Chapter 12
Service Network Design

Teodor Gabriel Crainic and Mike Hewitt

1 Introduction

The term *Service Network Design* (*SND*) is generally used to designate a set of issues and decisions aimed to plan the activities and resources of the *supply* side of a transportation system, in order to satisfy a given or estimated *demand* efficiently, profitably, and within the quality standards agreed upon with the customers generating this demand. *Service* is then understood as operating a vehicle, or a convoy, e.g., a railroad train, between two stations/terminals in the network, with or without intermediary stops, to transport a single or a group of people or freight loads. The service follows a given route on the appropriate infrastructure, and displays a number of physical, e.g., vehicle type and capacity, and operational, e.g., departure time, total trip duration and cost, characteristics. While all transportation systems and carriers offer "services" to their customers, SND occurs mainly in the context of *consolidation*-based transportation, an umbrella term for companies and systems, the *carriers*, which group and transport within the same vehicle several people who contracted the trip separately or several freight loads of different customers. In all cases, the alternative of a dedicated, direct transport is not economically justifiable or even feasible. Public-transport carriers in urban areas, by bus, light rail, and collective taxi, and those providing interurban transport by coach, train or airplane "consolidate" passengers who do not want or can move by

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

M. Hewitt
Information Systems and Supply Chain Management Department, Quinlan School of Business, Loyola University Chicago, Chicago, IL, USA
e-mail: mhewitt3@luc.edu

T. G. Crainic et al. (eds.), *Network Design with Applications to Transportation and Logistics*, https://doi.org/10.1007/978-3-030-64018-7_12

a dedicated vehicle between their respective origins and destinations. Postal and small-package transportation companies, less-than-truckload (LTL) motor carriers, railroads, ocean/maritime liner navigation companies, and land- and water (coastal, river, etc)-based intermodal carriers perform similar services for freight. Noticeable are the "new" transportation-system types introduced for urban, e.g., City Logistics, and interurban, e.g., Physical Internet and truck platooning, settings, which are heavily based on consolidation and resource sharing. Carriers may be publicly or privately owned/operated, while groups of carriers, operating under some form of cooperation agreement, may also be involved.

Carriers need to be profitable, while consolidation raises two challenges. First, that the vehicle movements, that is, the services offered, cannot be planned to address the demand of individual potential customers, but must satisfy as closely as possible the requirements of as many potential customers as possible (while probably not satisfying any of them entirely; contrasting taxi and public-transport services illustrates the point). This has implications for the service network, including on topology, i.e, where to propose and operate services, timelines, i.e., when to operate services, and performance measures, e.g., cost, efficiency, and quality of service. Second, that operations need to be efficient from the point of view of using the carrier's material and human resources. Indeed, one observes, on the one hand, a continuous increase in the size of vehicles on long-haul routes, e.g., mega container ships (capacity exceeding 20,000 twenty-foot equivalent containers), 120 to 130-car long trains running on the North American rail networks, and large passenger aircraft types. One also notices, on the other hand, that the cost of operating a service is greatly dependent upon the costs of vehicles and power units used for transport. System efficiency and cost reductions may then be achieved through economies of scale capacity utilization, obtained by assigning the most appropriate vehicles, and other associated resources (power units, people, etc.), to each movement, filling them well with the passengers or freight requiring transport, and routing them through multi-service *itineraries* and inter-service transfers at terminals. The availability of resources constrains the range of alternatives, however, while multi-service itineraries may imply additional costs and delays at terminals.

Trade-offs must thus be achieved in planning the service network, to balance customer demand for faster and cheaper transportation, on the one hand, and the pursuit of economies of scale and profitable and efficient carrier activities, on the other hand. Trade-offs must also be achieved among the various components of the carrier transportation system and operations as improving one aspect often has negative implications for other aspects, e.g., increasing the number of times a service is operated during a certain time interval improves customer service but may decrease the availability of resources for other services as well as increase congestion in terminals, thus deteriorating customer service. *Service Network Design* aim to address these issues network-wide and determine the services and itineraries to operate.

SND is closely related to network design. We emphasize these relations in this chapter, as well as the particular characteristics applications bring to SND. Several chapters in this book address SND in the context of such applications,

namely, public transport (Chap. 17), motor carriers (Chap. 14), railroads (Chap. 13), navigation (Chap. 15), and City Logistics (Chap. 16). The goal of this chapter is to present a comprehensive overview of the general SND methodology, in terms of models, solution methods and utilization, that cuts across application fields. To focus the presentation, however, we will use in the following the vocabulary of consolidation-based freight carrier planning.

The chapter is organized as follows. Section 2 recalls the structure and main components of the physical and service networks of consolidation-based freight carriers, as well as the associated tactical planning issues, and the main *Service Network Design* (*SND*) formulation classes with their utilization within the carrier-planning processes. Section 3 is dedicated to the static problem setting and formulations, while Sect. 4 introduces the explicit representation of time and time-related attributes in the basic SND models. Section 5 broadens the scope of SND methodology to integrate the management of the resources required to operate the selected services. Addressing uncertainty within SND is the topic of Sect. 6. Section 7 proposes an historical view of the field, in terms of the models and main solutions methods specifically developed for various SND settings. We conclude in Sect. 8 with a number of research issues we deem important and challenging.

2 Problem Settings

We initiate this section with a brief description of the physical and service networks typical of consolidation-based freight carriers. We then proceed to discuss the associated planning of operations and introduce the main classes of service network design models proposed to address them.

2.1 Consolidation-Based Freight Carriers

Carriers providing consolidation-based services operate on an infrastructure network made up of terminals connected by physical, e.g., highways and rail tracks, or conceptual, e.g., maritime and air corridors, links. Terminals come in several designs and sizes, targeting particular transportation modes, e.g., rail marshaling/consolidation yards and stations, LTL motor-carrier breakbulk and regional terminals, and maritime and river ports. Terminals may be owned/managed by and dedicated to the carrier, e.g., railroad yards and LTL breakbulk terminals, or may be shared by several carriers irrespective of ownership and management, e.g., maritime ports and terminals, intermodal terminals, passenger airports, etc. Inter-terminal links may also be proprietary (but may still be used by other carriers for a fee), e.g., rail tracks in North America, or shared, e.g., rail tracks in Europe and roads and highways mostly everywhere.

Carriers operate single or multi-modal networks on the infrastructure. LTL motor carriers and railroads operating exclusively trucks and trains, respectively, are usually identified as single-mode. Postal/express-courier services, City Logistics systems, and container intermodal transportation often involve more than one transportation mode, the transfer of loads from one to the next taking place at intermodal terminals. Notice, however, that many carriers traditionally classified as single mode actually operate multi or intermodal networks, the latter occurring when freight packaged at origin, e.g., in containers, is not handled before it is unpacked at destination. Railroads, owning LTL motor carriers, and maritime shipping companies, owning railroads or motor carriers, illustrate this case when they plan services and freight movements on the entire network. Moreover, particular vehicle and convoy configurations (in terms of power, speed, capacity, etc.) are also often identified for planning purposes as "modes" with their own tariffs, due to their different performances in terms of costs and travel time. We therefore address multi-modal networks in this chapter, each service being of a particular "mode" according to the infrastructure, vehicle and convoy configuration, speed and priority, etc .

Consolidation transportation carriers are organized into so-called hub-and-spoke networks. One identifies two main categories of nodes in such a network. The largest category consists of *local/regional terminals* where most of the demand from the corresponding regions is brought in to be transported by the system, and where the demand flows terminate their trips before being distributed to their final destinations. Rail stations, LTL regional terminal, most deep-sea and river/canal ports belong to this type. The *hubs* make up the second category. One finds in this category LTL breakbulks, major classification/blocking railroad yards, and major maritime ports for intermodal (container-based) traffic such as Hong Kong, Singapore, and Rotterdam. While these terminals play the same role as the regional terminals for their hinterlands, their main role is to *consolidate* the flows in and out of their associated regional terminals for efficient long-haul transportation and economies of scale.

Carriers offer service between origin and destination (*OD*) points corresponding to their terminals. The volume (or value or both) of most of the OD demands, identified in the following as *commodities* to recall that each may concern a specific product with specific transportation requirements, is too low, however, to justify a profitable direct service with reasonable service quality. Thus, for example, when the volume is too low with respect to the capacity of the usual vehicle for the corresponding distance, the cost of the transportation would yield tariffs few customers are willing to pay. Alternatively, waiting to fill up the vehicle with other demands to the same destination generally requires delays customers are not ready to accept. The combination of such phenomena gave rise to consolidation-based transportation, for freight and people, the number of commodities (OD demands) being significantly larger than the number of direct, origin to destination services operated by the carrier, which aims for economies of scale. Carriers thus first move low-volume loads available at a regional terminal to a hub, through what is known as feeder services. At hubs, loads are sorted (*classified* is the term used in several settings, e.g., freight railroads) and consolidated into larger flows, which are routed

to other hubs by high-frequency, high-capacity services. Loads may thus go through more than one intermediary hub before reaching the regional-terminal destination, being transferred from one service to another or undergoing re-classification and re-consolidation. Notice that, when the level or value of demand justifies it, high-frequency, high-capacity services may be run between a hub and a regional terminal or between two regional terminals. Notice also that, more than one service, of possibly different modes, may be operated between consolidation and regional terminals.

A *service* follows a route through the physical network. It may be direct, without intermediary stops from the origin of the service to its destination, or it may include stops at one or several terminals to drop and pick up loads and, eventually, vehicles, e.g., car and blocks for railroads and trailers for LTL motor carriers operating multi-trailer road trains. The route may also include stops that serve purposes other than consolidation. As an example, governmental highway safety regulations often limit the number of hours a driver may drive before resting. Yet a transportation service may be longer, in terms of drive-time, than that limit. Thus, if one driver executes the service, the duration of the service would have to reflect the driver's need to rest while en route. To reduce the service's duration, the vehicle could instead stop at an intermediate location, wherein an exchange of drivers occurs. Note this intermediate location need not be a terminal in the physical network. Instead, the driver exchange could occur at a rest stop on a highway.

The set of services the carrier selects to operate makes up the *service network*, which will be used to respond to the demand of customers who require their loads to be transported between particular origins and destinations. In most planning problems addressed with SND methodology, these locations are assumed to be the regional or hub terminals, planning processes not targeting the local pick up and delivery activities to bring loads to origin terminals to initiate transportation and to distribute them at destination. We follow this approach in this chapter.

Demand is thus multi-commodity, each commodity being defined by its specific origin (carrier terminal), destination (a different terminal), as well as commodity (product) related physical characteristics (e.g., weight and volume) and service requirements in terms of delivery conditions, type of vehicle (e.g., refrigerated, multi-platform for containers or vehicles, etc.), and so on. Two additional attributes are usually associated to each commodity. First, a unit profit or (transportation) cost, the latter often related to the vehicle type used. Second, time-related requirements, i.e., a date when it is delivered to the origin terminal, as well as a due date (or time interval) to be delivered at destination. The latter is often linked to the level of service quality required; a unit penalty cost for late delivery or a unit cost for the total delivery time or delay is generally associated to this service-quality level.

Carriers respond to demand by offering a network with more or less scheduled services. Demand itineraries will move the corresponding loads through this service network, each *itinerary* being defined by the sequence of services used and the operations to be performed (e.g., transfer or re-classification and consolidation) and intermediary terminals. The service schedule could simply be a certain *frequency of service* or *number of departures*, i.e., the number of times the "same" service

is run during the length of time the carrier uses to define its recurring operations (e.g., 1 day for LTL motor carriers, 1 week for railroads, longer for containership liners), also called *schedule length* in the following. A more precise service schedule gives the departure time from the origin terminal, arrival and departure times at each intermediary terminal, and the arrival time at destination. This information may be strict, as for most European and Canadian railroads and regular containership liners, or more of an "indicative" nature, the schedule being eventually modified to account for particular events (e.g., the need to pass a direct service for an important customer) or how much freight is already loaded. Often independent of its precision, a schedule is effective for a certain period of time, often related to seasonal variations in demand and operation conditions. We use in this chapter the term *season*, often found in the literature, to refer to this period of time during which the schedule and the services it contains are repeatedly performed.

