CIRRELT

CIRRELT-2024-30

Service Network Design for Consolidation-based Transportation – The Fundamentals

Teodor Gabriel Crainic

September 2024

Bureau de Montréal
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal (Québec) H3C 3J7
Tél : 1-514-343-7575
Télécopie : 1-514-343-7121

Bureau de Québec
Université Laval,
2325, rue de la Terrasse
Pavillon Palasis-Prince, local 2415
Québec (Québec) G1V 0A6
Tél : 1-418-656-2073
Télécopie : 1-418-656-2624

Service Network Design for Consolidation-based Transportation The Fundamentals[ł]

Teodor Gabriel Crainic

Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT) and School of Management, Université du Québec à Montréal

Abstract. The goal of this and companion chapters is to present a comprehensive overview of the Service Network Design modeling methodology to address the planning of consolidation-based freight transportation. This chapter focuses on the fundamental problem settings and modeling issues. It recalls the structure and main components of the physical and service networks of consolidation-based freight carriers, the associated tactical planning challenges, and the basic Service Network Design models addressing them.

Keywords: Service Network Design, modeling, freight transportation, consolidation, tactical planning.

Acknowledgements. While working on the project, the author held the UQAM Chair on Intelligent Logistics and Transportation Systems Planning, and was Adjunct Professor, Department of Computer Science and Operations Research, Université de Montréal. He gratefully acknowledges the financial support provided by the Natural Sciences and Engineering Council of Canada (NSERC), through its Discovery grant program, and of the Fonds de recherche du Québec through their infrastructure grants. The author is grateful to Ms. Diane Larin for her invaluable linguistic and editorial assistance, and to Professor Pirmin Fontaine, Dr Sara Khodaparasti, as well as the Editors and two anonymous referees, for their enriching comments and suggestions.

ł Part 1 of 2 – Part 2 CIRRELT-2024-31 “Service Network Design for Consolidation-based Transportation – Advanced Topics”.

* Corresponding author: teodorgabriel.crainic@cirrelt.net

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2024

1 Introduction

This chapter and its companion (Crainic, 2024) discuss *Service Network Design* (*SND*) models for the planning of consolidation-based freight transportation.

Freight transportation, and the supply chains it empowers, constitute a vital activity domain with major economic and social impacts, from the local neighborhood to the global trade exchanges generated by the needs and requirements of people, and public and private organizations making up the human society. As illustrated by the aftermaths of the political and health crises humanity faces regularly, one may argue that modern economies may not adequately perform and prosper without efficient transportation and logistics.

Traditionally defined in terms of delivering on time at the lowest possible cost to maximize profits, efficiency increasingly aims at reducing the "external" costs of transportation, pollution, congestion, and energy consumption, while continuing to support economic and social development. How to achieve these objectives is a major challenge that compounds the inherent complexity of the field where numerous and diverse stakeholders operate and interact within complex systems. Operations Research (OR) provides the scientific framework for the advanced models and methods required by the analysis and optimized planning and management of complex transportation and logistics systems that, in turn, challenge the discipline and motivate its continuous development.

Transportation, and trade, result from the interplay between *supply* and *demand*, within a given political, social, cultural, and economic, environment. To simplify the presentation, we thus identify *shippers*, which generate the multi-commodity, origin-destination demand side of the system, and *carriers* supplying the transportation and terminal resources, services, and capacity required to satisfy the demand. We focus on *consolidation-based* carriers and the core O.R. methodology addressing the planning of their operations.

Consolidation is a widely spread strategy aiming for increased operational and economic efficiency for shippers and carriers, by combining freight with different origins and destinations for loading into the same units (vehicles, containers, etc.) for their complete or partial journeys. The unit shipment cost and the journey duration should thus be reduced, benefiting all parties involved. Railroads, Less-than-Truckload (LTL) motor carriers, shipping companies moving containers on oceans, seas, rivers, and canals, postal services and express couriers, logistics-service providers, as well as synchromodal, City Logistics, and Physical Internet systems are prime examples of consolidation-based carriers moving a large and valuable part of the world trade (consumer goods in particular) over short, medium, long, and intercontinental distances.

