, but everything applies to time-dependent settings too. In fact, the latter case presents even greater challenges, the time dimension of the problem setting exacerbating the combinatorial multiplication of the number of potential services and blocks.

The cardinality of these sets may be very large. Consider, for example, that a service may, in theory, be defined between every pair or yards in the network, on every possible physical path, with every possible combination of stops at the yards on that physical path, as well as for every type of service in terms of power, capacity, speed, priority, and so on and so forth. Blocks may then be defined similarly but on the network made up of all those potential service legs. Obviously, full enumeration of all potential services and blocks is not more realistic for railroads than for the other modes or other situations of a similar nature, e.g., crew scheduling in passenger and freight transportation. On the one hand, full enumeration yields problem dimensions extremely difficult to manipulate and solve, even when stringent feasibility checks are enforced. On the other hand, trying to generate "good" services and blocks only, with respect to limits on costs and time, for example, generally eliminates elements contributing to very good or optimal solutions. Hence, a systematic service and block generation procedure tightly linked to or part of SND formulations is needed.

Partial targeted enumeration is appropriate in many practical cases when the plan for the next season is based on the previous one, adjusted for the trends and predictions in demand, prices, and the regulatory environment identified by management. The past service and block networks are then enriched with a number of additional possibilities reflecting these trends and predictions. Yet, even in such situations, one faces the problem of missing elements required for very good solutions, and a more systematic procedure is required.

The goal is thus to include the generation of the service and block sets into SND formulations. We illustrate the difficulty of arc-based formulations focusing on the case when one starts with the set of potential services, the blocks are to be generated together with the tactical plan, at most one block is created for each pair of yards.

The problem description and notation of Sect. 3.3.1 (and previous ones) apply except for the block definition, which is reduced to the origin and destination yards, $O(b)$ and $D(b)$, respectively, of block $b \in \mathscr{B}$. The path in the service network $\mathscr{L}^{\text{PH}}$ is thus not part of the input, but is an output of the optimization problem. Thus, at most $|\mathscr{B}| = |\mathscr{N}^{\text{PH}}|^2 - |\mathscr{N}^{\text{PH}}|$, which is relatively small. This gain in problem dimensions and number of integer block-selection variables is paid for, however, in increasing numbers and complexity of constraints, as shown in the following.

Given the updated definition of $\mathscr{B}$, the fixed cost f_b includes only the cost relative to building the bloc at the origin yard and dismantling it at destination. The inter-service transfer costs must be identified and, then, computed separately. We model this through a function Φ, which can be of any form but accounts for the characteristics and operating policies of the yard and the number of blocks to transfer. Other than the $y_\sigma, \sigma \in \Sigma$, $y_b, b \in \mathscr{B}$, and $x_b^k, b \in \mathscr{B}, k \in \mathscr{K}$, decision variables of Sect. 3.3.1, we define

- $y_{bl_i(\sigma)} = 1$ if block $b \in \mathscr{B}$ is moving on service leg $l_i(\sigma) \in \mathscr{L}^{\text{PH}}(\sigma), \sigma \in \Sigma$, and 0 otherwise;
- $z_\eta^b = 1$ if block $b \in \mathscr{B}$ is transferred at yard $\eta \in \mathscr{N}^{\text{PH}}$ from one service to another, and 0 otherwise.

The SND formulation with block generation minimizes the total system cost (13.25), computed as the service- and bloc-selection costs, plus the cost of moving cars on blocks given the service leg used to haul the block, the car classification cost at yards where blocks are generated, and the cost of transferring blocks between services.

$$\begin{aligned}\text{Minimize} \quad & \sum_{\sigma\in\Sigma} f_\sigma y_\sigma + \sum_{b\in\mathscr{B}} f_b y_b + \sum_{k\in\mathscr{K}} \sum_{b\in\mathscr{B}} \sum_{l_i(\sigma)\in\mathscr{L}^{\text{PH}}} c^k_{l_i(\sigma)} x_b^k y_{bl_i(\sigma)} \\ & + \sum_{k\in\mathscr{K}} \sum_{\eta\in\mathscr{N}} \sum_{b\in\mathscr{B}^+(\eta)} c^{\text{C}}_\eta x_b^k + \sum_{\eta\in\mathscr{N}} \sum_{b\in\mathscr{B}} \Phi(z_\eta^b) \end{aligned} \tag{13.25}$$

Subject to (13.10)–(13.13), and

$$\sum_{\sigma\in\Sigma} \sum_{l_i(\sigma)\in\mathscr{A}_\eta^+} y_{bl_i(\sigma)} - \sum_{\sigma\in\Sigma} \sum_{l_i(\sigma)\in\mathscr{A}_\eta^-} y_{bl_i(\sigma)} = \begin{cases} y_b, & \text{if } \eta = O(b), \\ -y_b, & \text{if } \eta = D(b), \\ 0, & \text{otherwise,} \end{cases} \quad \forall\, \eta \in \mathscr{N}, b \in \mathscr{B}, \tag{13.26}$$

$$\begin{aligned} y_{bl_i(\sigma)} - y_{bl_{i+1}(\sigma)} &\le z^b_{\eta_i}, \\ \forall b \in \mathscr{B}, i = 1, \ldots, n(\sigma) - 1,\ & l_i(\sigma) = (\eta_{i-1}, \eta_i) \in \mathscr{L}^{\text{PH}}(\sigma), \sigma \in \Sigma, \end{aligned} \tag{13.27}$$

$$\sum_{k\in\mathscr{K}} \sum_{b\in\mathscr{B}} x_b^k y_{bl_i(\sigma)} \le u_{l_i(\sigma)} y_\sigma, \ \forall l_i(\sigma) \in \mathscr{L}^{\text{PH}}, \sigma \in \Sigma, \tag{13.28}$$

$$y_{bl_i(\sigma)} \in \{0,1\}, \quad \forall\, b \in \mathscr{B}, l_i(\sigma) \in \mathscr{L}^{\text{PH}}, \sigma \in \Sigma, \tag{13.29}$$

$$y_\sigma \in \mathbb{Z}_+, \quad \forall\, \sigma \in \Sigma, \tag{13.30}$$

$$y_b \in \{0,1\}, \quad \forall\, b \in \mathscr{B}, \tag{13.31}$$

$$z_\eta^b \in \{0,1\}, \quad \forall\, b \in \mathscr{B}, \eta \in \mathscr{N}, \tag{13.32}$$

$$x_b^k \ge 0, \quad \forall\, b \in \mathscr{B}, k \in \mathscr{K}. \tag{13.33}$$

Constraints (13.26) and (13.27) enforce the building of blocks conditions. The former are the block-building constraints ensuring that a single path is selected in $\mathscr{L}^{\mathrm{PH}}$ for each block from its origin to its destination. The latter, (13.27), are linking relations ensuring that block transfers are accounted for, and that the corresponding costs will be paid, by setting the transfer decision variable z_{η}^{b} to 1 whenever block b is transferred from a service to a different one at yard η, except at the origin and destination of the block. Constraints (13.28) are the flow-service linking and capacity constraints, given the service legs moving the block transporting the cars. Restrictions on the decision variables are enforced by constraints (13.29)–(13.33).

It is noteworthy that a sleeker set $\mathscr{B}$ and, thus, fewer y_b selection variables, translates into a large number of constraints, namely (13.26) and (13.27), and decision variables, $y_{bl_i(\sigma)}$, required to build the blocks out of service legs and transfers. It is also worth noticing that both the objective function (13.25) and constraints (13.28) are non linear. This is not surprising and the issue can be addressed, but it does not make the problem dimensions smaller nor the formulation easier to address. These observations are not unique to railroad planning, but have been made in many other settings, e.g., crew scheduling and vehicle routing. Dynamic path generation techniques, based on Column Generation techniques, have been applied in such settings and appear promising for service network design and railroad tactical planning. Most work has still to be undertaken in this field, which constitutes a challenging but interesting research direction, particularly when the time dimension is explicitly considered as in the models of the following sections.