2.2 Planning and Service Network Design Models

The planning activities consolidation-based carriers undertake may be broadly classified into three levels, similarly to most complex systems. Strategic planning involves long-term decisions on system design, operation strategies, and acquisition of major resources. Tactical planning is dedicated to building an efficient service network and schedule. Short term planning involves monitoring activities and performance, adjusting plans, managing resources and operations.

We focus on tactical planning in this chapter, as it involves the arguably strongest connection to network design, through the service network design modeling framework. We discuss at the end of this section the utilization of SND in varied contexts, including the other levels of planning.

A hub-and-spoke network concentrates the multi-commodity flows and allows a much higher frequency of service for the consolidated demand loads, while providing a more efficient utilization of resources, economies of scale for the carrier, and lower tariffs for the customers. The drawbacks of this type of organization are possibly increased delays for demand due to longer routes and more time spent going through terminals, which play a major role within consolidation-based transportation systems. This role is significantly broader than the loading and unloading of freight. Vehicle and freight sorting and consolidation, convoy make up and break down, and vehicle transfer between services are all time and resource-consuming operations performed in terminals. Indeed, if not planned properly, these additional activities and delays may cancel the benefits of the hub-and-spoke strategy.

Tactical planning aims to build a transportation plan and schedule to mitigate the drawbacks of consolidation, satisfy customer demand and service-quality requirements, and operate profitably and efficiently. It addresses the system-wide planning of operations to decide the selection and scheduling of services, the transfer and consolidation activities in terminals (as well as the convoy makeup

and dismantling for railroads, and road and barge trains), the assignment and management of resources to support the selected services, and the routing of freight of each particular demand through the resulting service network. The goal is cost-efficient operation together with timely and reliable delivery of demand according to customer specifications and the service-quality targets of the carrier.

Such planning problems are difficult due to the strong interactions among system components and decisions and the corresponding trade-offs between operating costs and service levels that need to be achieved. Consider, for example, strategies based on re-consolidation and routing through intermediate terminals, which could be more efficient when direct services are offered rarely due to low levels of traffic demand. Such strategies would then probably result in higher equipment utilization and lower waiting times at the original terminals; hence, in a more rapid service for the customer. The same strategies would also result, however, in additional unloading, consolidation, and loading operations, creating larger delays and higher congestion levels at terminals, as well as a decrease in the delivery reliability of the shipment. Alternatively, offering more direct and frequent services would imply faster and more reliable service for the corresponding traffic and a decrease in the level of congestion at some terminals, but at the expense of additional resources, thus increasing the costs of the system.

Service network design is the methodology of choice to support tactical planning of consolidation-based carriers. A SND model integrates the issues discussed above and addresses them jointly at a network-wide level. It assumes a given physical system, infrastructure, resources, operation strategies, and it optimizes for an estimation for the season of the *regular demand* (e.g., 75–80% of the pick demand on a normal operating day). It integrates two major sets of decisions, the selection of the service network, that is, the routes—origin and destination terminals, physical route and intermediate stops—and schedules, or frequencies, on which services will be operated, and the itineraries, sequences of services, terminals, and operations, used to move the freight of each demand. Operating rules specifying, for example, how resources may be assigned and handled and how cargo and vehicles may be sorted and consolidated, are often specified as part of the service network. The SND model yields a *transportation plan* specifying operations for the given *schedule length*, to be repetitively applied for the next *season*.

Static problem settings, Sect. 3, assume that neither demand, nor any other problem characteristic varies during the schedule length considered. *Time-dependent* problem settings, Sect. 4, include an explicit representation of demand and activities in time and target the selection of *scheduled* services to support decisions related to *when* services leave and arrive at terminals on their routes. In all cases, the minimization of the total operating costs is the primary optimization criterion, reflecting the traditional objectives of freight carriers and expectations of customers to "get there fast at lowest possible cost". Increasingly, however, customers not only expect low rates, but also high-quality service, measured by speed, flexibility, and reliability. Service performance measures reflecting these expectations and modeled, in most cases, by delays incurred by freight and vehicles or by the respect of predefined performance targets are then added to the objective function of the

network optimization formulation. The resulting generalized cost function thus captures the trade-offs between operating costs and service quality.

The two sections that follow detail the issues and models associated to the two major problem settings for service network design described above (static and time-dependent). We then turn to problem settings and SND models addressing the needs and challenges of integrating resource-management concerns into tactical planning (Sect. 5). Section 6 continues this discussion addressing the issue of the explicit representation of the *uncertainty* inherent to any system and human endeavor.

We conclude this general presentation with a short discussion on the utilization of tactical planning SND methodology. One first must recognize that there are different contexts and mindsets with respect to service network design, from proposing a new plan yearly or twice a year (alternating between Summer and Winter) by railroads and shipping companies, to much shorter seasons of 3–4 months, or even solving a model weekly, as is typically performed by LTL motor carriers. The same model may then be used for a much shorter period, weekly (e.g., railroads) and daily (e.g., LTL trucking and City Logistics) to re-optimize and adjust the plan and operations to current conditions. What may be adjusted depends strongly on the application context. For example, while LTL motor carriers may quite freely cancel and add truck departures, such strategies are normally much more difficult for railroads, which will rather update the actual demands assigned to blocks and trains. Obviously, the scope of the SND model may be more focused when in plan-adjustment mode, parts of the system which should not be modified being fixed.

A different class of problem settings calls upon SND models to yield plans to be applied once only. Consider, for example, the case of City Logistics when one has little or no restrictions on calling up for duty on very short notice facilities, vehicles, and people. The planning of such systems is better performed close to operation-time. The so-called *day-before* SSND models, similar to those presented in this chapter, are then used before each operation period based on updated data. Note that, in this context, there are no impacts of today's decisions on the system status and capability for the next days. When this is not the case, e.g., when transport or storage activities require several periods, the time-dependent SND models may be used in a rolling-horizon approach. Then, the SND yields decisions for "now" (a somewhat limited number of periods) and for a number of following periods. The latter are not to be implemented, but bring to the model an evaluation of the consequences of today's actions on future capabilities. Then, today's proposed actions are implemented, time is advanced, information is updated, and the process repeats.

SND models may also be used as policy and performance-evaluation tools for strategic scenarios. Operational details need to be abstracted in such cases as well as, according to the planning horizon contemplated, the demand and cost figures. Governmental institutions and funding or control organizations, such as the World or Asian Development banks, may also use SND models as a simulation tool in the context of cost-benefit analyses, with appropriate approximation of carrier and shipper characteristics. Finally, generalized service network design models

may be built to answer strategic-level decisions such as the number, locations, and characteristics of terminals to build, rent or use, the construction of dedicated infrastructure, the types of vehicles to use and the dimensions of the fleets, etc.

3 Static SND

Let $\mathscr{G}^{\mathrm{PH}} = (\mathscr{N}^{\mathrm{PH}}, \mathscr{A}^{\mathrm{PH}})$ represent the physical infrastructure network, where $\mathscr{N}^{\mathrm{PH}}$ stands for the set of facilities, hubs and regional terminals, connected by the the physical or conceptual links of set $\mathscr{A}^{\mathrm{PH}}$. The goal of such models is to select from a set of potential services $\Sigma = \{\sigma\}$ either the service network only or the services and their frequencies to satisfy the demand for transportation of a set $\mathscr{K}$ of origin-destination (OD) commodities, each $k \in \mathscr{K}$ requiring to move a quantity of freight d^k between its origin $O(k)$ to its destination $D(k)$.

A static SND model is built on a network $\mathscr{G} = (\mathscr{N}, \mathscr{A})$ that is defined in the static case on the physical nodes of the system, i.e., $\mathscr{N} = \mathscr{N}^{\mathrm{PH}}$. With respect to the arc set $\mathscr{A}$, its composition depends on whether the potential services are *single-leg*, with no intermediary stop between the origin and destination terminals, or *multi-leg*. In the former case, $\mathscr{A} = \Sigma$. In the latter, $\mathscr{A} = \mathscr{L} = \bigcup_{\sigma \in \Sigma} \mathscr{L}(\sigma)$, where each multi-leg service is defined by a sequence of n service legs, collected in set $\mathscr{L}(\sigma) = \{l_i(\sigma) \mid i = 1, \ldots, n\}$, each service leg being a path in the physical network connecting two consecutive terminals on the route of service σ. We let $\sigma_a \in \Sigma$ denote the service associated with arc $a \in \mathscr{A}$.

Associated with service σ is a capacity $u(\sigma)$, which can be leg specific $u(l_i(\sigma))$, $l_i(\sigma) \in \mathscr{L}(\sigma)$, representing the total volume of freight the service may load and haul, as well as the cost f_σ incurred when doing so. In terms of arcs $a \in \mathscr{A}$, the attribute u_a takes the value $u(l_i(\sigma))$ associated with the service leg, $l_i(\sigma)$, modeled by that arc. Note that capacity may be measured in volume, tonnage, length (particularly for railroads), and number of units (e.g., containers for intermodal navigation and rail), that more than one capacity measure may be active in any given problem setting, and that particular capacities for particular products can also be imposed. To simplify the presentation and if not otherwise indicated, however, we continue with a single capacity restriction in this chapter.

Note that with both single-leg and multi-leg services, $\mathscr{A}$ may contain multiple arcs that have the same origin and destination, but differ in one of these attributes. In a single-leg setting, the carrier may choose from a market of third party carriers for the execution of the same service, with each carrier offering a different cost and capacity. In a multi-leg setting, two services may involve different sequences of legs, but those sequences may overlap.

In applications wherein a service can be executed multiple times, the SND models the frequency with which a service is executed with the non-negative integer variable $y_\sigma \in \mathbb{Z}_+, \sigma \in \Sigma$. When the decision is whether a service should be executed, binary variables $y_\sigma \in \{0, 1\}, \ \sigma \in \Sigma$ are used. We note that adapting the SND to applications wherein vehicle capacity is measured along multiple dimensions (e.g., weight and volume) is straightforward.

The arc $a = (i, j) \in \mathscr{A}$ also models the opportunity to transport a commodity on the transportation leg (i, j), which incurs a per-unit cost c_a. In some applications, this cost can depend on the commodity being transported, and thus the cost parameter is also indexed by the commodity, k, yielding c_a^k. We can consider different types of *flow* variables to model the routing of commodities on such arcs. The first type of variable is of the form $x_a^k \geq 0,\ a \in \mathscr{A}, k \in \mathscr{K}$, and prescribes the amount of commodity k that travels on arc $a \in \mathscr{A}$. The second is named similarly, but is instead defined over the range [0, 1], and models the percentage of commodity $k's$ demand that flows on arc (i, j). Modifying the SND to accommodate one type of flow variable instead of another is an exercise in ensuring the model correctly calculates the total flow on each arc. Both sets of flow variables allow a commodity to be split, and then routed along multiple paths from its origin to its destination.

In settings wherein this is inappropriate or undesirable (e.g., breaking down a sealed pallet is not allowed), the model must restrict a commodity to travel on a single path from its origin to its destination. This can be done by restricting the x_a^k variables to be binary, wherein they model whether commodity k travels on arc a. In this chapter, we focus on the first form of flow variable, which represents the amount of a commodity that flows on a leg.

As noted in the description of the problem setting, shipments travel on itineraries, which in the context of our network $\mathscr{G} = (\mathscr{N}, \mathscr{A})$ can be represented by paths. As a result, flow variables can be defined in terms of paths. The same options (continuous, fractional, binary) for the domains of path-based flow variables exist as for arc-based flow variables with each option having an analogous modeling implication. The formulation we present next can be modified to prescribe decisions in terms of paths, similar to what was presented in Chap. 2.