To achieve their economic and service-quality goals, consolidation-based carriers organize their operations according to a *transportation plan* optimizing the resource utilization and deployment through a (more or less scheduled) service network that answers the shipper demand. The plan is built as part of the tactical planning process, and is therefore sometimes called *tactical plan* in the literature. It supports strategic planning and also guides operations and the management of resources and flows. Building it requires to address and conciliate, at the system/network level, many strongly interrelated planning and operation decisions and activ-

ities. It hence rises significant methodological challenges, particularly as systems grow more complex, requirements more constraining, and the world more uncertain.

Service Network Design (*SND*) is the O.R methodology of choice to address these challenges. It has been largely applied to consolidation-based freight transportation, as witnessed by a rich body of literature (reflected in, e.g., Crainic et al., 2021b). SND is part of the larger and important combinatorial-optimization family of *Network Design* (ND) problems and methodology (Crainic et al., 2021a; Crainic and Gendron, 2021; Crainic and Gendreau, 2021). One may say that SND is ND applied to transportation planning, and that SND and ND are part of a virtuous circle of methodological innovation targeting scientific development and performance-enhancing applications. Indeed, one observes that, any SND formulation may be described as a ND model by suitably (re-)defining the service network, and that many ND methodological developments were inspired by transportation planning, SND in particular.

The goal of this and companion chapters (Crainic, 2024) is to present a comprehensive overview of the general SND modeling methodology to address the planning of consolidation-based carrier activities. This chapter focuses on the fundamental SND problem settings and models, while the next is dedicated to more advanced topics.

The interested reader may consult a number of general and domain-specific surveys and syntheses (for brevity reasons, only the most recent are mentioned) and references within: Crainic and Rei (2024); Crainic and Hewitt (2021) (general syntheses), Chouman and Crainic (2021) (railroads), Bakir et al. (2021) (LTL motor carriers), Christiansen et al. (2021) (maritime transportation), and Crainic et al. (2021c) (City Logistics).

The chapter is organized as follows. Section 2 recalls the operations structure and associated planning issues of consolidation-based freight carriers. Section 3 introduces the fundamental SND concepts, notation, and formulations. Finally, Section 4 discusses the modeling of the main components of consolidation-based carriers for SND formulations, namely, the physical system, the demand, and the service network. We conclude in Section 5.

2 Consolidation-based Freight Transportation

Bruni et al. (2024) discuss in some depth the concepts of shipper and carrier used to classify demand and supply stakeholders, respectively, in consolidation-based freight transportation. Briefly, the generic “shipper" term designates an entity requiring transportation for its freight. It includes producers, traders, brokers, importers/exporters, logistics service providers, as well as whole- and retail buyers, sellers, and distributors of goods. On the supply side, the general “carrier" term encompasses the actual modal and intermodal freight transport firms, logistics-service providers, as well as integrated multi-stakeholder Physical Internet, synchromodal, and City Logistics systems.

We now recall the basic concepts of the supply facet of consolidation-based freight transportation, namely, the carrier physical and service networks (Section 2.1) and the tactical planning scope and objectives (Section 2.2).

2.1 Carriers, Consolidation, and Services

Carriers providing consolidation-based services operate on infrastructure-defined single, multi, or intermodal *physical networks*, the latter term being generally used when freight packaged at origin, mostly in containers, is not handled before it is unpacked at destination. Postal / express-courier services, container transport, as well as City Logistics, Physical Internet, and synchromodal systems often involve more than one transportation mode, the transfer of loads from one to the next taking place at intermodal terminals. LTL motor carriers, railroads, and shipping companies are generally identified as single-mode, as they often operate exclusively trucks, trains, and ships, respectively. Notice, however, that such carriers may also operate multi/intermodal networks, as illustrated by railroads owning motor carriers, and maritime shipping companies owning railroads or motor carriers.