4 Time-Dependent SND and Integrated Planning

We now turn to time-dependent SND formulations, also known as *Scheduled Service Network Design* (*SSND*) models. As discussed in Sect. 2.3, SSND targets time-dependent problem settings, explicitly representing the time-related characteristics of demand, in terms of availability time at origin and due time at destination. To answer the requirements of time-dependent demand, the time characteristics of the service the railroad offers is also explicitly represented, in terms of a schedule stating the departure and arrival times at each of the yards on the route of each individual service. The aim is thus not only to select the service network, but also the schedule of the selected services to address the time-dependent demand.

Most SSND models address a broader set of planning issues than selecting services only, and are generally qualified as *integrated-planning methods*. The general SSND modeling framework presented herein for the integrated freight-railroad planning problem addresses the main tactical-planning issues: service selection and scheduling, blocking and classification, train makeup, and freight routing. The goal is to minimize the total cost of the system, while satisfying demand with the available resources. The framework is extended in Sect. 5 to address exiting schedules, intermodal traffic, and resource management at the tactical planning level.

Most of the notation is introduced in Sect. 2.3 and Table 13.1. It is briefly recalled and completed in the following. SSND formulations are built on time-space networks, defined over the total duration of the schedule length, most of them using a discrete representation of time. The mixed-integer network design formulation presented in this section follows this classical approach, adapted for the multiple interrelated decisions involved in the problem setting. The model is thus built on a *multi-layer time-space* network $\mathscr{G} = (\mathscr{N}, \mathscr{A})$, using a discrete representation of the schedule length $\mathbf{T}$. The nodes are representations of the physical nodes (yards, mainly) at all time periods. Two nodes, $\eta_t^{\text{IN}}, \eta_t^{\text{OUT}} \in \mathscr{N}$, are created for each yard $\eta \in \mathscr{N}^{\text{PH}}$ and time period $t \in \mathscr{T}$, to capture all the traffic coming into the node and going out of the node, respectively.

Arcs represent movement on service legs and holding activities at nodes. Recalling that the scheduled service plan is to be applied repeatedly over the tactical-planning horizon, $\mathscr{G}$ takes on a *cyclic* nature. The network thus reflects the fact that activities and decisions do not stop with the end of the schedule length, but rather involve the next application of the scheduled plan. Thus, for example, when building a week-long schedule, a service may start on Friday and arrive at destination the following Tuesday. We model these situations by having the corresponding arcs *wrap around*. In modeling terms, this means that the destination for an arc with origin at time t and a duration which would make it arrive at a time $> \mathbf{T}$ is defined at a time $t' < t$ through a *modulo* computation.

We present the SSND model on a *three-layer* time-space network, schematically illustrated in Fig. 13.4. Multi-layer networks make up a general methodology with applications in transport and telecommunications. Integrated railroad planning offers a very good illustration. Each layer represents the activities and decisions which focus on the particular type of flow, cars, blocks, and services. The arcs

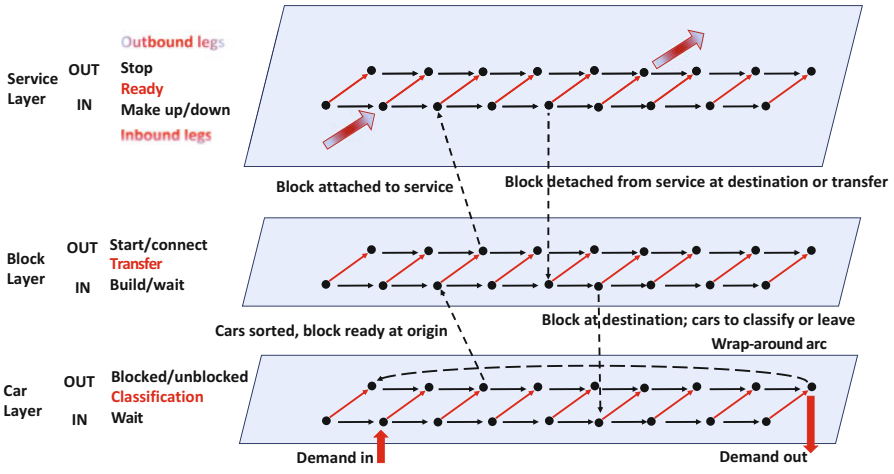

Fig. 13.4 Three-layer time-space SSND network

in each layer stand for the operations taking place in terminals, impacting principally the corresponding flows. The inter-layer arcs model the assignment and consolidation/de-consolidation of these flows, from cars to blocks to services and vice-versa, the classification, transfer, and makeup activities being modeled within each layer. Each layer is thus a "complete" time-space network. In the model we present in this chapter, the time and schedule length definitions are the same for all layers. Consequently, the node set $\mathscr{N}$ is the union of the node sets in all layers. Similarly, the arc set $\mathscr{A}$ is the union of all arcs, movement and holding, in all layers, plus the inter-layer arcs supporting the flow movements between layers. To simplify the presentation, we do not detail the notation based on layers, except when needed to avoid confusion (Sect. 6 points to literature with detailed notations).

Cars, loaded and empty, enter the network through an IN node within the *car layer*. They exit the network through an OUT node in the same layer. Section 5 adds a container loading/unloading layer within the context of intermodal rail transportation. The holding arcs between IN nodes represent the waiting time for classification. In this formulation, the yard classification capacity is defined by time period and is associated to the classification links. The interplay between this limit and the flow of cars requiring classification determines for how long (given by the number of waiting links) cars have to wait before being sorted. The blocked/unblocked links capture the time waiting, once classified, on the appropriate block track for the appropriate number of cars to accumulate and the block to be ready. They also represent the possible waiting at the final destination when arrived too early. The car layer illustration also shows a wrapped-around (blocked/unblocked) arc (for clarity of illustration we do not show such arcs on the other layers).

The car and the block layers are linked through two types of arcs. The first moves the sorted and blocked cars, i.e., the block, to the block origin in the block layer. Symmetrically, block-to-car arcs move the cars on a block at destination back to the car layer, to be either re-classified (not at their destination, yet) and put on a new block, or to exit the yard for final distribution (cars at their final destination).

The *block layer* focuses on selecting the blocks, the block-to-service assignment, and the associated operations of attaching a block to a train or detaching the block from a train. The attach-to-train operation involves the new blocks at their origins, and blocks transferring at intermediary yards. The detach-from-train operation applies to blocks at their destinations, and blocks requiring transfer to a different train. The build/wait arcs in the block layer capture the blocks ready to be attached/transferred, while the start/connect arcs capture the waiting for the departure service. Transfer arcs stand for the physical operations of moving blocks to or between trains and may model limited capacity and force waiting.

Block-to-service and service-to-block arcs link the block and service layers. The former connect OUT nodes in the block layer to IN nodes in the service layer and represent adding the blocks, new or transferred, to trains at the given period. The latter connect IN service nodes to IN block nodes, taking the blocks off their current services for transfer or dismantling (when at destination).