Formally, the SND seeks to

$$\text{Minimize} \quad \sum_{\sigma \in \Sigma} f_\sigma y_\sigma + \sum_{k \in \mathscr{K}} \sum_{a \in \mathscr{A}} c_a^k x_a^k \tag{12.1}$$

Subject to

$$\sum_{a \in \mathscr{A}_i^+} x_a^k - \sum_{a \in \mathscr{A}_i^-} x_a^k = \begin{cases} d^k, & \text{if } i = O(k), \\ -d^k, & \text{if } i = D(k), \\ 0, & \text{otherwise,} \end{cases} \qquad \forall i \in \mathscr{N}, \forall k \in \mathscr{K}, \tag{12.2}$$

$$\sum_{k \in \mathscr{K}} x_a^k \leq u_a y_{\sigma_a}, \qquad \forall a \in \mathscr{A}, \tag{12.3}$$

$$y_\sigma \in \mathbb{Z}_+, \qquad \forall \sigma \in \Sigma, \tag{12.4}$$

$$x_a^k \geq 0, \qquad \forall a \in \mathscr{A}, \forall k \in \mathscr{K}. \tag{12.5}$$

where for each $i \in \mathscr{N}$ we define the sets $\mathscr{A}_i^+ = \{(i', j) \in \mathscr{A} : i' = i\}$ and $\mathscr{A}_i^- = \{(j, i') \in \mathscr{A} : i' = i\}$.

The objective of the SND is to minimize the sum of the fixed costs associated with selecting and executing transportation services (the first term in (12.1)) and the variable costs associated with transporting commodities on legs associated with

those services (the second term in (12.1)). Constraints (12.2) are often referred to as *flow-balance* constraints and ensure that all of a commodity's demand departs from its origin (the first case), arrives at its destination (the second case), and departs any other locations at which it arrives (the third case). The expression on the left-hand side of the *linking* constraints (12.3) computes the total amount of demand that travels on arc $a \in \mathscr{A}$, whereas the expression on the right-hand side computes the total amount of capacity on that arc that is provided by the selected services. Thus, the constraint ensures that sufficient capacity is paid for. Constraints (12.4) define the domain of variables that indicate how often services are executed, while constraints (12.5) define the commodity routing decision variables, as well as their domains. .

Variants of the SND use commodity flow variables that represent the percentage (not the portion) of a commodity's demand that flows on an arc. This necessitates changing constraints (12.5) to require that $x_a^k \in [0, 1]$, replacing the expression on the left-hand-side of constraints (12.3) with $\sum_{k \in \mathscr{K}} d^k x_a^k$, dividing the right-hand-sides of constraints (12.2) by d^k, and multiplying the second expression in the objective (12.1) by d^k. Alternately, modeling a problem wherein the demand of each commodity flows along a single path necessitates changing Constraints (12.5) to instead require that $x_a^k \in \{0, 1\}$, in addition to the other changes noted above. Finally, for variants of the SND wherein the decision is whether a service should be executed, and not how many times it should be executed, constraints (12.4) should be changed to $y_\sigma \in \{0, 1\}$.

4 Time-Dependent SND

In *time-dependent* problem settings, demand $k \in \mathscr{K}$ is further characterized by an *availability time* $o(k)$ at origin $O(k)$ and a due date $d(k)$ at destination $D(k)$. Services are also characterized not only by their origin $O(\sigma)$, destination $D(\sigma)$, and set of legs $\mathscr{L}(\sigma) = \{l_i(\sigma) \mid i = 1, \ldots, n\}$, but also by a schedule indicating the departure and arrival times, $o(l_i(\sigma))$ and $d(l_i(\sigma))$, at the origin and destination terminals, respectively, of each of its legs $l_i(\sigma) \in \mathscr{L}(\sigma)$. Services are further characterized by a total duration $\tau(\sigma)$, that includes the time spent in terminals and the moving time associated to each leg $\tau(l_i(\sigma))$.

To capture these time-related characteristics of demand and service, SND models are generally defined on a time-space network $\mathscr{G} = (\mathscr{N}, \mathscr{A})$, that is typically built by extending the network $(\mathscr{N}^{\text{PH}}, \mathscr{A}^{\text{PH}})$ along the dimension of time for the fixed duration of the *schedule length*. The selected service network specifies in this case the movements through space and time of the vehicles and convoys of the various modes considered, while itineraries perform the same role for the transportation of time-dependent demand. When formulated on such a network, the SND is often referred to as a *Scheduled Service Network Design* (*SSND*) model.

The time-space network is built by partitioning the schedule length into non-overlapping periods of time, wherein all activities at terminals during a period will

be modeled as occurring at the same time. Often, these periods are of the same length. As an example, a planning horizon consisting of seven 24-h days may be partitioned into 14 half day (of 12 h each when the terminal operate continuously) or 168 one hour time periods. Then, demand that arrives at a terminal during a period, e.g., between 04:01 and 05:00, is modeled as being there by the end of the preiod, e.g., 05:00, and hence can be consolidated and loaded on services that leave the terminal in the same or following periods.

The time periods represented at one terminal may differ, however, from those represented at another terminal. This is often due to different terminals serving different purposes within a network. For example, some terminals may primarily serve as the interface between customers and the transportation network. For these terminals, it may be sufficient to only represent the time periods during which shipments become available or are due. Other terminals may primarily serve as consolidation centers. At these terminals vehicles may arrive and depart throughout the day, and thus more time periods may need to be modeled. Similarly, the length of a time period modeled at a node may depend on both the physical terminal and the start of the time period itself.

More formally, for each terminal $i \in \mathcal{N}^{\text{PH}}$, the network is based on a set, $\mathcal{T}_i = \{t_1^i, t_2^i, \ldots, t_{m_i}^i\}$ of periods of time during which activities may occur at that terminal. We let $\mathcal{T} = \cup_{i \in \mathcal{N}^{\text{PH}}} \mathcal{T}_i$ denote the set of all such time periods. The node set $\mathcal{N}$ then consists of nodes of the form $(i, t_p^i), i \in \mathcal{N}^{\text{PH}}, t_p^i \in \mathcal{T}_i$, that represent a terminal during a period in time.

The arc set $\mathcal{A}$ consists of multiple types of arcs. The first represents the execution of the service legs. Specifically, for service $\sigma \in \Sigma$ and leg $(i, j) \in \mathcal{L}(\sigma)$, $\mathcal{A}$ will contain an arc of the form $a = ((i, t_p^i), (j, t_q^j))$, which represents the departure of that service leg from terminal i at time t_p^i to arrive at location j at time t_q^j. As with the SND, we denote the underlying service associated with arc a by σ_a. The time point t_p^i is usually chosen so that $t_p^i \leq o(l_i(\sigma))$. Namely the arc models that the leg departs no later than its scheduled departure time. Similarly, the time point t_q^j is usually chosen so that $t_q^j \geq d(l_i(\sigma))$. Namely, the arc models that the service arrives no earlier than its scheduled arrival time. As with the SND, associated with a service, its legs, and the corresponding arcs are capacity and cost attributes.

Recall that we focus on a time-expanded network that enables the SSND to prescribe a plan that is repeatable. This is done by modeling activities (e.g., transportation) that would end after the end of the scheduling period as instead ending after the beginning. This is done by making the relevant arcs of $\mathcal{A}$ *wrap-around*. To be more precise, consider when the departure of service leg (i, j) at time t_p^i would imply arriving at j at a time period that is later than the last time period, $t_{m_j}^j$, modeled for terminal j. To model that the schedule is assumed to be repeated, an arc $a = ((i, t_p^i), (j, t_q^j))$ is then created wherein $t_q^j < t_p^i$. For example, consider a schedule length that is a business week and time periods that correspond to days. If service leg (i, j) has a 2 day duration, then a Friday departure from i

could be modeled with an arc that arrives at the node that models terminal j on Tuesday.

The second type of arc, often referred to as a *holding arc*, is of the form $a = ((i, t_p^i), (i, t_{p+1}^i))$ and represents the opportunity to hold goods or a resource at terminal i from period t_p^i to period t_{p+1}^i. As with transportation arcs, wrap-around arcs of the form $a = ((i, t_{m_i}^i), (i, 1))$ are created to represent holding a shipment (or allowing a resource to idle) from one schedule period to the next. The attribute u_a of a holding arc can be used to model the capacity terminal i has for holding goods. Similarly, the variable cost attribute c_a may be used to model the cost associated with holding goods for a period at terminal i (which may also depend on the commodity, k). The fixed cost attribute, f_a, may be used to model the cost associated with a resource idling at a location from one period to the next. However, we only consider variable costs in the model presented below.

Like the SND, the SSND considers two sets of decision variables. The first type of decision variable, $y_\sigma \in \mathbb{Z}_+, \sigma \in \Sigma$, models the number of times the transportation service, σ, is executed, which in turn implies the number of times its scheduled legs, $\mathscr{L}(\sigma)$, are executed. Selection-type decision variables are not typically associated with holding arcs in $\mathscr{A}$. However, situations wherein storage capacity at a location for a fixed period of time is paid for in fixed lot sizes (e.g., a storage cage) could be modeled with similarly-defined y variables. Like the SND, the domain of these variables, either service or holding, can be binary.

The second, $x_a^k \geq 0, a \in \mathscr{A}, k \in \mathscr{K}$, represents the amount of commodity k's demand that travels on arc $a \in \mathscr{A}$. Note that these commodity flow variables are defined over both types of arcs, those that represent transportation services, and those that represent the commodity being held at a terminal. As in the SND, these x variables can also be restricted to take on binary values. Alternately, and again like the SND, these x variables can instead be used to model the fraction, of k's demand that travels on the arc.

Thus, the SSND seeks to

$$\text{Minimize} \sum_{\sigma \in \Sigma} f_\sigma y_\sigma + \sum_{k \in \mathscr{K}} \sum_{a \in \mathscr{A}} c_a^k x_a^k \tag{12.6}$$

Subject to

$$\sum_{a \in \mathscr{A}^+_{(i,t_p^i)}} x_a^k - \sum_{a \in \mathscr{A}^-_{(i,t_p^i)}} x_a^k = \begin{cases} d^k, \text{ if } i = O(k), t_p^i = o(k), \\ -d^k, \text{ if } i = D(k), t_p^i = d(k) \\ 0, \text{ otherwise,} \end{cases}$$

$$\forall (i, t_p^i) \in \mathscr{N}, \forall k \in \mathscr{K}, \tag{12.7}$$

$$\sum_{k \in \mathscr{K}} x_a^k \leq u_a y_{\sigma_a}, \forall a \in \mathscr{A}, \tag{12.8}$$

$$y_\sigma \in \mathbb{Z}_+, \quad \forall \sigma \in \Sigma, \tag{12.9}$$

$$x_a^k \geq 0, \quad \forall a \in \mathscr{A}, \forall k \in \mathscr{K}, \tag{12.10}$$

where for each $(i, t_p^i) \in \mathscr{N}$ we define the sets $\mathscr{A}^+_{(i,t_p^i)} = \left\{a = \left((i', t_p^{i'}), (j, t_q^j)\right) \in \mathscr{A} : i' = i, t_p^{i'} = t_p^i\right\}$ and $\mathscr{A}^-_{(i,t_p^i)} = \left\{a = \left((j, t_q^j), (i', t_q^{i'})\right) \in \mathscr{A} : i' = i, t_p^{i'} = t_p^i\right\}$.

Each constraint set in the SSND has a direct analog in the SND. Note that the right-hand-side values of the flow balance equations (12.7) depend on the available and due times for the commodity. Also, note that constraints (12.8) and (12.9) are only defined for arcs in $\mathscr{A}$ that correspond to transportation services.

We complete this introduction to SSND models recalling that the time-space networks are a modeling tool and cannot capture all the temporal aspects of the problem. A continuous-time representation of the schedule length and associated events and decisions may be contemplated. It its most general setting, the SND model appears very complicated, however, as the time attributes of each service (and, thus, each itinerary) becomes part of the decision variable, including when several occurrences of the service are looked for (as in the model of this section). Such a model would also require a significant amount of constraints governing arrivals, departures, and activity synchronization at terminals. Moreover, such an approach does not fit well the applications where schedules are not strict and one only search for a number of departures within a given time interval.