The carrier physical system is made up of *terminals* connected by physical, e.g., highways and rail tracks, or conceptual, e.g., maritime and air corridors, infrastructure. Inter-terminal links may 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 as well as roads and highways mostly everywhere. The left side of Figure 1 illustrates a multimodal physical network, where highways (black dashed arrows), regular roads (black dotted arrows), and rail tracks (red full arrows) connect eleven terminals.

One identifies two main categories of terminals. The largest category consists of *regional* terminals, where most of the demand from the surrounding regions is brought in to be transported by the carrier, and where the freight flows coming from other regions terminate their trips before being distributed to their final destinations. Most rail stations, LTL terminals, airports, and deep-sea, river/canal ports belong to this type, illustrated by numbered disks in Figure 1.

Hubs make up the second category. While also being the regional terminal for its hinterland, the main role of a hub is to *consolidate* the flows in and out of its associated regional terminals for efficient long-haul transportation and economies of scale. LTL major terminals (often called breakbulks), main classification/blocking/marshaling railroad yards, and large maritime ports for intermodal (container-based) traffic belong to this group. 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, and airports. Three hubs are part of the physical network on the left of Figure 1, where terminals 1, 2, 3, and 4 are associated to hub A, 5, 6, and 7 to B, and 4, 7, and 8 to C, the latter illustrating the possibility og a regional hub to be associated to more than one hub. The designation of particular terminals as “hubs", together with most regional-terminal-to-hub association, decisions are part of strategic planning, while their utilization is determined as part of the tactical plan.

The multi-commodity, *origin-destination* (*OD*) *demand* is defined as a quantity of freight to be moved between specific origin and destination terminals. For planning purposes, this quantity is expressed in the same units used to define the capacity of the services offered by the carrier (e.g., total weight, volume, or both, or total number of homogeneous items). Many

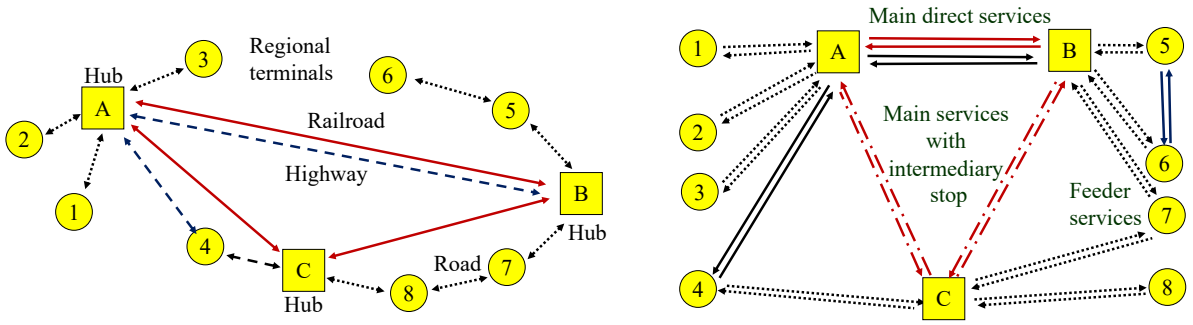

Figure 1: Hub-and-spoke service network

other attributes may be associated to each demand according to the particular context, including shipper demand type (e.g., priority, express, regular delivery), timing (e.g., availability at origin and due-date at destination), fare and penalties for missing the agreed-upon service-quality targets, particular product with related physical characteristics and requirements in terms of vehicle type (e.g., refrigerated, multi-platform for containers or vehicles, etc.), and so on.