Arcs in the *service layer* represent service selection, schedules, and operations (Sect. 2.3). Services start from an OUT node at the specified starting time period. Direct services terminate their routes at an IN node at the period specified in its schedule. Multi-stop services, arrive at the first stop at an IN node, stop at the yard for the specified number of periods, leave from the corresponding OUT node, move to the IN node of the next stop, and so on and so forth until the final destination. Differently from the two previous layers, the service layer thus includes explicitly moving arcs representing the service legs of the potential services according to their schedules. Two such arcs only are sketched in Fig. 13.4. The figure also shows the make up/down arcs on which trains stay while taking out blocks at destination or being transferred or taking in blocks for the outbound move. The stop arcs complete the stop length until the departing service leg. The ready arcs complete the modeling of the service yard activity and may be used to model yard capacity limitations.

A few notes before introducing the SSND formulation. A demand *itinerary* is then a path in the three-layer time-space network between the IN and OUT nodes in the car layer representing the corresponding origin and destination yards at the availability and due dates, respectively. When the cars are delivered before the due date, they wait on the blocked/unblocked arcs (the model may be easily modified to account for late deliveries and penalties). The itinerary then includes the wait arcs, a classification arc, reaching an OUT node, where they wait for accumulation of cars on blocked/unblocked arcs. Once the block is formed, the traffic goes up to the block layer, where the block journeys to its destination yard, where the block is dismantled and the cars return down to the car layer at an IN node for final delivery or re-classification. The journey continues as described in the latter case until the final destination of the demand.

A block journey may be similarly described, from an IN node in the block layer, through build/wait arcs, a transfer arc, and start/connect arcs until the OUT node when the block is ready to be put on the train through an inter-layer arc. Once in the service layer, the block journeys through a sequence of service legs interspersed with movements down to the block layer, the arcs involved in the transfer operation, and then back up to the service layer and the next segment on the block route.

It is noteworthy that "moving arcs" may be shown in the car layer by projecting on it the appropriate blocks making up each itinerary. Similarly, moving arcs in the block layer are obtained by projecting the corresponding service legs. Notice, finally, that parallel arcs may exist in the service layer standing either for train movements following different physical routes between the two yards with the same departure and transit times, or for services of different types (e.g., regular cargo and intermodal) sharing the same infrastructure.

The parameter and decision-variable definitions follow the pattern of all other models in this chapter, with the provision that, all service, block, and itinerary sets follow the time-space network definition with IN and OUT nodes for each yard. Thus

- f_σ: Fixed selection cost for service $\sigma \in \Sigma$;
- $y_\sigma \in \mathbb{Z}_+$: Frequency of service $\sigma \in \Sigma$. The selection decision and its fixed cost concern, as usual, the complete service definition in the service layer. For representation purposes, they may be associated to the moving arc of the first leg of the service;
- f_b: Fixed cost of building and moving block $b \in \mathscr{B}$;
- $y_b = 1$, if block $b \in \mathscr{B}$ is built, and 0, otherwise. Similarly to service selection, block-selection decision and fixed cost apply to the complete block definition in the block and service layers. For representation purposes, they may be associated to the arc out of the $O(b)$ IN node on the block layer;
- $x_a^k \geq 0$, continuous flow variable representing the volume of commodity $k \in \mathscr{K}$ on arcs $a \in \mathscr{A}$, becoming x_b^k, when $a = b \in \mathscr{B}$, and $x_{l_i(\sigma)}^k$, when $a = l_i(\sigma) \in \mathscr{L}, \sigma \in \Sigma$;
- c_a^k: Hauling and time unit cost for commodity $k \in \mathscr{K}$ on the service legs, i.e., on the moving arcs of the service layer; the handling cost on car-classification (car layer) and block-transfer (block layer) arcs; and the time cost on the holding arcs of the car, block and service layers;
- u_a: Capacity of classification (u_η^{C}, car layer) and transfer arcs (u_η^{B}, block layer).

The integrated scheduled service network design model may be formulated as:

$$\text{Minimize} \quad \sum_{\sigma \in \Sigma} f_\sigma y_\sigma + \sum_{b \in \mathscr{B}} f_b y_b + \sum_{k \in \mathscr{K}} \sum_{a \in \mathscr{A}} c_a^k x_a^k \tag{13.34}$$

Subject to

$$\sum_{a \in \mathscr{A}_\eta^+} x_a^k - \sum_{a \in \mathscr{A}_\eta^-} x_a^k = \begin{cases} d^k, \text{ if } \eta = O(k), \\ -d^k, \text{ if } \eta = D(k), \\ 0, \text{ otherwise, } \forall \eta \in \mathscr{N}, k \in \mathscr{K}, \end{cases} \tag{13.35}$$

$$\sum_{k \in \mathscr{K}} x_b^k \leq u_b y_b, \qquad \forall\, b \in \mathscr{B}, \tag{13.36}$$

$$\sum_{k \in \mathscr{K}} \sum_{b \in \mathscr{B}(l_i(\sigma))} x_b^k \leq u_{l_i(\sigma)} y_\sigma, \qquad \forall l_i(\sigma) \in \mathscr{L}, \sigma \in \Sigma, \tag{13.37}$$

$$\sum_{k \in \mathscr{K}} x_a^k \leq u_a^{\mathrm{C}}, \qquad \forall a \in \text{classification arcs} \subset \mathscr{A}, \tag{13.38}$$

$$y_b \leq \delta_\sigma^b y_\sigma, \qquad \forall b \in \mathscr{B}, \sigma \in \Sigma, \tag{13.39}$$

$$\sum_{b \in \mathscr{B}} y_b \leq u_a^{\mathrm{T}}, \qquad \forall a \in \text{transfer arcs} \subset \mathscr{A}, \tag{13.40}$$

$$\sum_{b \in \mathscr{B} \mid O(b) = \eta} y_b \leq u_\eta^{\mathrm{B}}, \qquad \forall \eta \in \text{in nodes on block layer} \subset \mathscr{N}, \tag{13.41}$$

$$\sum_{\sigma \in \Sigma \mid O(\sigma)=\eta} y_\sigma \leq u_\eta^{\mathrm{M}}, \qquad \forall \eta \in \text{in nodes on service layer} \subset \mathscr{N}, \tag{13.42}$$

$$y_\sigma \in \mathbb{Z}_+, \qquad \forall \sigma \in \Sigma, \tag{13.43}$$

$$y_b \in \{0, 1\}, \qquad \forall b \in \mathscr{B}, \tag{13.44}$$

$$x_a^k \geq 0, \qquad \forall a \in \mathscr{A}, k \in \mathscr{K}, \tag{13.45}$$

where the objective function (13.34) computes the total cost of selecting and operating services, building, transferring, and hauling blocks, and classifying, blocking, transferring, and hauling cars. Constraints (13.35) enforce flow conservation at all nodes on all layers. Constraints (13.36) and (13.37) limit the loads of blocks and service legs in terms of cars hauled, respectively, while constraints (13.38) perform the same task on the yard classification arcs on the car layer. Constraints (13.39) link the building of blocks to the selection of the services moving them, while constraints (13.40) limit the number of blocks which can transfer simultaneously at a yard on the block layer. Finally, constraints (13.41) and (13.42) limit the number of blocks and trains, respectively, which can be built at each yard during the schedule length (extending the formulation to enforce capacities by time periods is straightforward).

5 Extending the SSND

The previous model is general and may be extended to account for additional railroad features and planning issues. It may be extended, for example, to path-based models, integrating the handling of non-additive characteristics and penalty or congestion representations of capacity limits. We focus in this section on three extensions, the first handling given service schedules and continuous-time representation, the second addressing the intermodal railroad case, while the third is concerned with the integration of resource-management concerns into tactical SSND.