Discretization of time, as described in this section, is thus the preferred methodology in time-dependent settings. Then, the question is what discretization granularity should one use. On the one hand, a fine granularity, yielding short time periods, provides the means to a detailed representation of time and time-related activities. But, it makes for huge time-space networks with dire consequences on the problem-solving capabilities of the current exact and metaheuristic methods, even when mathematical techniques and the restrictions of the application are used to reduce the network. A coarser granularity alleviates partially this problem, but may result in a poorer representation of decisions and operations in time. In most cases reported in the literature, the granularity is decided based on the application at hand, the experience of the researcher, and the power of the solver and computer available. We present in Sect. 7 the *Dynamic Discretization Discovery*, a new and very promising algorithmic strategy introduced recently to address this issue by an iterative generation the time-space network.

5 Broadening the Scope of SND: Integrating Resource Management

A number of additional considerations may characterize the carrier transportation system and its planning, enriching and challenging service network design methodology. Several such issues are particularly relevant for specific transportation modes

(rail, trucking, navigation, public transport) or problem settings (City Logistics) and are discussed in the associated chapters. In this section, we focus on an important issue of general relevance, namely, the integration resource-management concerns into SND models for tactical planning.

Carriers need resources to execute services, including equipment and manpower. For example, even though automation is becoming more and more prevalent in terminals, human resources are still needed to load/unload freight into/from outbound/inbound vehicles or containers, as well as handle and classify freight, vehicles or containers. Transportation activities also require resources. All modes require some type of power unit (tractors or trucks in trucking, locomotive engines in rail, planes in air, and vessels in maritime), one or several carrying units (trailers for trucking, rail cars for rail), an operator (truck driver, railroad engineer also called engine or train driver outside North America, air pilot, and ship captain), and sometimes a whole crew (particularly in air and sea).

These resources are generally scarce. There are several reasons for this fact. First, carriers aim continuously to control and hopefully reduce their operating costs to improve their market share and profitability. Consequently, the number of the power units and vehicles a carrier maintains has been drastically reduced to fit the forecast level of activity. There are precious few units available in most air, rail, trucking, and navigation carrier systems in case "something happens". Most resources are also expensive to acquire (e.g., power units), or there are few available for renting or acquisition, the shortage of truck drivers in North America being a perfect illustration. Leasing the appropriate number of the "right" type of intermodal rail cars is also an increasingly serious issue, at least in North America.

A second phenomenon impacting the availability of resources where and when needed is the unbalance inherent in trade. Indeed, the very nature of why trade is initiated (the desire for something available somewhere else), makes the commercial exchanges among countries, regions, and cities unbalanced not only in monetary value, but also in the type and quantity of products exchanged. This results in an unbalanced resource distribution among the terminals of the system. Power units and vehicles at destination and unloaded, become empty and available for the next operation. Yet, very often, these vehicles are not of the appropriate type or not available at the appropriate moment to load the outgoing freight. There is therefore a shortage of the appropriate vehicles, while those on location are needed somewhere else. Moving power units and vehicles "empty" to re-balance the system is called *repositioning* (deadheading for crews, especially in the air industry), and may represent a significant cost item for the carrier.

Not surprisingly, carriers have always aimed to minimize such operations and costs. Traditionally, however, the literature identified this problem as operational and addressed it, through more or less sophisticate network flow models, over rather short planning horizons, given the tactical plan. A somewhat more integrative approach computed an origin-destination matrix of empty vehicles, based on the OD-demand matrices, and distributed over the network jointly with the regular demand. None of these approaches works directly on the design of the service network.

The first integrative approaches focused on the most expensive resources, planes, ships, and rail engines. It assumed one unit of resource for each service and required that the number of selected services entering a terminal equals the number exiting the terminal. The corresponding *design-balanced SND* formulations extends the previous models by adding the set of node-degree constraints

$$\sum_{(i,j)\in\mathscr{A}_i^+} y_{ij} - \sum_{(j,i)\in\mathscr{A}_i^-} y_{ji} = 0, \quad \forall\, i \in \mathscr{N}. \tag{12.11}$$

Notice that, adding design-balanced constraints to SND formulations greatly complicates the search for high-quality solutions as, for example, even finding an initial solution is no longer straightforward (the rounding of the linear relaxation no longer guarantees a feasible solution). Moreover, the size of the formulation is increased, as is the computational effort to address arc-based models. On the other hand, note that such constraints naturally imply that resources move on cycles. The cycles may be of different time lengths (controlling cycle duration requires appropriate constraints) and may start at different periods during the schedule length. They are all, however, anchored at the terminal to which the resource is assigned. *Cycle-based* formulations thus appear natural.

Let $\Theta = \{\theta\}$ stand for the set of feasible cycles the units of the resource considered may perform, f_θ the "fixed" cost of selecting and operating the resource cycle $\theta \in \Theta$, and δ_θ^σ the cycle-to-service assignment indicator, where $\delta_\theta^\sigma = 1$ if the resource performing cycle $\theta \in \Theta$ may support service $\sigma \in \Sigma$, and 0 otherwise. Define the binary decision variable $y_\theta = 1$, if cycle $\theta \in \Theta$ is selected, and 0 otherwise,. The SSND with single resource management then becomes (to simplify the presentation, we display the formulation for the single-leg service case):

$$\text{Minimize} \sum_{\sigma\in\Sigma} f_\sigma y_\sigma + \sum_{\theta\in\Theta} f_\theta y_\theta + \sum_{k\in\mathscr{K}} \sum_{a\in\mathscr{A}} c_a^k x_a^k \tag{12.12}$$

subject to constraints (12.7)–(12.9) enriched with

$$y_\sigma \le \sum_{\theta\in\Theta} \delta_\theta^\sigma y_\theta, \quad \forall \sigma \in \Sigma, \tag{12.13}$$

$$y_\theta \in \mathbb{Z}_+, \quad \forall \theta \in \Theta, \tag{12.14}$$

where the objective function aims to minimize the selection and operation costs of services and resources, plus the cost of moving the demand flows, while constraints (12.13) link the existence of services and the resources required to operate them.

A more general *Scheduled Service Network Design with Resource Acquisition and Management*, *SSND-RAM*, problem includes not only several types of resources, but also integrates tactical, service network design-related decisions, and strategic, resource-acquisition and allocation decisions. In the SSND-RAM model presented herein, resources are differentiated by relevant characteristics, e.g.,

capacity, traction power, speed, energy and emission, scheduling rules, etc. The model also considers the additional "resource" of executing a service by a third party rather than by a resource owned or leased. Calling on such a resource incurs costs that are greater than executing the service with an owned resource, but may be valuable when, for example, moving a resource into and out of a somewhat remote region is costly. Moreover, the carrier does not have to worry about how the utilization of the third-party resource outside the execution of the designated service. Tactical planning is thus selecting services with costs and capacities that can be influenced by the type of resource supporting them, including the outsourcing possibility. To simplify the presentation, services in the following model require one unit of resource only to operate. Resources, on the other hand, are assigned to specific terminals and must return to their home terminals at least once during the tactical planning horizon.

The model also addresses strategic decisions related to fleet acquisition and management, e.g., how many resources of each type should be acquired (or rented, depending on the resource type and supplier), to what terminal new resources should be assigned, and between which terminals currently existing resources should be reassigned. Costs associated with these strategic decisions include, e.g., the unit purchase or renting cost, the additional salary or signing bonus associated with hiring the required personnel to operate the resource, and the transportation costs associated with re-allocating a resource from a home terminal to another.

The problem and decisions may be represented schematically as in Fig. 12.1. An integrated SSND-RAM formulation captures those decisions through a two-layer time-space network, illustrated in Fig. 12.2 for the decisions related to a single resource type. The SSND layer, on the right of the figure, corresponds to the tactical-planning decisions on service choice and commodity transportation. It is similar to the SSND models of the previous sections. The resource acquisition and allocation decisions are modeled on the strategic RAM layer, on the left of the figure.

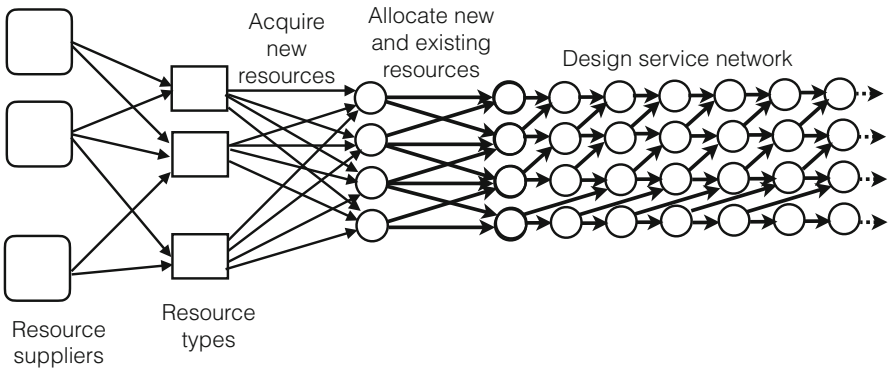

Fig. 12.1 Network model of SSND-RAM strategic and tactical decisions

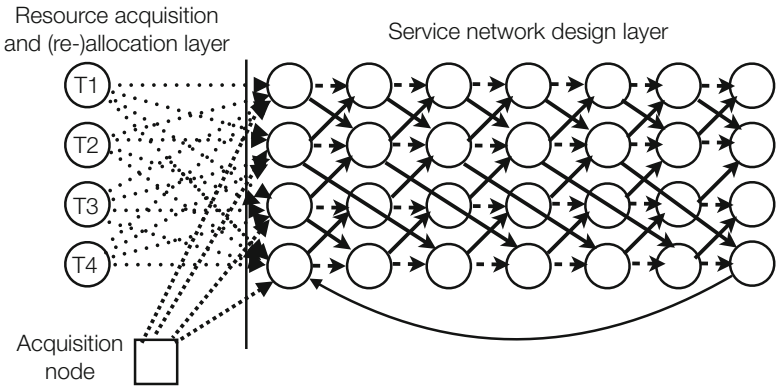

Fig. 12.2 SSND-RAM network model of strategic and tactical decisions

The SSND layer and notation is mostly similar to that of Sect. 4, adjusted for multiple resources. Let $\mathscr{R}$ stand for the set of available resources, f_i^r the fixed cost (salaries, maintenance, etc.) of operating a unit of resource of type $r \in \mathscr{R}$ that is assigned to terminal $i \in \mathscr{N}^{\text{PH}}$, and I_i^r the quantity of resources of type r initially assigned to terminal i. Let also Θ_i^r be the set of potential cycles a resource of type r assigned to terminal i can execute, $\Theta^r = \cup_{i \in \mathscr{N}} \Theta_i^r$ and $\Theta = \cup_{r \in \mathscr{R}} \Theta^r$. The cycle-to-service assignment indicator δ_θ^σ, links services and resources as previously. When service costs and capacities vary according to the assigned resource, the notation becomes f_σ^r and $u(\sigma, r), \sigma \in \Sigma,\ r \in \mathscr{R}$, respectively. Notice that a resource-independent fixed service selection cost, f_σ, may still be associated to a service modeling, e.g., the salaries of the officers of a liner ship. Finally, F_σ^r represents the fixed cost of operating service σ with a third party-owned resource of type r.

The RAM layer adds a few nodes, $\mathscr{N}'$, and arcs, $\mathscr{A}'$, to the time-space network, together with associated parameters and decision variables. There are two types of nodes in this layer, which are (1) symbolically defined at period 0, before the first period of the schedule length, and (2) connected to all first representations of the terminal nodes in the SSND layer. To simplify the presentation, and without loss of generality, we do not indicate the period 0, unless necessary.

A unique node, A, represents the acquisition of new resources. The corresponding arcs $(A,\ i, t_1^i),\ (i, t_1^i) \in \mathscr{N}$, represent the allocation of newly acquired resources to terminal i at the first period of activity at that terminal. Let h_i^r be the total cost of acquiring a new unit of resource $r \in \mathscr{R}$ and allocating it to terminal $i \in \mathscr{N}^{\text{PH}}$.