Carriers aim to profitably and efficiently satisfy the requests of many shippers by consolidating their loads and moving them together within appropriate vehicles. Moreover, they also want to offer reasonable service quality (e.g., not waiting beyond customer willingness for vehicles to fill up with other demands), and address shipper concerns relative to demands of too low a volume or value to justify paying the tariffs associated with a direct, dedicated transport.

To address these challenges and achieve these goals, consolidation-based carriers organize their operations into *hub-and-spoke service networks* determined as part of the tactical plan.

Each *service* is defined by a mode and physical route between a pair of origin and destination terminals, a possible sequence of intermediate stops on that route where loads may be picked up / dropped off, a schedule indicating more or less precisely the arrival and departure times at/from the concerned terminals, as well as particular equipment, operations, and economic characteristics. A service hub-and-spoke network is illustrated on the right of Figure 1. Defined on the physical network on the left of the figure, the thick black and red full arrows illustrate eight direct main *long-haul* services. Four of these connect the A and B hubs. The four others, illustrate the possibility of defining main services between regional terminals (5 $\leftrightarrow$6) or between those and hubs (A$\leftrightarrow$4) when justified by the volume or value of goods. The red point-dash arrows stand for services between hubs A and B, with a stop at hub C, while the black dotted arrows illustrate *feeder* services moving freight between regional terminals and their assigned hubs.

When operating hub-and-spoke service networks, carriers first call on feeder services to move the low-volume/value loads from regional terminals to hubs. There, loads are sorted (*classified* is the term used in several settings, e.g., freight railways) and consolidated into larger flows, which are routed to other hubs by high-frequency, high-capacity long-haul 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. Once at the last hub on their itineraries, loads are unloaded, possibly sorted, and loaded on feeder services to be moved to their destination regional terminal, to be distributed from there to their final destinations.

A hub-and-spoke service network concentrates the multi-commodity demand flows, providing the means for economies of scale and more efficient utilization of resources for the carrier, high-frequency service for the consolidated demand, and lower tariffs for shippers. Possible drawbacks of this type of organization are increased delays for demand due to longer routes and more time spent going through terminals, as well as more sophisticated terminal layouts, equipment, and human resources. Indeed, terminals play a role significantly broader than simply loading/unloading freight, as vehicle and freight classification, consolidation, and inter-service transfer are time, cost, and resource-consuming operations performed in terminals.

The adequate, hopefully optimal, design of the service network and planning of its operations and resource utilization, is required to avoid such pitfalls and to reap the full benefits of consolidation-based transportation. This is the scope of carrier tactical planning.

2.2 Tactical Planning

One may describe *planning* as the preparation of operations in anticipation of a future situation, whereas *execution* is the acting, including updating/adjusting, of the pre-established plans. Consolidation-based carriers engage into a rather extensive set of strategic, tactical, and operational planning activities prior to executing the resulting plans (Crainic and Rei, 2024). We focus on tactical planning in this chapter, but briefly touch on the other two given the role tactical plans and SND methodology play in those contexts.

Strategic planning addresses long-term decisions on market deployment, system design (e.g., hub selection and design), operation strategies, and acquisition of major resources. It concerns long planning horizons and involves rather high-level management. Tactical-planning SND models and methods may be used as policy and performance-evaluation tools for strategic scenarios. One may, e.g., adapt the SND for an aggregation of operational details, demand, and costs appropriate for strategic planning. Alternatively, one may simulate the scenario impact through the performance of the corresponding tactical plan.

The goal of operational planning is to prepare the tactical plan for execution given the observed conditions, in terms of demand and operations. The adjustment may be performed at application-specific intervals, every week and day, for example. Notice that, the same SND methodology may be often used in this case, albeit with different degrees of aggregation and time frames (Crainic et al., 2009).

The scope and goal of tactical planning is to build a *transportation plan* to operate profitably and efficiently while addressing simultaneously the cost and service-quality transportation requirements of a large group of shippers. The system-wide plan involves deciding on the selection and scheduling of services, the transfer and consolidation activities of freight and

vehicles in terminals (as well as the convoy makeup and dismantling for rail, road, and barge trains), the assignment and management of resources to support the selected services, and the routing of the freight of each particular demand through the resulting service network.