It is not unusual for tactical railroad planning to be performed by two different teams within the railroad, one focusing on the service design, with somewhat rough blocking concerns, the other starting from the service network selected and focusing on the detailed classification, blocking, and final train makeup decisions. This case is particularly observed when intermodal traffic is concerned. The service design still addresses the entire system and all traffic classes, while only the services dedicated to intermodal freight are within the scope of classification and blocking planning.

A given service network and schedule induces a continuous time discretization of the schedule length, corresponding to the departure and arrival time instants of each service at each of the yards in its route. Figure 13.5 illustrates the service layer of a multi-layer SSND, at a particular yard, for two services stopping at that yard for different lengths of time. The network is greatly simplified, as the only IN and OUT nodes in the layer correspond to the arrival and departure time instances,

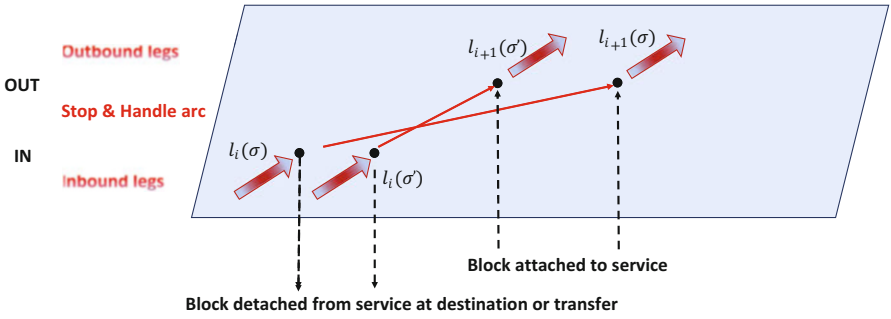

Fig. 13.5 Service layer with given SSND service schedule

respectively, of each service. The block-to-train attach and detach activities are concentrated in the IN and OUT nodes of the service. This defines the time instances in all the other layers. The network is further simplified as the activities related to each service while in the yard are associated to a unique Stop & Handle service-specific arc, the arcs in the other layers being simplified in a similar way. Note that holding arcs capturing waiting periods on the car or block layer are not eliminated. The duration and costs of the inter- and intra-layer arcs may be adjusted to account for this discretization. Thus, for example, when the availability time of demand $k \in \mathcal{K}$ is before the first node defined in the network, $o(k)$ is set to the time instance of that node and the waiting cost and time of the demand are adjusted accordingly. The resulting SSND still presents the same degree of complexity as the general network design problems. The simplification of the network provides the means, however, to address much larger problem dimensions with commercial mixed-integer software, corresponding, for example, to the cases of several North American railroads.

Intermodal traffic and operations are the topic of the second extension of the SSND we discuss (to simplify the presentation, car classification activities are not included). As already indicated, intermodal demand must be loaded onto cars at the origin terminal and must be unloaded at destination. This major difference with regular traffic, induces an additional layer to the SSND network. The container layer corresponds to the entry and exit of container OD demand into and out of the railroad system. Holding arcs capturing waiting prior to blocking or prior to final delivery at destination are part of the container layer. Inter-layer arcs between the container and car layers support the container-to-car assignment and loading, in one direction, and the container unloading, in the opposite direction (when the service network and schedule is given, as above, the time instance of those arcs correspond to a possible block for the cars, which corresponds to a possible service for the block.)

Representing the many types of containers and container-compatible cars, as well as the large number of rules governing the loading of containers on railroad cars is a major challenge for tactical (and strategic) modeling. Indeed, one cannot explicitly include the huge number of feasible loading patterns into aggregated

service network design formulations. This is still a challenging research issue. We present an approximation procedure, which proved appropriate for tactical planning in the North American context, where double-stacking is largely used (the approximation may be easily adapted for the simple case of single staking). Double stack means that two containers may be loaded one on top of the other, following very stringent rules, e.g., a 53-feet long container may be positioned on top of a 40-feet container or two 20-feet containers, but the opposite patterns are generally not allowed.

The approximation is based on the observations that (1) cars built to haul containers (also called Well cars) come in several multi-platform configurations, a platform providing two slots, one on the bottom and one on top, for containers of a given length; (2) 40-feet and 53-feet are the two main container types, the former world wide (with the 20-feet ones, which can be approximated as two 20s = one 40), the latter mainly in North America (43-feet and 45-feet may be modeled through the 53s), and correspond to the main platform-loading capability of cars; (3) cars able to carry 53-feet containers are more expensive (to rent and operate) than the other, more regular, cars and one therefore aims to use not more than necessary; (4) *length* is a major constraining feature for trains and blocks both in terminals and when put on trains. Then, given that the length of a car is determined for the most part by the number of platforms it provides, the approximation makes use of the *platform*, of a given type, as a loading unit, and considers 40- and 53-feet long container and platform types.

Consider the basic loading rules for these container and platform types:

- *40-feet platform*: (1) single 40-feet container in the bottom slot; (2) two 40-feet containers in the two slots; (3) one 40-feet container in the bottom slot and one 53-feet in the top slot; 4) empty;
- *53-feet platform*: (1) all the configurations of a 40-feet platform; (2) single 53-feet container in bottom slot; (3) two 53-feet containers in the two slots.

The procedure aims to "maximize" the number of forty-feet platforms per unit of train length. Consequently, when the number of 53-feet containers (nb_{53}) is greater than or equal to the number of 40-feet containers (nb_{40}), the 40s should be placed in bottom slots and the 53s on top, as much as possible. The numbers of 53-feet (nbp_{53}) and 40-feet (nbp_{40}) platforms are given by (13.46) and (13.47), respectively. These four parameters are then the basis for decision-variable definitions for the numbers of containers and platforms, of each type, assigned to each block. The values of these variables are governed by constraints implementing relations (13.46) and (13.47), and are used to compute costs and enforce capacities.

$$nbp_{53} = \max\{0, \lceil (nb_{53} - nb_{40})/2 \rceil\}, \tag{13.46}$$

$$nbp_{40} = \lceil (nb_{53} + nb_{40})/2) \rceil - nbp_{53}. \tag{13.47}$$

Similar to any other transportation mode (Chap. 12), rail services require resources to operate, people, cars, and locomotives, in particular. While a rich

literature addresses resources-management issues (Sect. 6), most research and contributions target operational planning issues, in which the service network and schedule are given. Multicommodity network flow optimization (linear formulations with integer flow variables) is the methodology of choice in those cases. Given the scope and length limits of this book, we do not detail this methodology. We rather focus on the challenge of representing resource-management concerns at the level of tactical planning and within SSND models. The goal is not to integrate the details of scheduling and managing resources, but rather to capture the main impacts of resource management on the tactical plan. This is a broad and largely unexplored research area, which needs significant work, not only for railroads, but for consolidation-based transportation in general.

The initial and current developments focus mainly on the availability and routing of material resources (also sometimes called "assets"). They follow from the operational needs and developments aimed at balancing assets, as well as from the growing requirements of efficiently running a scheduled railroad without a level of resources higher than what is strictly needed. With respect to the first aspect, recall that trade and, thus, demand flows are unbalanced, different products and quantities flowing, say, from West to East than vice-versa. This results in vehicles, cars and locomotives, of certain types, becoming available after providing service at yards where they are not needed for the next cycle of operations but missing at others. Empty cars and locomotives must therefore be moved, *repositioned*, for the next cycle. So-called "full-asset utilization" policies illustrate the second aspect, where resources are ideally expected to circulate continuously in the network (accounting for the maintenance requirements, of course) supporting the scheduled services.