The second type of node is used to model the re-allocation of existing resources. A node i' is added at period 0 for each terminal $i \in \mathscr{N}^{\text{PH}}$, the arcs $(i',\ (j, t_1^j)), (j, t_1^j) \in \mathscr{N}$, connecting that node to each terminal representing the re-allocation of the resources initially at terminal i to terminal $j \in \mathscr{N}^{\text{PH}}$.

The corresponding cost of repositioning a unit of resource $r \in \mathscr{R}$ from terminal $i' \in \mathscr{N}^{\mathrm{PH}}$ to terminal $j \in \mathscr{N}^{\mathrm{PH}}$ is noted $h^r_{i'j}$ (with $h^r_{i'j} = 0$ for $i' = j$).

The cycle definition is extended over the RAM layer to capture the acquisition and re-allocation activities within the resource-routing decisions. Cycles are thus associated to nodes in $\mathscr{N}'$ and include the arcs of $\mathscr{A}'$, yielding the set $\Theta^r_{i'}$ of potential cycles a resource of type r can execute out of each respective terminal.

The decision variables of the SSND, $y_\sigma, \sigma \in \Sigma$, and $x^k_a \geq 0, a \in \mathscr{A}, k \in \mathscr{K}$, are also defined for the SSND-RAM. Define the additional decision variables

$y^r_\sigma = 1$ if service $\sigma \in \Sigma$ is operated with a third party-owned resource $r \in \mathscr{R}$ and 0, otherwise;

$z^r_\theta = 1$ if cycle $\theta \in \Theta^r,\ r \in \mathscr{R}$, is selected and 0, otherwise;

w^r_i: The number of new units of resource $r \in \mathscr{R}$ acquired and assigned to terminal $i \in \mathscr{N}^{\mathrm{PH}}$;

$w^r_{i'j}$ The number of units of resource $r \in \mathscr{R}$ positioned from terminal ($i' \in \mathscr{N}^{\mathrm{PH}}$ to terminal $j \in \mathscr{N}^{\mathrm{PH}}$.

The *Scheduled Service Network Design with Resource Acquisition and Management* formulation for the single-leg-service case may be then written as follows:

$$\begin{aligned}\text{Minimize} \quad & \sum_{r\in\mathscr{R}}\left(\sum_{i\in\mathscr{N}} h^r_i w^r_i + \sum_{i'\in\mathscr{N}}\sum_{j\in\mathscr{N}} h^r_{i'j} w^r_{i'j}\right) + \\ & + \sum_{\sigma\in\Sigma}\left(f_\sigma y_\sigma + \sum_{r\in\mathscr{R}} f^r_\sigma \sum_{\theta\in\Theta^r} \delta^\sigma_\theta z^r_\theta\right) + \sum_{\sigma\in\Sigma}\sum_{r\in\mathscr{R}} F^r_\sigma y^r_\sigma \\ & + \sum_{r\in\mathscr{R}}\sum_{i\in\mathscr{N}} f^r_i \sum_{\theta\in\Theta^r} z^r_\theta + \sum_{k\in\mathscr{K}}\sum_{a\in\mathscr{A}} c^k_a x^k_a\end{aligned} \tag{12.15}$$

Subject to

$$\sum_{i'\in\mathscr{N}'} w^r_{i'j} = I^r_i, \ \ \forall r \in \mathscr{R}, \ \forall (j, t^j_1) \in \mathscr{N}, \tag{12.16}$$

$$\sum_{\theta\in\Theta^r_{i'}} z^r_\theta \leq \sum_{(j,t^j_1)\in\mathscr{N}} h^r_{i'j}, \ \ \forall r \in \mathscr{R}, \ \forall i' \in \mathscr{N}', \tag{12.17}$$

$$\sum_{a\in\mathscr{A}^+_{(i,t^i_p)}} x^k_a - \sum_{a\in\mathscr{A}^-_{(i,t^i_p)}} x^k_a = d^k, \ \ \forall (i, t^i_p) \in \mathscr{N}, \ \forall k \in \mathscr{K}, \tag{12.18}$$

$$\sum_{k\in\mathscr{K}} x^k_a \leq \sum_{r\in\mathscr{R}} u(\sigma, r)\left(\sum_{\theta\in\Theta^r} \delta^\sigma_\theta z^r_\theta + y^r_\sigma\right), \ \ \forall a \in \mathscr{A}, \tag{12.19}$$

$$y_\sigma \leq \sum_{r \in \mathscr{R}} \sum_{\theta \in \Theta^r} \delta_\theta^\sigma z_\theta^r, \quad \forall \sigma \in \Sigma, \tag{12.20}$$

$$y_\sigma + y_\sigma^r \leq 1, \quad \forall \sigma \in \Sigma, \tag{12.21}$$

$$w_i^r, \ w_{i'j}^r \in \mathbb{Z}^+, \quad \forall r \in \mathscr{R}, \ i \in \mathscr{N}', \tag{12.22}$$

$$z_\theta^r \in \{0, 1\}, \quad \forall r \in \mathscr{R}, \ \forall \theta \in \Theta^r, \tag{12.23}$$

$$y_\sigma^r \in \{0, 1\}, \quad r \in \mathscr{R}, \ \forall \sigma \in \Sigma, \tag{12.24}$$

$$x_a^k \geq 0, \quad \forall a \in \mathscr{A}, \ \forall k \in \mathscr{K}. \tag{12.25}$$

The objective minimizes the total cost of the system. The first term models the cost of acquiring new and re-allocating existing resources. The second term computes the cost of selecting services and operating them with owned resources on particular cyclic routes. The third term models the costs incurred to secure third-party resources. The fourth term represents the costs associated with putting a resource into use, while the fifth and last term models shipment transportation costs.

Constraints (12.16) ensure that all resources of type r that are initially allocated to terminal i are either left at i or re-allocated. Constraints (12.17) link the strategic resource acquisition and allocation/re-allocation decisions that determine the number of resources available at each terminal with the tactical decision of how many resources from that terminal are to be used to execute services. Note the summation over $\mathscr{N}'$ in constraint s(12.17) enables the use of resources that are newly acquired.

Constraints (12.18) and (12.19) enforce classical network design relations. The former are commodity-specific flow conservation constraints. The latter link the existence of flow on owned or outsourced services to the corresponding service-selection decision. Constraints (12.20) indicate that at most one resources is used for each owned service, while constraints (12.21) specify that each service cannot be selected more than once, either supported by the carrier's resources or outsourced. Finally, constraints (12.22)–(12.25), define the domains of the variables in the formulation.

6 Managing Uncertainty

The SND and SSND are parameterized mathematical models of consolidation-based transportation systems. Using the methodology for planing and management purposes requires not only the model to accurately represent the system, but also the values of the model parameters to adequately predict the variations in the state of the system over the contemplated planning horizon. Of course, in reality, the validity of this assumption is rarely certain. Accounting explicitly for uncertainty in SND and SSND models aims to address this issue. An in-depth discussion of uncertainty and

network design may be found in Chap. 9. We briefly recall the fundamental concepts in this section, focusing on their application to service network design.

In general, researchers have classified uncertainty into one of three types based upon their likelihood and impact. The first type, *randomness*, refers to events whose likelihood can be described and is reasonably high, but whose impacts can usually be mitigated within normal operations. The classic example of such uncertainty in SND contexts is fluctuations in the shipment volume between a given origin and destination. The second type, *hazards*, refers to events whose likelihood can be described, but are quite rare. An example in SND contexts is vehicle failure. The third type, *deep uncertainty*, refers to events whose likelihood can not be described and is extremely impactful. An example in SND contexts is a maritime port closing down due to a threat of terrorist attack.

Much (if not all) of the research on SND problems has focused on the first type of uncertainty, randomness, and specifically uncertainty with respect to model parameter values. This uncertainty is modeled by extending one of these deterministic models to a two-stage stochastic program . Such an optimization model presumes that some decisions must be made and implemented at a time when information regarding instance parameter values is incomplete. Specifically, that some decisions must be made at a time when only statistical distributions are known for the values of some parameters. In the context of a two-stage stochastic program, these decisions are referred to as *first stage* decisions. Then, at some point after the first stage decisions are implemented, the realizations of the uncertain parameter values is revealed. At that point, the remaining decisions can be made, in light of both the realized parameter values and the first stage decisions. These remaining decisions are often referred to as *second stage*, or, *recourse* decisions. As the second stage decisions are functions of random variables, they are random variables as well. Thus, the objective of such a model is to minimize the sum of the costs associated with the first stage decisions and the expected costs associated with second stage decisions.

In the context of service network design, most stochastic models prescribe the selection of services in the first stage and the routing of commodities, given those services and the realized parameter values, in the second stage. It is important to note that with two-stage stochastic programs in general, as well as those for service network design, the presumption behind these models is that from a practical planning perspective only the first stage decisions must be determined. The second stage decisions are not expected to be implemented. They may be used as guidelines (e.g., the itineraries and terminals for the main demand flows) when repeatedly applying the plan during the planning horizon. They primarily serve, however,as a means of approximating the impact of the first stage decisions on the performance of the system over the planning horizon. Specifically, the second stage approximates the expected cost of transporting demand loads given a network design.

To that effect, much of the research involving stochastic service network design models includes in the second stage the option to *outsource* all, or a part of, the delivery of a commodity from its origin to its destination, wherein the cost of outsourcing is proportional to the amount of the commodity's demand that is

outsourced. Outsourcing may mean calling on an external service provider, or using an owned service which is not within the scope of the current problem and SND model. Thus, e.g., empty and loaded container-dedicated rail cars that cannot be accommodated on intermodal train services when the intermodal SND is being built, can be moved by general trains not in the scope of the planning problem. Then, by outsourcing a commodity, its delivery does not require the carrier designing a transportation plan to execute transportation services. Thus, in total, the stochastic SND formulation seeks to minimize the cost of executing services together with the expected cost of routing commodities and calling on external resources.

In addition, most research involving stochastic service network design models presumes that the joint probability distribution for uncertain parameter values can be approximated with a finite set of scenarios, wherein each scenario contains a realization of each uncertain parameter value and has a probability of occurring. With these scenarios, the expectation in the objective function can be expressed as a linear function, and the stochastic program can be formulated as a deterministic mixed integer program. In this section, we first focus on what types of stochastic programs have received the most attention for the SND. Namely, models that explicitly recognize uncertainty in shipment volumes due to randomness. We then discuss other potential sources and types of uncertainties that can be modeled.

6.1 Uncertainty in Shipment Volumes

The most commonly modeled source of uncertainty is demand. This is in part because it is the most prevalent in practice. Fundamentally, the SND and SSND presume that the size of a commodity is known and constant over the planning horizon during which the transportation plan prescribed by the model is implemented. In many logistics settings, a commodity models the orders of some customer (or the aggregation of multiple customers' orders). Thus, one source of uncertainty in commodity demand is due to variation in customer orders during that horizon. Another source of uncertainty has to do with the actual amount of vehicle capacity required by a customer order. The commodity demand value derived from a customer order is often just an estimate that is based on physical dimensions that are communicated by the customer to the transportation carrier. Thus, the actual amount of vehicle capacity required by an order may not be known with certainty until the order is picked up. We next present a stochastic programming variant of the SND above that is based on the premise that there is uncertainty in shipment volumes.

Uncertainty in shipment volumes is typically represented by treating the demand quantities, d^k, as random variables. A joint probability distribution for those random variables is presumed known, and is represented with a finite set of scenarios, $\mathscr{S}$. Each scenario $s \in \mathscr{S}$ represents a realization of the values, d^{ks}, of each of the random variables d^k. In addition, associated with scenario s is a probability, p_s, that it occurs.

The service network design under uncertainty (SND-U) problem is typically formulated on the same network, $\mathscr{G} = (\mathscr{N}, \mathscr{A})$, as the SND and considers the same set of services, Σ. As service selection is determined in the first stage, before demand information is completely known, the SND-U involves the same y variables, $y_\sigma, \sigma \in \Sigma$, as the SND. Like the SND, the domains of these y_σ variables are usually either binary or integer numbers.