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. Two examples to illustrate these challenges and necessary trade-offs. First, increasing the number of services operated during a certain time interval between two terminals improves customer service but may decrease the availability of resources elsewhere, as well as increase congestion in terminals and on certain infrastructure links, such as rail tracks, thus increasing costs and deteriorating customer service. Consider, second, routing freight through intermediate terminals, to be re-classified and consolidated before transfer to the next service. On the positive side, this generally results in faster freight departure from the origin terminal and better rolling-equipment utilization. On the negative one, it would also result in additional unloading, consolidation, and loading operations, involving larger delays and creating higher congestion levels at terminals. Transfers may also decrease the shipment delivery reliability. The alternative, 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 and, thus, higher costs for the system and tariffs for the shippers.

Tactical planning is performed for a medium-term planning horizon, which may extend from a few (e.g., LTL motor carriers) up to six or twelve months, and which we call *season*. It builds the tactical plan considering the consistency of the so-called *regular demand*, that is, shippers that are strongly believed to bring a consistent level of business on a regular basis for the coming season. This prediction, which may or may not follow from formal forecasting methods, is based on a combination of long-term contracts, informal understanding with long-standing and trustful customers, and market estimation by sales and customer-relation personnel. In terms of volume, regular demand is expected to make up a good part, e.g., 75% - 80%, of the pick demand to be serviced on a “normal activity period” (e.g., day). In terms of consistency, demand, and, hence, service, is expected to be repetitive according to a certain pattern, e.g., every day or week. The plan produced by the tactical-planning process is thus for a given time duration, called *schedule length*, and is to be applied repetitively for the duration of the season. Of course, particular service networks may be build for specific moments, e.g., for week days different from weekends, and be combined to form the repetitive plan and schedule length. The tactical plan-building methodology is presented next.

3 Service Network Design

We start with a brief presentation of SND core concepts, followed by the model formulation for the basic problem-setting case (Section 3.1) and extensions (Section 3.2).

The input of a service network design model for tactical planning of a consolidation-based carrier follows the transportation-planning structure and includes demand and supply com-

ponents. The former gives the origin-destination demand matrix (or matrices, when several products are considered) of the freight to be moved between particular pairs of origin and destination terminals. The latter specifies the physical network, the set of potential services, and the available resources, out of which the service network is to be built to answer this demand. The physical and operational attributes of demand and supply elements, and the rules associating them and specifying how the system is to be operated, are also part of the input of the problem settings and SND models. Section 4 discusses the modeling of these elements and rules within various carrier-planning contexts.

The SND decision variables represent and integrate two major sets of decisions, namely, the design of the service network and the utilization of that network to service demand. The main *design* or *selection* decisions build the service network by selecting the services to be operated out of a set of potential services. When resource management is explicitly accounted for within the tactical-planning decision process, decisions relative to the selection, association to services, and work rules of those resources are also part of the design decision set (Crainic, 2024).

The main *utilization* or *flow* decisions concern how each individual demand is satisfied using the designed service network, that is, its *itinerary* specifying the sequence of services, terminals, and terminal operations (loading, unloading, inter-service transfer, classification and consolidation), used to move the corresponding freight. Several itineraries may be used simultaneously for a given demand, when its shipment may be split among several service paths between the respective origin and destination terminals.

The objective function of an SND formulation reflects the carrier's economics, as well as, increasingly, its customer-service and societal concerns and objectives. Carriers aim to maximize their profit, generally defined as revenue minus cost. Most tactical-planning applications assume the set of shippers, corresponding commercial relations, and expected/estimated revenue to be known and fixed for the tactical planning horizon. Hence, the transportation plan aims for the minimization of the total operating cost only, which also reflects the traditional carrier and shipper objectives to "get there fast at the lowest possible cost".