The basic translation of the previous discussion into a mathematical formulation are the *design-balancing* constraints

$$\sum_{a \in \mathscr{A}_\eta^+} \sum_{\sigma \in \Sigma} \delta_\sigma^a y_\sigma - \sum_{a \in \mathscr{A}_\eta^-} \sum_{\sigma \in \Sigma} \delta_\sigma^a y_\sigma = 0, \ \forall \eta \in \mathscr{N}, \tag{13.48}$$

where $\delta_\sigma^a = 1$ if service $\sigma \in \Sigma$ operates on arc $a \in \mathscr{A}$ (i.e., one of its legs defines arc a, which terminates or initiates at node η), and 0, otherwise.

The design-balancing constraints (13.48) state that the number of resources brought into a yard by all incoming services equals the number of resources taken out of the yard by the selected outgoing services. The constraints assume that one unit of resource is required by each occurrence of each service (recall that $y_\sigma \in \mathbb{Z}_+$), where the unit may represent a locomotive, a group of locomotives, a car, a group of cars, or a crew. In this sense, resources are assimilated to services. The formulation is still rich, however. Several resource types may be defined, with the corresponding design-balancing constraints. Moreover, the δ_σ^a parameters may be refined and tailored for particular resources and restrictions of services or yards. On the other hand, it is difficult to represent cost and utilization characteristics of particular resource types, such as the assignment of resources to particular home yards, which

is a regular feature of transportation systems, and the maximum duration before returning home for maintenance.

A path-based formulation provides the modeling tool to address these shortcomings. The idea is to explicitly define the sequence of tasks a resource has to undertake as a *cycle* θ of service legs, and holding arcs in time-dependent formulations. Let Θ represent the set of resource cycles, and u^{R} the quantity of resources available in the system. Let $\mathscr{L}(\theta)$ be the set of service legs $l_i(\sigma) \in \mathscr{L}(\sigma), \sigma \in \Sigma$, the resource cycle $\theta \in \Theta$ supports in sequence, with the definitional parameter $\delta^{\theta}_{l_i(\sigma)} = 1$ if service leg $l_i(\sigma) \in \mathscr{L}(\sigma), \sigma \in \Sigma$, is supported by resource cycle $\theta \in \Theta$, and 0, otherwise. Let f_θ represent the fixed cost of selecting and operating resources on cycle $\theta \in \Theta$, and let us define the resource selection decision variable $y_\theta \in \mathbb{Z}_+$ as the number of resources executing cycle $\theta \in \Theta$.

The total resource cost $\sum_{\theta \in \Theta} f_\theta y_\theta$ is then added to the objective function of the SSND model. Constraints (13.49) are added to the formulation to connect the selection of the sufficient number of resources and the requirements of the selected services.

$$\sum_{\theta \in \Theta} \delta^{\theta}_{l_i(\sigma)} y_\theta = y_\sigma, \qquad \forall l_i(\sigma) \in \mathscr{L}(\sigma), \sigma \in \Sigma, \tag{13.49}$$

$$\sum_{\theta \in \Theta} y_\theta \leq u^{\mathrm{R}}, \qquad \forall \eta \in \mathscr{N}. \tag{13.50}$$

Notice that constraints (13.50), limiting the number of resources selected to the availability of resources, is not needed as stated, the resource cost driving the number of selected resources to the minimum required to run the system. The constraints may be refined, however, to represent resource availability at each yard, when cycles (i.e., the resources executing them) are linked to a home yard as its respective domicile, from where it originates and where it returns at the end. Notice also that the attributes of resource cycles may be controlled during generation, e.g., one may forbid generating cycles longer than permitted by the rules governing the resource type and its home yard (see also the discussion of Sect. 3.4). As discussed in Chap. 12, this approach is extremely promising for linking resource management and service network design for tactical planning, but much more research is required on modeling the various cases and objectives and on developing efficient solution methods for large problem instances.

6 Bibliographical Notes

The literature on railroads and railroad planning goes many years back, generally presenting application-based contributions and reflecting often industry practice. Several survey papers synthesize the story and contributions of operations research, including network design methodology, to railroad planning, e.g., Assad (1980b);

Dejax and Crainic (1987); Crainic (1988); Crainic and Laporte (1997); Cordeau et al. (1998); Crainic (2000); Newman et al. (2002); Crainic (2003); Ahuja et al. (2005a); Crainic and Kim (2007); Bektaş and Crainic (2008); Crainic (2009); Yaghini and Akhavan (2012).

Early contributions focus on single problems or combinations of a limited number of issues. These include the pioneering service selection, routing and makeup model of Assad (1980a), the train routing and the scheduling model of Morlok and Peterson (1970). Huntley et al. (1995) developed a computerized routing and scheduling system for CSX Transportation, while Ireland et al. (2004) developed a planning system for Canadian Pacific Railway that brought together several separate procedures without building a comprehensive model.

Blocking has often been addressed as a separate problem to be solved before the selection of services. Bodin et al. (1980) proposed one of the first such models, a non-linear mixed-integer formulation, blocking delays being dependent on the number of cars assigned to each block. Newton (1996); Newton et al. (1998); Barnhart et al. (2000) formulates the blocking problem as a network design model, arcs representing candidate blocks among classification yards. No fixed costs are associated to blocks, the number of blocks which can be build at each yard being limited through budget constraints. A path-formulation and a branch-and-price algorithm (Barnhart et al. 1998) are proposed in the first two contributions, while a dual-based Lagrangian relaxation is used in the latter to decompose the problem into easier-to-address subproblems, namely a continuous multicommodity flow problem and an integer block formulation that selects blocks satisfying yard capacity constraints (addressed by a branch-and-cut algorithm). Ahuja et al. (2007) follows the same approach in an arc-based formulation, proposing a large neighborhood search algorithm aimed at addressing large problem instances. Jha et al. (2008) then proposes arc and path-based time-space formulations for the block-to-train assignment problem. The latter formulation proved the most flexible and amenable to be efficiently solved either with an *a priori* set of paths, or a dynamic-path generation procedure. Metaheuristics for the arc or path-based formulations are proposed by, e.g., Yaghini et al. (2011, 2012); Yue et al. (2011). Uncertainty has been rarely addressed in models targeting freight railroad planning. A few contributions addressing blocking problems have been proposed (e.g., Yang et al. 2011; Hasany and Shafahi 2017), but much more research is required in this area.

Service selection was also often treated separately of the other planning problems (Assad 1980a; Morlok and Peterson 1970; Martinelli and Teng 1996; Yaghini et al. 2014). It has also been addressed in two steps, service routes and frequencies being determined first (e.g., Marín and Salmerón 1996a,b; Goossens et al. 2004), the schedule being constructed in a second step, based on the routing patterns yielded by the first step (e.g., Nozick and Morlok 1997; Brännlund et al. 1998; Caprara et al. 2002, 2006; Cacchiani et al. 2010; Cacchiani and Toth 2012).

Models aiming for integration of tactical planning issues were proposed simultaneously with those targeting individual issues described above. Crainic et al. (1984) presents what is probably the first service network design model addressing

simultaneously the selection of services and their frequencies, car classification and blocking, train makeup, and freight routing. It is noteworthy that the model integrates the distribution of empty cars through one or several origin-destination demand matrices (generated through demand-distribution models from the surplus and penury levels at yards, which were derived from the loaded car demand). These matrices become commodities to be handled simultaneously with all other OD commodities in the problem. The model takes the form of a static path-based, nonlinear network design formulation accounting for congestion and accumulation-delay phenomena in yards and on rail tracks, service-quality targets, and trade-offs between operating and time-related costs. Block fixed costs were not included; they were approximated through the accumulation-delay costs and the limits on yard-specific block dimensions. A heuristic solution method was used to address realistically-sized problem instances derived from the case of a large North-American railroad.