The SND-U models that commodity routing decisions occur after demand information has been fully revealed and design decisions are made. As a result, these decisions may depend on the scenario observed, and are modeled by indexing commodity flow variables by scenario, x_a^{ks}. Like the service selection variables, y_σ, these variables have the same possible domains as in the SND. Thus, the SND-U seeks to

$$\text{Minimize} \sum_{\sigma \in \Sigma} f_\sigma y_\sigma + \sum_{s \in \mathscr{S}} p_s \sum_{k \in \mathscr{K}} \sum_{a \in \mathscr{A}} c_a^k x_a^{ks} \tag{12.26}$$

Subject to

$$\sum_{a \in \mathscr{A}_i^+} x_a^{ks} - \sum_{a \in \mathscr{A}_i^-} x_a^{ks} = \begin{cases} d^{ks}, \text{ if } i = O(k), \\ -d^{ks}, \text{ if } i = D(k), \\ 0, \text{ otherwise,} \end{cases} \quad \forall i \in \mathscr{N}, \forall k \in \mathscr{K}, \forall s \in \mathscr{S}, \tag{12.27}$$

$$\sum_{k \in \mathscr{K}} x_a^{ks} \le u_a y_{\sigma_a}, \quad \forall a \in \mathscr{A}, s \in \mathscr{S}, \tag{12.28}$$

$$y_\sigma \in \mathbb{Z}_+, \quad \forall \sigma \in \Sigma. \tag{12.29}$$

$$x_a^{ks} \ge 0, \quad \forall a \in \mathscr{A}, \forall k \in \mathscr{K}, \forall s \in \mathscr{S}. \tag{12.30}$$

The objective of the SND-U seeks to minimize the cost associated with executing services along with the expected cost of routing commodities given the services selected. As the SND-U models commodity routing decisions that vary by scenario, constraints (12.27)–(12.30) enforce the same logical conditions as constraints (12.2)–(12.5) of the SND, albeit with a set of constraints for each scenario and demands that depend on the scenario. However, note that the right-hand side of constraints (12.28) represents that the same design is used to route commodities in each scenario.

As noted above, the SND-U is sometimes formulated under the assumption that the transportation of a commodity from its origin to its destination may be outsourced, and at a cost that is proportional to the amount outsourced. This is often modeled by adding the arc $(O(k), D(k))$ to $\mathscr{A}$ for each $k \in \mathscr{K}$. For these arcs, the cost c_a^k represents the outsourcing costs. As the transportation options modeled by these arcs do not involve a service executed by the carrier, constraints (12.28) are not formulated for such arcs.

6.2 Other Uncertainties in SND

We next discuss models that recognize other uncertainties that can be present in service network design. However, we note that many of these models have received little academic attention and some none at all. On the supply side, there can be uncertainty regarding the capacity to route commodities provided by a service that first stage decisions indicate should be executed. In practice, there are two sources for this uncertainty. In the first, unforeseen events (i.e., hazards) such as equipment failures can prevent the execution of a service that the first stage decisions prescribe. Thus, the capacity of the service effectively becomes zero. The second is similar, in that the capacity is different from what was anticipated in the first stage of the model, but less dramatic. Such uncertainty can occur, e.g., when a service is executed by a third party transportation carrier and the capacity provided by that service is shared with other carriers. As a result, when other carriers use more capacity than anticipated, the capacity available to the organization solving the SND is reduced. Such a drop in capacity may occur even with owned resources, such as partial equipment failure, e.g., cars on trains or compartments on liner ships.

Both sources can be modeled by treating the quantities u_a as random variables. However, the distributions used to model the two different sources are likely different. Regardless, given a set of scenarios to approximate the joint distribution of arc capacities (and potentially other random variables such as commodity demands), a SND-U similar to the one presented above can be formulated wherein u_a^s represents the capacity of arc a in scenario s. Then, the right-hand-side of constraints (12.28) is replaced with the term $u_a^s y_{\sigma_a}$.

There can also be uncertainty related to the costs incurred, either when routing a commodity or executing a service. Regarding routing a commodity, the SND may model the opportunity to use services that are executed by a third-party carrier that charges on a per-unit-of-demand basis (e.g., per pallet). In such a situation, there may be variability in the variable costs due to market forces. Modeling such variability can be done by treating the quantities c_a as random variables, which can be easily done as the variables associated with those cost coefficients are already in the second stage. By again presuming a set of scenarios representing the joint distribution of random variables, and c_a^s representing the variable cost on arc a in scenario s, a SND-U similar to the one above can be formulated, albeit with a slightly modified second term in the objective.

Regarding executing a service, as the associated cost is generally a function of transportation, variability from what was estimated can be driven by variability in the resources needed for transportation (e.g., fuel). Alternately, when a service is executed by a third party that provides transportation services to multiple customers, but charges on a per-service basis, variability may be driven by market forces. Such variability can be modeled by treating the costs f_{ij} as random variables. As these coefficients are associated with first stage decisions, calculating the total, expected fixed cost is not as straightforward as treating the variable costs c_a as random variables. No known research considers models that recognize this source of uncertainty.

Finally, and specific to the SSND, there may be uncertainty in the timings of activities. For example, there may be uncertainty related to the time, e_k, at which a commodity is available or to the time, l_k, at which it is due for delivery. In constraints (12.7), the SSND presumes these times are known with certainty, when in fact both may vary from what is expected. Issues with a manufacturing process may mean that goods to be transported are not always available by the time e_k. Alternately, a customer may sometimes need to rush an order, requiring the goods to be delivered before the time l_k. The SSND can be easily extended to a stochastic program that models both these uncertainties. Specifically, the right-hand-side values of constraints (12.7) can be modeled as random variables, d_{it}, with a distribution that is approximated by scenario. Then, in each scenario s, there must be a single t such that the $d^s_{O(k)t} = d^k$, a single t' such that $d^s_{D(k)t'} = -d^k$, and for all other i, t'', $d^s_{it''} = 0$.

Lastly, there may be uncertainty in the departure and arrival times of services. Note that time-dependent service travel times may be accommodated in the construction of the time-expanded network, $\mathscr{G}_{\mathscr{T}}$. Variability in service departure and arrival times may occur due to traffic congestion, weather conditions, or unforeseen events in terminal operations. Recently, there has been research that seeks to design transportation networks that meet a "service quality" target, wherein service quality refers to the probability of a commodity reaching its destination on time.

7 Bibliographical Notes

There is a broad and extensive literature on the Service Network Design problem. General surveys of the literature can be found in Crainic and Laporte (1997); Crainic (2000, 2003) and Wieberneit (2008). There are also surveys that focus on the use of SND models in specific contexts. Examples include intermodal freight transportation (Crainic and Kim 2007; Bektaş and Crainic 2008), City Logistics (Bektaş et al. 2017), and several chapters of this book. In the remainder of this section we review some of the most significant contributions to the literature. As this chapter was focused primarily on modeling up to now, we pay particular attention to solution approaches. Many ideas proposed for more general network design problems have been successfully adapted or applied to service network design problems. However, we focus our discussion on ideas that were primarily proposed in the context of service network design.

Some of the earliest work, both in terms of modeling and algorithmic development, can be found in Crainic et al. (1984); Crainic and Rousseau (1986) and Crainic and Roy (1988). The static path-based SND formulation minimizes a nonlinear generalized objective function combining operating and time-related costs for services and shipments, as well as penalty costs for non compliance with service targets (e.g., market-specific delivery times) or the capacity limitations of terminals and services. The latter are cast as quadratic functions of the excess flow or duration. Moreover, the duration of terminal activities is modeled through convex

approximations of average (and standard deviation) delays derived from queuing models accounting for the capacity and operation characteristics of the terminal. A similar approach is used for inter-terminal travel times when vehicles are captive of the infrastructure (e.g., rail and barges) or congestion phenomena are considered.

Many of the early solution methods proposed for SND problems are iterative local improvement heuristics. Examples of such methods can be found in the previous papers as well as in Powell (1986); Farvolden and Powell (1994). These methods search for an improving solution at an iteration by first adding or dropping services from the current network, and then routing the flows on the resulting network. Adding/dropping services from a network in the context of searching for an improving solution continues to be an effective algorithm strategy (Pedersen et al. 2009).

Kim et al. (1999) study a service network design problem in the context of express package delivery via a transportation network that connects ground and air movements. They leverage the structure of this transportation network and the nature of potential vehicle routes to derive a reduced time-space network on which they formulate an integer programming based on service and package flow route variables. Due in part to the scale of the delivery operation they solve this reduced formulation with column and row generation techniques. This exact, integer programming-based method is also used as the basis of a computationally effective heuristic. Armacost et al. (2002) study a similar problem and also approach the problem with integer programming methodology. However, motivated in part by the notoriously weak linear programming relaxations of service network design problems, they propose a formulation that does not model package flows directly. Instead, they propose a formulation based solely on design variables that represent aircraft routes, and show that with the right constraint set such a formulation can ensure sufficient capacity to transport all package demands, even though those demands are not explicitly modeled. They further strengthen the formulation by defining a specific type of design variable called a *composite variable*, which encodes the selection of multiple aircraft routes.

Jarrah et al. (2009) studies the service network design problem in the context of the less-than-truckload freight transportation industry. They leverage the single-path per shipment policy desired by carriers to propose a new formulation to the problem. Specifically, because the paths for shipments destined for the same terminal must induce a directed in-tree rooted at that terminal, the problem can be formulated with variables that represent flows on such trees. The proposed solution approach generates destination in-trees in a column generation-fashion in the context of a heuristic scheme. This in-tree structure was also exploited in Erera et al. (2013) in the context of a matheuristic scheme which at each iteration chooses a destination terminal and then solved an integer program to route freight destined for that terminal, holding fixed the routes for freight destined for other terminals.

Crainic et al. (1984); Crainic and Rousseau (1986); Crainic and Roy (1988); Powell (1986); Armacost et al. (2002); Jarrah et al. (2009), and Erera et al. (2013), to name but a few, consider models wherein the need to reposition empty vehicles is explicitly modeled. This has also been more generally referred to as *asset*

management or *design-balance* (Andersen et al. 2009b,a; Pedersen et al. 2009; Chouman and Crainic 2015; Vu et al. 2013). Generally speaking, these models seek to ensure that the number of services that arrive at a node in a network equal the number of services that depart. This requirement introduces a challenge to linear programming-based heuristics for the SND as rounding up a fractional solution to the linear programming relaxation is no longer guaranteed to yield a feasible solution to the original problem. However, it also induces a structure to solutions. Specifically, that a design can be decomposed into cycles. Andersen et al. (2011) exploit this structure in a branch-and-price-based scheme for the problem wherein vehicles flow on cycles and commodities flow on paths, with both cycles and paths generated dynamically via column generation.

Many of the earliest service network design models do not consider assets at all. They seek to ensure there is sufficient capacity dispatched to transport shipments. Models that incorporate asset management constraints recognize that resources are needed to transport shipments and thus may have to move empty to be positioned for future moves. However, these models do not recognize that there may be a fixed fleet of resources, or that resources may need to periodically return to a specific "home" terminal. These types of issues are studied in models that incorporate *resource management* considerations (Crainic et al. 2014, 2018; Hewitt et al. 2019). The solution methods proposed in these papers combine a column generation scheme for generating resource cycles with another scheme for choosing cycles and routing shipments given the capacity created by those cycles. Crainic et al. (2018); Hewitt et al. (2019) also consider resource acquisition, allocation, and re-allocation decisions. Unlike the papers discussed so far, Hewitt et al. (2019) considers a model that explicitly recognizes uncertainty. Specifically, shipment volumes are presumed to be uncertain and resource and service network design decisions are made before complete demand information is known.

SND and SSND problems that recognize uncertainty have been studied. Lium et al. (2007, 2009) analyze the value of recognizing uncertainty in such models as well as how doing so leads to different structures in solutions. Both papers consider models that recognize uncertainty in shipment volumes, with the first focusing specifically on situations where there are correlations between those volumes. Turning to algorithms for such problems, Hoff et al. (2010) proposes a metaheuristic for a SND problem wherein there is uncertainty in shipment volumes. Wang et al. (2019) consider a different algorithmic approach to stochastic service network design problems and instead focus on the potential of creating a solution to a stochastic SND from a solution to its deterministic counterpart.

Stochastic programming-based approaches to SND that recognize uncertainty typically prescribe design decisions in the first stage and shipment flow decisions in the second. Thus, most of these models presume that the design remains unchanged after demands are revealed. Some (e.g., Crainic et al. 2016; Hewitt et al. 2019) model the opportunity to slightly adjust, in departure times, for example, or augment the chosen design after demands are revealed. Bai et al. (2014) model the opportunity to instead change the design (albeit at a penalty). Lastly, we note that while the vast majority of stochastic service network design models presume

that statistical distributions exist for demands, robust optimization-based approaches have recently been proposed (Wang and Qi 2019, 2020).

Other sources of uncertainty have received attention as well. Specifically, Lanza et al. (2018, 2021) study a model wherein there is uncertainty in travel times. Such uncertainty introduces an additional component to the objective of the model that measures quality of service. Namely, the objective incorporates penalty factors based upon the total expected lateness of services and shipments. Demir et al. (2016) also study a model that recognizes uncertainty in travel times. Instead of general SND, they focus on intermodal transportation wherein fluctuations in travel times can interfere with the need to synchronize different transportation modes. One source of variability in travel times is the potential for vehicles and shipments to be delayed at a terminal. Estimating the lengths of these delays has received some attention (Crainic and Gendreau 1986).

Most SNDs consider a single level of consolidation. Namely, the consolidation of shipments into a container that is transported by a vehicle. However, some modes of transportation (e.g., rail, sea liners, and intermodal barges) necessitate multiple levels of consolidation as a vehicle may transport many containers, vehicles may be grouped into so-called *blocks* or *convoys* (rail and barge trains), or both. Such multi-layer models are considered in Kazemzadeh et al. (2019) and Zhu et al. (2014). We defer a deeper discussion of this topic to Chap. 12 of this book that focuses on rail network design. However, we note that a similar phenomenon has been considered in papers on SND that model motor-carrier platooning for long-haul movements (Albinski et al. 2020), and autonomous vehicles that can only travel autonomously in certain geographic regions (Scherr et al. 2018, 2019). The autonomous vehicles instead have to be pulled (called *platooning*) by a manned vehicle to such regions wherein they can then operate autonomously.

One of the computational challenges associated with solving instances of SSND models inspired by real-world operations is that the time-space networks on which these instances are based end up being very large. As a result, the numbers of variables that model shipments and vehicles moving through that network in those instances are very large as well, leaving mathematical programs that are too large to be solved in reasonable run-times. The network reduction techniques proposed in Kim et al. (1999) leveraged the specifics of that logistics context in an attempt to mitigate this issue. However, the size of these networks is due in part to the enumerative nature in which they are created and the process by which they are used.

First, the node set of such a network is created by enumerating each physical location at every time point when operations can occur at that location. Second, one portion of the arc set is created by enumerating each physical transportation move (the service) at every time point when it can depart. The other portion of the arc set consists of arcs that connect two nodes that represent the same location at different time points. Then, the instance is formulated on this network and solved. In such a static approach, much of the network that is created may not be needed by high-quality solutions.

Motivated by this observation, Boland et al. (2017) propose a different algorithmic strategy, named *Dynamic Discretization Discovery* (DDD), for using time-space networks in the context of SSND models. Specifically, they propose an iterative approach that begins with a time-space network wherein each location is represented at a small subset of the time points wherein operations may occur. Similarly, each physical transportation move is represented at a small subset of the time points at which it may depart. Boland et al. (2017) refer to such a network as a *partially time-expanded network* and formulate it in such a way that a SSND formulated on such a network is a relaxation of the SSND formulated on the time-space network derived from complete enumeration. To ensure that it is a relaxation given that not all locations are represented at all potential times, the network may need to contain arcs that underestimate actual travel times.

Thus, at an iteration of DDD, a SSND is solved on the current partially time-expanded network and the solution is examined to see if it can be converted to an optimal solution to the SSND formulated on a time-space network derived from complete enumeration. If it can be converted, the algorithm stops. If it can not, the current partially time-expanded network is refined and the algorithm continues. While Boland et al. (2017) present DDD for general SSNDs, it has also been adapted to other SSND-related problems. Medina et al. (2019) and He et al. (2019) propose adaptations of the algorithm to SSND problems that also determine local delivery routes. Hewitt (2019) proposes speed-up techniques for DDD when used to solve instances of SSNDs inspired by the less-than-truckload freight transportation industry. Marshall et al. (2021) propose a variant of DDD based on a differently-formed partially time-expanded network.

Another computational challenge that is often encountered when solving either SNDs or SSNDs inspired by real-world operations is that the number of shipments to be transported can be very large. This in turn can yield a large number of shipment flow variables and a mathematical program that is too large to be solved in reasonable run-times. One way to mitigate this issue is by defining shipment flows on paths instead of arcs (e.g., Crainic and Rousseau 1986; Crainic et al. 2009; Andersen et al. 2011; Hewitt et al. 2019). The downside to this approach is that it typically necessitates a scheme for dynamically generating paths as there are usually far too many to enumerate a priori. Another approach is a Benders decomposition-based method, wherein design decisions are made by a master problem based on estimates of the resulting shipment routing costs. These estimates are reflected in a constraint set present in the master problem that is iteratively added to as new designs are discovered. The downside of this type of approach is that these estimates are typically very poor in the early stages of the algorithm. As a result, Belieres et al. (2020) propose strengthening the master problem with the need to route a single, aggregated, *super-product*. Fontaine et al. (2016, 2021) take advantage of the problem structure to propose a different Benders decomposition, which includes tailored partial-decomposition technique for deterministic mixed-integer linear-programming formulations and specialized valid inequalities. In this approach, the master problem selects the services to generate a lower bound, while the slave problem solves a multiple knapsack problem with precedence constraints.

8 Conclusions and Perspectives

The chapter presented an overview and synthesis of the main classes of Service Network Design models aimed at supporting decision-making in planning the activities and managing the resources of consolidation-based freight carriers and systems. Applications of SND models and associated solution methods to particular transportation modes and system organizations are described in many chapters of this book, notably Chaps. 12–17. The chapter focused rather on issues and model structures of general interest and relevance. We continue this approach in identifying a number of challenging research perspectives of importance for both service network design and its applications and the broader network design field.

Extending the scope of SND, and the related modeling challenges, makes up a first research field. The aim is (1) to enhance the representation power and relevance of our models and solution methods, and (2) to extend the applicability of SND methodology beyond planning, to the short-term adjustment of plans to today's or this week's environment in terms of demand, state of the physical network, weather, and so on and so forth. Identifying the relevant issues and modeling them adequately requires a significant research effort, similar to the ones evoked in the following.

We have discussed the integration of resource-management concerns in Sect. 5, but many challenges remain. Different services and resources have different requirements and limitation. Thus, for example, North American trains are generally long and heavy and require traction power which can be provided only by combining two or three engines. Several such combinations are possible and each engine type has particular operational, maintenance, and fleet-size characteristics. Human resources also come with particular qualifications and work rules, including limitations on working hours and the types of vehicle individuals are authorized to operate. And, irrespective of mode and setting, transportation services require resources of several types, governed by particular compatibility rules to operate. Integrating the management of several heterogeneous interlinked resources into SND and SSND formulations challenges modeling and solution-method development alike.

Similar challenges characterize a better, more refined representation of terminal activities within tactical and strategic-level formulations. Most contributions so far model terminals through "simple" single node or arc representations and associated unit cost measures. Global capacity and unit time-related measures are appended to the node or arc representation in some cases. Yet, most terminals are complex infrastructures performing several operations on vehicles, power units, and loads in various parallel or sequential activity and waiting/queuing combinations. Average node or arc measures per unit of flow or service do not adequately represent this complexity. Obviously, one cannot integrate into a network-wide tactical or strategic model the full detailed representation of an operating terminal through, e.g., a network of queues. A few authors explored more detailed terminal representations replacing the node or arc with a small network capturing the main activities and waiting times of the terminal (e.g., Andersen et al. 2009b; Pedersen and Crainic 2007). These models provided more accurate estimations of the performance of the terminal, in terms of time and cost.

Explicitly addressing time and delay-related issues enlarges and refines the scope of SND models while raising significant modeling and algorithmic challenges. Consider, for example, the congestion one frequently observes in terminals and the resulting delays to vehicles and freight, which have to, first, enter the terminal and, then, go through the sequence of operations. These congestion conditions and delays are the result of high volumes of vehicles and freight "competing" for the terminal resources, i.e., its "capacity", within more or less the same time interval. As briefly mentioned above, the first challenge is to adequately represent these delays in terms both of model representativity and algorithmic efficiency. Working with a more refined terminal representation is part of the response to this challenge. Then, there is the issue of approximating the delays with linear or non-linear, ideally convex, functions. The former makes for an easier algorithm development, while the latter offers a more refined and adequate representation.

Adopting a non-linear formulation provides opportunities for modeling a broader range of issues and criteria compared to linear formulations with fixed capacities. Two cases to illustrate the point. First, representing infrastructure or service capacity in tactical-planning models through traditional constraints ignores the flexibility of translating plans into daily operations and of adjusting the former to the reality of the second. Thus, depending on the mode and carrier, extra freight one cannot load on a service either waits for the next departure or forces the dispatch of an extra vehicle. Both actions come at a cost best represented through a non-linear penalty, which may increase with the volume waiting, the length of the delay, etc. Consider, second, the quality targets carriers set and often publicize, e.g., A to B in X days. One may refine the selection of services and freight itineraries by representing potential deviations from target in the SND objective function. Non-linear penalties accounting for deviations from schedule for services and from due dates for commodity paths model such situations and guide the SND solution method. Standard deviations of activity, waiting, and travel times may also be included in computing the penalties, as well as the generally unpublicized percentage of error in attaining the targets the carrier allows for itself. Research is needed into SND formulations with non-linear objective functions and the associated solution methods (e.g., Bektaş et al. 2010).

Addressing uncertainty in SND models and methods for transportation and logistics planning constitutes a broad and important research area, challenging modeling and algorithmic development alike. As discussed in Sects. 6 and 7, SND models and solution models have already been proposed to address a number of uncertainty-related issues. One may state, however, that research in this area is still in its infancy. Research is still required in adequately representing demand uncertainty in the various problem settings evoked in this chapter and the other chapters of the book. Almost totally overlooked, although of great operational and economic importance, is the uncertainty in travel and terminal-activity times. The solution often adopted in practice of adding large buffers to the planned delivery times is not only scientifically unsatisfactory, but also less and less economically viable and impracticable in many cases (City Logistics to name but one example). Moreover, one should not overlook that both demand and time uncertainty (and

heavy correlations) characterize operations and their simultaneous presence, and interactions, should be reflected in the planning models proposed. Research on this challenging topic is needed.

The previous issues, the discussion and the model of Sect. 6 refer to what is known as business-as-usual cases, when uncertainty can be somewhat easily represented with probability distributions. Other sources of uncertainty exist, however, and should be studied. Reliability and robustness are two such issues, as is resilience, i.e., the capability to rebound following an incident, and the operation plans to perform the recovery and return to a desired state of system and operation behavior. Advancing in this direction would also lead to a broader exploration of information-revelation mechanisms and multi-stage formulations.