Crainic and Rousseau (1986) generalizes the model for the tactical planning of consolidation-based multicommodity multimode freight transportation systems. Bektaş et al. (2010) later studied Lagrangean-based relaxation and decomposition algorithms. The authors show that, first, non-linearities may be handled efficiently through decomposition and, second, that the relaxation of the flow constraints, which yields an arc decomposition, has computationally better convergence properties than the dualization of the capacity constraints. These results are very encouraging for this demanding but important research topic.

A number of contributions followed toward the end of the 80; and during the 90's. Haghani (1989) presents a model which attempts to combine train routing and scheduling, make-up, as well as empty car distribution on a space-time network with fixed travel times and pre-specified traffic rules. A heuristic is used to address a somewhat simplified version of the model and illustrate the interest of integrated planning. The model proposed by Keaton (1989, 1992) aims to determine the pairs of yards to connect by direct services, and whether to offer more than one train a day, as well as the routing of freight and the blocking of rail cars. The service network is made up of one network for each pair of yards in the system with positive demand. Arcs represent trains and connections in yards, as well as *a priori* determined blocking alternatives. Gorman (1998) starts from the previous model aiming to design a scheduled operating plan that followed as much as possible the particular operation rules of a given railroad. An innovative tabu-enhanced genetic search metaheuristic is used to generate candidate train schedules, which are evaluated on their economic, service, and operational performances. On relatively small but realistic problems, the metaheuristic performed well and was used for strategic scenario analysis for a major North-American railroad. All these contributions model blocking through classification costs, rather than explicit blocking decision variables.

Zhu et al. (2014) propose a cyclic multi-layer time-space SSND model, which appears to be the first comprehensive formulation to select the train services and schedules to operate for a given schedule length, the car classification policies,

the blocks to build in each terminal with their routes within the service network, the train makeup, and the demand itineraries using these services and blocks. The authors also introduce a matheuristic solution methodology combining slope scaling, a dynamic block-generation mechanism, long-term memory-based perturbation strategies, and an ellipsoidal search, i.e., a new intensification mechanism to thoroughly explore very large neighborhoods of elite solutions in an efficient way using information from the history of the search. Experimental results show that the proposed solution method is efficient and robust, yielding high-quality solutions for realistically-sized problem instances. The model of Sect. 4 is based on this work.

As already mentioned, the management of resources, or assets, has a long history of research and applications, yielding a rich corpus of literature, starting with the pioneering work on empty cars and containers (Bomberault and White 1966; White 1968; White and Bomberault 1969; White 1972) and locomotives (Florian et al. 1976). Dejax and Crainic (1987); Cordeau et al. (1998); Piu and Speranza (2014) present detailed surveys and syntheses of the literature until the end of the 80's. Most of this literature and developments address operational planning issues, e.g., distribution and routing. Network flow optimization is the methodology of choice in this field, evolving from the initial transportation problem models to the contemporary integer-flow time-space multicommodity formulations integrating various practical rules and constraints (e.g., Ahuja et al. 2005b; Vaidyanathan et al. 2008b,a; Balakrishnan et al. 2016; Bouzaïene-Ayari et al. 2016; Piu et al. 2015; Ortiz-Astorquiza et al. 2021; Miranda et al. 2020).

Few contributions aimed until rather recently to integrate resource management concerns into tactical planning service network design models. We mentioned the modeling of empty cars as an additional demand proposed by Crainic et al. (1984). Close to the network design methodology, Joborn et al. (2004) proposes a time-space formulation to select kernel paths to move groups of empty cars between pairs of yards by using the residual capacity of a given set of scheduled services. A kernel path corresponds to a sequence of services, which can move the group of cars between its origin and destination, plus waiting and inventory arcs. A particular characteristic of the formulation is that fixed costs and capacity constraints are not associated to the design arcs of the network (services), but rather to the kernel path, that is, to a set of design arcs, which increases the difficulty to solve it. A tabu search metaheuristic was proposed to efficiently address the problem and to show that the proposed model achieves the looked-for economies of scale.

Resource-management considerations were integrated into service network design models through the contributions of Andersen et al. (2009a,b); Pedersen and Crainic (2007); Pedersen et al. (2009) (see also Andersen and Christiansen 2009, where the modeling framework of Crainic et al. (1984) is used for the strategic analysis of a new intermodal service in Europe). Pedersen and Crainic (2007); Pedersen et al. (2009) focus on the management of one asset type, namely, locomotives. The authors introduce the concept of design-balanced SND and present a tabu search metaheuristic to address it (see also Vu et al. 2013; Chouman and Crainic 2015, for metaheuristics targeting the same problem). Andersen et al. (2009a,b) enlarges the scope of the models to include resource cycles, cyclic

schedules and the coordination/synchronization of several railroads and navigation services at particular junction points. The authors also show that cycle-based formulations provided more modeling flexibility and computational efficiency. A branch-and-price algorithm is proposed by Andersen et al. (2011) for the cycle-based SSND formulation. It is also noteworthy that the papers mentioned in this paragraph also offer insights into modeling tightly time-constrained systems, as well as rail and rail-road intermodal terminals. Car classification and blocking issues were not addressed, however, nor the train makeup problem, or the case of multiple resource types with complex resource-to-service assignment rules. Integrating resource management and service network design for tactical and strategic planning is still a very active, important, and challenging research area.

We complete this brief literature survey with the case of planning intermodal railroad transport. As indicated previously, intermodality presents additional challenges. In particular, the assignment and loading of containers to cars must be explicitly integrated into the planning methods, while accounting for the multiple and complex loading rules (see, e.g., Mantovani et al. 2017, for a detailed description of the complex rules governing the loading of containers on rail cars of diverse characteristics, particularly when double stacking is performed). An additional layer to the SSND time-space network illustrates this additional complexity. Yet, one finds few contributions targeting rail intermodal transport. Newman and Yano (2000) proposes a day-of-week uncapacitated (in the number of trains one may make up in a yard and operate on a line) train scheduling model, to determine whether intermodal OD demand should be moved by a direct or indirect, through a main yard, service. A decomposition method yielding simpler problem settings provides encouraging results. Morganti et al. (2020) proposes a blocking SSND model for intermodal services, when the service network and schedule are given, which determines container-to-car assignments and loading, blocking, service makeup, and demand itineraries. The model of Sect. 5 is inspired by this paper. Very good experimental results data from a large North American railroad were obtained using a well-known commercial software. Much research work is still needed in this area. Two directions in particular. First, integrate resource management concerns (see Kienzle et al. 2021, for very encouraging developments) and service selection decisions. Second, similar to all the other facets of research on railroad planning, algorithmic developments are needed to address efficiently large problem instances.

7 Conclusions and Perspectives

Rail transportation is very important in economic and environmental terms. Its many benefits follow, however, from a complex organization with several levels of consolidation, e.g., freight into cars, cars into blocks, and blocks into trains. Network design, through its service network design formulations with or without explicit schedules, offers models and methods to address these challenges and efficiently support the planning of freight railroad operations to achieve economic

and service-quality objectives. Rail is actually more complex than most other consolidation-based transportation modes and, thus, challenges both the modeling and algorithmic facets of operations research, in general, and network design, in particular.

This chapter presents these issues, challenges, and contributions. It illustrates the long and successful history of the connections between rail tactical planning and operations research development. This is a vibrant research area, in continuous development based on the mutually beneficial interactions between new realities in the field and new methodological developments.