We complete this "modeling" discussion noticing that most SND models, including those discussed in this chapter, assume that the behavior of customers, that is, of demand, is known with respect to economic, e.g., tariffs, and service-level criteria. This is true even when uncertainty in these elements is explicitly represented. Simply put, customers react to tariffs and quality-of-service levels and, consequently, so is the demand the carrier will ultimately service and the revenues it can potentially earn. Extending the SND to address such issues requires considering not only a profit-maximizing objective, but also modeling in mathematical terms the behavioral relations between tariffs, service-quality levels, and the willingness of customers to give a carrier their business. The revenue management literature is the starting point of this line of research noticing, however, that most of it targets people-servicing industries and that one cannot simply transpose those results to the freight transport environment (Bilegan et al. 2021). Bi-level SND programming, modeling the interactions between the carrier setting of tariffs and service levels and its self-interested customers, appears promising for a strategic-type of decision making on service level and pricing. Equally promising are the developments related to aggregated but accurate customer-behavior representations that could be integrated into carrier system-optimization SND models. An initial approach could define the response, through weights or probabilities, of customer types (e.g., regular, occasional, ad-hoc) to the carrier's discriminative service and tariff classes. The approximations, and their consequence in terms of demand volumes and revenues, would then become part of deterministic or stochastic SND formulations.

Addressing large SND models, particularly when several layers of design decisions are present and scheduling is involved, makes up another important research area. A first research direction in this area concerns the Dynamic Discretization Discovery (DDD) approach, which has shown its value when applied to standard SSND settings. More research is required, however, to refine and accelerate the method. We also need to extend it to the cases involving particular scheduling rules and patterns for some or all services according to, e.g., operation practices, modes or geographic/administrative zones. Another important extension, including for the DDD, concerns the problem settings with several design layers as one encounters when management of resources is considered or when, as in the case of railroads where one consolidates freight into cars, cars into blocks, and blocks into trains, each with potentially different temporal characteristics.

A second algorithmic research direction focuses on the dynamic generation of services (paths), resource work assignments (cycles), blocks (paths), and demand-flow itineraries (paths). With a few exceptions for resource management, this area has received little attention so far. Recall that, as illustrated in the models of this chapter, services are selected out of a set of potential services; the same discussion is relevant for the other system components as well. Authors rarely elaborate on the construction of the potential set. Obviously, it may correspond to all possible services, of all possible types, at all possible time instances. Two main issues with building a priori such a set. First, it would be of dimensions one could not address in most practical cases. Moreover, many of the potential services would be totally useless. But, which ones? How to avoid "bad" ones? The research on crew scheduling has clearly demonstrated that ad-hoc rules are not appropriate even when based on a company's own policies. The second issue is that, in practice, such a set would mean that the complete service structure and schedule of the carrier is to be built from scratch each time. This is generally not the case. Indeed, the demand structure of the next season is not totally different in most cases from the one at the last similar (e.g., Summer or Winter) season. Carriers then aim to update their previous schedule to adapt it to changes in demand patterns without imposing dramatic changes to their customers. Part of the service network is thus more or less fixed and a set of potential services must be built to reflect the changes in demand. The same question as previously stated arises with respect to building such a set. Research efforts have thus to be dedicated to extending the column-generation methodology to the SND and SSND cases with simultaneous generation of several types of paths and cycles.

With respect to solution-method approaches, recall that network design problems are NP-Hard in most cases of interest, and service network design ones are not different. Consequently, heuristic-type solution methods must be constructed. Yet, the research in this area is still not sufficiently developed. Particularly promising, and challenging, are matheuristics combining exact and meta-heuristic solution principles, ideally coupled with parallel optimization strategies, such as the Integrative Cooperative Search (Crainic 2019).

The development of efficient solution methods for stochastic SND and SSND is particularly challenging, even for the two-stage formulations of business-as-usual demand uncertainty case, which has been studied the most. The adequate representation of the "future", through sets of scenarios for example, is one the aspects contributing to this challenge. It raises issues regarding, the required number of scenarios, the purpose of scenario generation (represent the solution space or the optimal-solution neighborhood?), and how to generate the scenarios to serve this purpose. These questions are even more challenging when correlations and uncertainty in several problem parameters are considered. A scenario-based representation generally yields deterministic formulations of very large dimensions, very challenging to address as discussed above. The contributions mentioned in Sect. 6 and Chap. 9 make up the starting point of what should be a significant research effort on exact and matheuristic solution methods for stochastic service network design.

References

Albinski, S., Crainic, T. G., & Minner, S. (2020). The day-before truck platooning planning problem and the value of autonomous driving. Publication CIRRELT-2020-04, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et les transports, Université de Montréal, Montréal, QC, Canada.

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., 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.

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.

Armacost, A. P., Barnhart, C., & Ware, K. A. (2002). Composite variable formulations for express shipment service network design. *Transportation science, 36*(1), 1–20.

Bai, R., Wallace, S. W., Li, J., & Chong, A. Y. L. (2014). Stochastic service network design with rerouting. *Transportation Research Part B: Methodological, 60*, 50–65.

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.

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., & Van Woensel, T. (2017). From managing urban freight to smart city logistics networks. In K. Gakis, & P. Pardalos (Eds.), *Networks design and optimization for smart cities, series on computers and operations research* (Vol. 8, pp. 143–188). Singapore: World Scientific Publishing.

Belieres, S., Hewitt, M., Jozefowiez, N., Semet, F., & Van Woensel, T. (2020). A Benders decomposition-based approach for logistics service network design. *European Journal of Operational Research, 286*(2), 523–537.

Bilegan, I. C., Crainic, T. G., & Wang, Y. (2021). Scheduled service network design with revenue management considerations for intermodal barge transportation. Publication CIRRELT-2021-23, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport, Université de Montréal.

Boland, N., Hewitt, M., Marshall, L., & Savelsbergh, M. W. F. (2017). The continuous-time service network design problem. *Operations Research, 65*(5), 1303–1321.

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

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 ed., pp. 451–516). Norwell, MA: Kluwer Academic Publishers.

Crainic, T. G. (2019). Parallel metaheuristics and cooperative search. In M. Gendreau, & J.-Y. Potvin (Eds.), *Handbook of metaheuristics* (3rd ed., pp. 419–451). Berlin: Springer

Crainic, T. G., & Gendreau, M. (1986). Approximate formulas for the computation of connection delays under capacity restrictions in rail freight transportation. In *Research for Tomorrow's Transport Requirements, Fourth World Conference on Transport Research* (Vol. 2, pp. 1142–1155). Vancouver.

Crainic, T. G., & Kim, K. H. (2007). Intermodal transportation. In C. Barnhart, & G. Laporte (Eds.), *Transportation, Handbooks in Operations Research and Management Science*, chap 8 (Vol. 14, 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.

Crainic, T. G., & Roy, J. (1988). O.R. tools for tactical freight transportation planning. *European Journal of Operational Research, 33*(3), 290–297.

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., Ricciardi, N., & Storchi, G. (2009). Models for evaluating and planning city logistics transportation systems. *Transportation Science, 43*(4), 432–454.

Crainic, T. G., Hewitt, M., Toulouse, M., & Vu, D. M. (2014). Service network design with resource constraints. *Transportation Science, 50*(4), 1380–1393.

Crainic, T. G., Errico, F., Rei, W., & Ricciardi, N. (2016). Modeling demand uncertainty in two-tier city logistics tactical planning. *Transportation Science, 50*(2), 559–578.

Crainic, T. G., Hewitt, M., Toulouse, M., & Vu, D. M. (2018). Scheduled service network design with resource acquisition and management. *EURO Journal on Transportation and Logistics, 7*(3):277–309

Demir, E., Burgholzer, W., Hrušovskỳ, M., Arıkan, E., Jammernegg, W., & Van Woensel, T. (2016). A green intermodal service network design problem with travel time uncertainty. *Transportation Research Part B: Methodological, 93*, 789–807.

Erera, A., Hewitt, M., Savelsbergh, M., & Zhang, Y. (2013). Improved load plan design through integer programming based local search. *Transportation Science, 47*(3), 412–427.

Farvolden, J. M., & Powell, W. B. (1994). Subgradient methods for the service network design problem. *Transportation Science, 28*(3), 256–272.

Fontaine, P., Crainic, T. G., Jabali, O., & Rei, W. (2016). The impact of combining inbound and outbound demand in city logistics systems. In *41st IEEE Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 2, pp. 766–770). Piscataway, NJ: IEEE.

Fontaine, P., Crainic, T. G., Jabali, O., & Rei, W. (2021). Scheduled service network design with resource management for two-tier multimodal city logistics. *European Journal of Operational Research, 294*(2), 558–570.

He, Y., Péton, O., Lehuédé, F., Hewitt, M., Medina, J. (2019) A continuous-time service network design and routing problem. In Program ROADEF 2019. On-line at: roadef2019.univ-lehavre.fr/programme/ROADEF2019_submissions/ROADEF2019_paper_195.pdf

Hewitt, M. (2019). Enhanced dynamic discretization discovery for the continuous-time load plan design problem. *Transportation Science, 53*(6), 1731–1750.

Hewitt, M., Crainic, T. G., Nowak, M., & Rei, W. (2019). Scheduled service network design with resource acquisition and management under uncertainty. *Transportation Research Part B: Methodological 128*, 324–343.

Hoff, A., Lium, A. G., Løkketangen, A., & Crainic, T. G. (2010). A metaheuristic for stochastic service network design. *Journal of Heuristics, 16*(1), 653–679.

Jarrah, A. I., Johnson, E., & Neubert, L. C. (2009). Large-scale, less-than-truckload service network design. *Operations Research, 57*(3), 609–625.

Kazemzadeh, M. R. A., Crainic, T. G., & Gendron, B. (2019). A survey and taxonomy of multilayer network design. Publication CIRRELT-2019-11, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport, Université de Montréal, Montréal, QC, Canada.

Kim, D., Barnhart, C., Ware, K., & Reinhardt, G. (1999). Multimodal express package delivery: A service network design application. *Transportation Science 33*(4), 391–407.

Lanza, G., Crainic, T. G., Rei, W., & Ricciardi, N. (2018). A study on travel time stochasticity in service network design with quality targets. *Lecture Notes in Computer Science, 11184*, 401–416.

Lanza, G., Crainic, T. G., Rei, W., & Ricciardi, N. (2021). Service network design problem with quality targets and stochastic travel times. *European Journal of Operational Research, 288*(1), 30–46.

Lium, A. G., Crainic, T. G., & Wallace, S. W. (2007). Correlations in stochastic programming: A case from stochastic service network design. *Asia-Pacific Journal of Operational Research 24*(2), 161–179.

Lium, A. G., Crainic, T. G., & Wallace, S. W. (2009). A study of demand stochasticity in service network design. *Transportation Science, 43*(2), 144–157.

Marshall, L., Boland, N., Savelsbergh, M., & Hewitt, M. (2021). Interval-based dynamic discretization discovery for solving the continuous-time service network design problem. *Transportation Science 55*(1), 29–51.

Medina, J., Hewitt, M., Lehuédé, F., & Péton, O. (2019). Integrating long-haul and local transportation planning: The service network design and routing problem. *EURO Journal on Transportation and Logistics, 8*(2), 119–145.

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 metaheuristics for service network design with asset-balance requirements. *Transportation Science, 43*(2), 158–177.

Powell, W. B. (1986) A local improvement heuristic for the design of less-than-truckload motor carrier networks. *Transportation Science, 20*(4), 246–257.

Scherr, Y. O., Neumann-Saavedra, B. A., Hewitt, M., & Mattfeld, D. C. (2018) Service network design for same day delivery with mixed autonomous fleets. *Transportation Research Procedia, 30*, 23–32.

Scherr, Y. O., Saavedra, B. A. N., Hewitt, M., & Mattfeld, D. C. (2019). Service network design with mixed autonomous fleets. *Transportation Research Part E: Logistics and Transportation Review, 124*, 40–55.

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

Wang, X., Crainic, T. G., & Wallace, S. W. (2019). Stochastic network design for planning scheduled transportation services: The value of deterministic solutions. *INFORMS Journal on Computing, 31*(1), 153–170.

Wang, Z., & Qi, M. (2019). Service network design considering multiple types of seervices. *Transportation Research Part E, 126*, 1–14.

Wang, Z., & Qi, M. (2020) Robust service network design under demand uncertainty. *Transportation Science, 54*(32), 676–689.

Wieberneit, N. (2008). Service network design for freight transportation: A review. *OR Spectrum 30*(1), 77–112.

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