Many research perspectives have been identified during the presentation, particularly in Sect. 6. We do not repeat them here. We recall the challenges of continuing to study the integration of the main components of railroad planning at the tactical level, from scheduled service selection to resource management. Among the other interesting and challenging research directions, we single out two. First, the study and explicit integration of uncertainty into the planning models. Uncertainty may be found in demand (e.g., volume, realization of temporal characteristics, etc.) as well as supply in terms of travel and yard-activity times. How one predicts these elements, both at the level of day-to-day operation and as more rare but disturbing incidents, and how one integrates them, and the options to alleviate their negative effects, into SSND models constitutes a significant research challenge. Second, revenue management starts to interest freight railroads. There is still little literature on revenue management in freight transportation (air cargo is somewhat of an exception) and even less when rail is concerned. Research is needed in the revenue mechanisms as applied to rail transport, as well as in the interaction between planning and these mechanisms.

We conclude recalling the challenge of efficient solution methods for large problem instances. Algorithms are required for multi-layer time-space networks, in both their linear and non-linear incarnations. Decomposition methods and parallel optimization offer one interesting avenue for development. Dynamic generation of paths – services, blocks, demand itineraries, resource cycles –, or of the time-space network, or a combination of both, offer an equally interesting complimentary avenue.

References

Ahuja, R. K., Cunha, C. B., & Şahin, G. (2005a). Network models in railroad planning and scheduling. In *Tutorials in Operations Research INFORMS 2005*, INFORMS (pp. 54–101), Published online: 14 Oct 2014.

Ahuja, R. K., Jha, K. C., & Liu, J. (2007). Solving real-life railroad blocking problems. *Interfaces, 37*, 404–419.

Ahuja, R. K., Liu, J., Orlin, J. B., Sharma, L. A., & Dand, S. (2005b). Solving real-life locomotive-scheduling problems. *Transportation Science, 39*(4), 503–517.

Andersen, J., & Christiansen, M. (2009) Designing new European rail freight services. *Journal of the Operational Research Society, 60*, 348–360.

Andersen, J., Crainic, T. G., & Christiansen, M. (2009a). Service network design with asset management: Formulations and comparative analyzes. *Transportation Research Part C: Emerging Technologies, 17*(2), 197–207.

Andersen, J., Christiansen, M., Crainic, T. G., & Grønhaug, R. (2011). Branch-and-price for service network design with asset management constraints. *Transportation Science, 46*(1), 33–49.

Andersen, J., Crainic, T. G., & Christiansen, M. (2009b). Service network design with management and coordination of multiple fleets. *European Journal of Operational Research, 193*(2), 377–389.

Assad, A. A. (1980a). Modelling of rail networks: toward a routing/makeup model. *Transportation Research Part B: Methodological, 14*, 101–114.

Assad, A. A. (1980b). Models for rail transportation. *Transportation Research Part A: Policy and Practice, 14*, 205–220.

Balakrishnan, A., Kuo, A., & Si, X. (2016). Real-time decision support for crew assignment in double-ended districts for U.S. freight railways. *Transportation Science, 50*(4), 1139–1393.

Barnhart, C., Jin, H., & Vance, P. H. (2000). Railroad blocking: A network design application. *Operations Research, 48*(4), 603–614.

Barnhart, C., Johnson, E. L., Nemhauser, G. L., Savelsbergh, M. W. F., & Vance, P. H. (1998). Branch-and-price: column generation for solving huge integer programs. *Operations Research, 46*(3), 316–329.

Bektaş, T., Chouman, M., & Crainic, T. G. (2010). Lagrangean-based decomposition algorithms for multicommodity network design with penalized constraintsm. *Networks, 55*(3), 272–280.

Bektaş, T., & Crainic, T. G. (2008). A brief overview of intermodal transportation. In G. D. Taylor (Ed.), *Logistics engineering handbook* (Chap. 28, pp. 1–16). Boca Raton, FL: Taylor and Francis Group.

Bodin, L. D., Golden, B. L., Schuster, A. D., & Romig, W. (1980). A model for the blocking of trains. *Transportation Research Part B: Methodological, 14*(1), 115–120.

Bomberault, A. M., & White, W. W. (1966). Scheduling empty box cars. Technical Report. IBM New York Scientific Center, Hawthorne, N.Y.

Bouzaïene-Ayari, B., Cheng, C., Das, S., Fiorillo, R., & Powell, W. B. (2016). From single commodity to multiattribute models for locomotive optimization: A comparison of optimal integer programming and approximate dynamic programming. *Transportation Science, 50*(2), 366–389.

Brännlund, U., Lindberg, P. O., Nõu, A., & Nielsson, J. E. (1998). Railway timetabling using lagrangian relaxation. *Transportation Science, 32*(4), 358–369.

Cacchiani, V., Caprara, A., & Toth, P. (2010). Scheduling extra freight trains on railway networks. *Transportation Research Part B: Methodological, 44*(2), 215–231.

Caprara, A., Fischetti, M., & Toth, P. (2002). Modeling and solving the train timetabling problem. *Operations Research, 50*(5), 851–861.

Caprara, A., Monaci, M., Toth, P., & Guida, P. L. (2006). A Lagrangian heuristic algorithm for a real-world train timetabling problem. *Discrete Applied Mathematics 154*, 738–753.

Cacchiani, V. & Toth, P. (2012). Nominal and robust train timetabling problems. *European Journal of Operational Research, 2019*, 727–737.

Chouman, M., & Crainic, T. G. (2015). Cutting-plane matheuristic for service network design with design-balanced requirements. *Transportation Science, 49*(1), 99–113.

Cordeau, J. F., Toth, P., & Vigo, D. (1998). A survey of optimization models for train routing and scheduling. *Transportation Science, 32*(4), 380–404.

Crainic, T. G. (1988). Rail tactical planning: issues, models and tools. In L. Bianco & A. La Bella (Eds.) *Freight Transport Planning and Logistics* (pp. 463–509). Berlin: Springer.

Crainic, T. G. (2000). Network design in freight transportation. *European Journal of Operational Research, 122*(2), 272–288.

Crainic, T. G. (2003). Long-Haul freight transportation. In R. W. Hall (Ed.), *Handbook of Transportation Science* (2nd edn., pp. 451–516). Norwell, MA: Kluwer Academic Publishers.

Crainic, T. G. (2009) Service design models for rail intermodal transportation. In L. Bertazzi, M. G. Speranza, & J. A. E. E. van Nunen (Eds.), *Lecture Notes in Economics and Mathematical Systems* (Vol. 619, pp. 53–67). Berlin: Springer.

Crainic, T. G., Ferland, J. A., & Rousseau, J. M. (1984). A tactical planning model for rail freight transportation. *Transportation Science, 18*(2), 165–184.

Crainic, T. G., Kim, K. H. (2007). Intermodal transportation. In C. Barnhart & G. Laporte (Eds.), *Transportation, Handbooks in Operations Research and Management Science* (Vol. 14, Chap. 8, pp 467–537). Amsterdam: North-Holland.

Crainic, T. G., & Laporte, G. (1997). Planning models for freight transportation. *European Journal of Operational Research, 97*(3), 409–438.

Crainic, T. G., & Rousseau, J. M. (1986). Multicommodity, multimode freight transportation: A general modeling and algorithmic framework for the service network design problem. *Transportation Research Part B: Methodological, 20*, 225–242.

Dejax, P. J., & Crainic, T. G. (1987). A review of empty flows and fleet management models in freight transportation. *Transportation Science, 21*(4), 227–247.

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

Goossens, J. W., van Hoesel, S., & Kroon, L. (2004). A branch-and-cut approach for solving railway line-planning problems. *Transportation Science, 38*(3), 379–393.

Gorman, M. F. (1998). An application of genetic and tabu searches to the freight railroad operating plan problem. *Annals of Operations Research 78*, 51–69.

Haghani, A. E. (1989). Formulation and solution of combined train routing and makeup, and empty car distribution model. *Transportation Research Part B: Methodological, 23*(6), 433–452.

Hasany, R. M., & Shafahi, Y. (2017). Two-stage stochastic programming for the railroad blocking problem with uncertain demand and supply resources. *Computers & Industrial Engineering, 106*, 275–286.

Huntley, C. L., Brown, D. E., Sappington, D. E., & Markowicz, B. P. (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, K. C., Ahuja, R. K., & Şahin, G. (2008). New approaches for solving the block-to-train assignment problem. *Networks, 51*(1), 48–62.

Joborn, M., Crainic, T. G., Gendreau, M., Holmberg, K., & Lundgren, J. T. (2004). Economies of scale in empty freight car distribution in scheduled railways. *Transportation Science, 38*(2), 459–464.

Keaton, M. H. (1989). Designing optimal railroad operating plans: lagrangian relaxation and heuristic approaches. *Transportation Research Part B: Methodological, 23*(6), 415–431.

Keaton, M. H. (1992). The impact of train timetables on average car time in rail classification Yards. *Journal of the Transportation Research Forum, 32*(2), 345–354.

Kienzle, J., Crainic, T. G., Frejinger, E., & Bisaillon, S. (2021). The intermodal railroad blocking & railcar fleet management planning problem. Technical Report. CIRRELT-2021, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et les transports, Université de Montréal, Montréal, QC, Canada

Mantovani, S., Morganti, G., Umang, N., Crainic, T. G., Frejinger, E., & Larsen, E. (2017). The load planning problem for double-stack intermodal trains. *European Journal of Operational Research, 267*(1), 107–119.

Marín, A., & Salmerón, J. (1996a). Tactical planning of rail freight networks. Part I: Exact and heuristic methods. *European Journal of Operational Research, 90*, 26–44.

Marín, A., & Salmerón, J. (1996b). Tactical planning of rail freight networks. Part II: local search methods with statistical analysis. *European Journal of Operational Research, 94*, 43–53.

Martinelli, D. R., & Teng, H. (1996). Optimization of railway operations using neural networks. *Transportation Research Part C: Emerging Technologies, 4C*(1), 33–49.

Miranda, P., Cordeau, J. F., & Frejinger, E. (2020). A time-space formulation for the locomotive routing problem at the Canadian National Railways. Technical Report. CIRRELT-2020-19, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et les transports, Université de Montréal, Montréal, QC, Canada

Morganti, G., Crainic, T. G., Frejinger, E., & Ricciardi, N. (2020). Block planning for intermodal rail: methodology and case study. *Transportation Research Procedia, 47*, 19–26.

Morlok, E. K., & Peterson, R. B. (1970). A final report on a development of a geographic transportation network generation and evaluation model. In *Proceedings of the Eleventh Annual Meeting, Transportation Research Forum* (pp. 99–103)

Newman, A. M., Nozick, L. K., & Yano, C. A. (2002). Optimization in the rail industry. In P. M. Pardalos & M. G. C. Resende (Eds.), *Handbook of Applied Optimization* (pp. 704–718), New York, NY: Oxford University Press.

Newman, A. M., & Yano, C. A. (2000). Centralized and decentralized train scheduling for intermodal operations. *IIE Transactions, 32*(1), 743–754.

Newton, H. N. (1996). Network design under budget constraints with application to the railroad blocking problem. Ph.D. Thesis. Industrial and Systems Engineering, Auburn University, Auburn, Alabama, U.S.A.

Newton, H. N., Barnhart, C., & Vance, P. H. (1998). Constructing railroad blocking plans to minimize handling costs. *Transportation Science, 32*(4), 330–345.

Nozick, L. K., & Morlok, E. K. (1997). A model for medium-term operations planning in an intermodal rail-truck service. *Transportation Research Part A: Policy and Practice, 31*(2), 91–108.

Ortiz-Astorquiza, C., Cordeau, J. F., & Frejinger, E. (2021). The locomotive assignment problem with distributed power at the Canadian National Railway Company. *Transportation Science, 55*(2), 510–531.

Pedersen, M. B., & Crainic, T. G. (2007). Optimization of intermodal freight service schedules on train canals. Publication CIRRELT-2007-51, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport, Montréal, QC, Canada.

Pedersen, M. B., Crainic, T. G., & Madsen, O. B. G. (2009). Models and tabu search meta-heuristics for service network design with asset-balance requirements. *Transportation Science, 43*(2), 158–177.

Piu, F., Première Kumar, V., Bierlaire, M., & Speranza, M. G. (2015). Introducing a preliminary consists selection in the locomotive assignment problem. *Transportation Research Part E: Logistics and Transportation Review, 82*, 214–237.

Piu, F., & Speranza, M. G. (2014). The locomotive assignment problem: A survey on optimization models. *International Transactions in Operational Research, 21*(3), 327–352.

Vaidyanathan, B., Ahuja, R. K., Liu, J., & Shughart, L. A. (2008a). Real-life locomotive planning: new formulations and computational results. *Transportation Research Part B: Methodological, 42*(2), 147–168.

Vaidyanathan, B., Ahuja, R. K., & Orlin, J. B. (2008b). The locomotive routing problem. *Transportation Science, 42*(4), 492–507.

Vu, D. M., Crainic, T. G., & Toulouse, M. (2013). A three-stage matheuristic for the capacitated multi-commodity fixed-cost network design with design-balance constraints. *Journal of Heuristics, 19*, 757–795.

White, W. W. (1968). A program for empty freight car allocation. Technical Report. 360D.29.002, IBM Contributed Program Library, IBM Corporation, Program Information Department, Hawthorne, N.Y.

White, W. W. (1972). Dynamic transshipment networks: an algorithm and its application to the distribution of empty containers. *Networks, 2*(3), 211–236.

White, W. W., & Bomberault, A. M. (1969). A network algorithm for empty freight car allocation. *IBM Systems Journal, 8*(2), 147–171.

Yaghini, M., & Akhavan, R. (2012). Multicommodity network design problem in rail freight transportation planning. *Procedia Social and Behavioral Sciences, 43*, 728–739.

Yaghini, M., Momeni, M., & Sarmadi, M. (2014). Solving train formation problem using simulated annealing algorithm in a simplex framework. *Journal of Advanced Transportation, 48*, 402–416.

Yaghini, M., Seyedabadi, M., & Khoshraftar, M. M. (2011). Solving railroad blocking problem using ant colony optimization algorithm. *Applied Mathematical Modelling, 35*(12), 5579–5591.

Yaghini, M., Seyedabadi, M., & Khoshraftar, M. M. (2012). A population-based algorithm for the railroad blocking problem. *Journal of Industrial Engineering International, 8*(8), 1–11.

Yang, L., Gao, Z., & Li, K. (2011). Railway freight transportation planning with mixed uncertainty of randomness and fuzziness. *Applied Soft Computing 11*, 778–792.

Yue, Y., Zhou, L., Yue, Q., & Fan, Z. (2011). Multi-route railroad blocking problem by improved model and ant colony algorithm in real world. *Computers & Industrial Engineering, 60*, 34–42.

Zhu, E., Crainic, T. G., & Gendreau, M. (2014). Scheduled service network design for freight rail transportations. *Operations Research 62*(2), 383–400.