The definition of "cost" is becoming broader, however. On the one hand, shippers expect not only low fares, but also high-quality service, which is generally measured by speed and reliability of service and delivery times. On the other hand, carriers are increasingly sensitive to social and legal pressure regarding energy consumption and environmental impacts. Service performance measures reflecting these expectations and concerns are then added to the objective function of the SND optimization formulation. This yields a generalized cost function that captures the trade-offs between operating costs, service quality, and societal impact. We discuss these issues in Section 4.3.

Notice that, profit maximization is the objective function of the SND model when the selection of shippers or demands is explicitly included in the problem setting, or when revenue-management policies are applied by the carrier (Crainic, 2024).

One generally classifies SND problems and models relative to two dimensions, time and uncertainty. Along the *time* dimension, one finds *static* and *time-dependent* (the term "time-

sensitive" is also used) problem settings and formulations. The former, discussed in this chapter, assumes that neither demand, nor any other problem characteristic varies during the schedule length and the planning horizon considered. Notice that, the schedule length is implicitly defined as the normal activity period, Time-dependent settings include an explicit or implicit representation of demand and activities in time, and are addressed in Crainic (2024).

With respect to uncertainty, one generally identifies *deterministic* and *stochastic* SND settings and models. Reflecting most contributions in the literature, this chapter focuses on deterministic SND, assuming known values for the system parameters, both on the demand and the supply facets, which do not change for the planning-horizon duration. This translates in fixed figures, which may come from historical and field-knowledge data (one finds a lot of average measures) or single-point estimations, when forecasting methods are used. (See Crainic, 2024, for SND addressing uncertainty).

3.1 The Basic SND Models

The basic SND problem setting is static and deterministic with linear costs, and direct services (i.e., operating without intermediate stops) only. Here as in the rest of the chapter, we follow the notation of Crainic and Hewitt (2021).

The fundamental system and model components of all SND problem settings and formulations, discussed in some depth in Section 4, are:

Physical network $\mathscr{G}^{\text{PH}} = (\mathscr{N}^{\text{PH}}, \mathscr{A}^{\text{PH}})$, where $\mathscr{N}^{\text{PH}}$ stands for the set of facilities, hubs and regional terminals, connected by the physical or conceptual arcs of set $\mathscr{A}^{\text{PH}}$ (illustrated on the left side of Figure 1, it is the topic of Section 4.1);

Demand for transportation of a set $\mathscr{K}$ of OD commodities, each $k \in \mathscr{K}$ requiring to move a quantity of freight vol_k from its origin $O(k)$ to its destination $D(k)$ (Section 4.2);

Service network $\mathscr{G} = (\mathscr{N}, \mathscr{A})$, defined based on the physical nodes of the system and the set of potential services Σ, within the context of the carrier resources, operation rules, economics, and service goals (the right side of Figure 1 and Section 4.3).

In the basic setting, $\mathscr{N} = \mathscr{N}^{\text{PH}}$ and $\mathscr{A} = \Sigma$, the service-arc association being encapsulated in $\sigma(a) \in \Sigma$, stating that service σ corresponds to (defines) arc a. Each service $\sigma \in \Sigma$ is characterized by a fixed cost f_σ, incurred when selecting and operating it, a unit freight-transportation cost c_σ or $c^k_\sigma, k \in \mathscr{K}$, when commodity characteristics are relevant, and a capacity u_σ, representing the total volume of freight the service may load and haul; commodity-specific capacities $u^k_\sigma, k \in \mathscr{K}$, are defined when relevant (e.g., hazardous material-loaded containers on trains or ships). The cost and capacity figures are thus inherited by the corresponding arc $a \in \mathscr{A}$, i.e., $c^k_a = c^k_\sigma$ ($c_a = c_\sigma$) and $u^k_a = u^k_\sigma$ ($u_a = u_\sigma$).

The design decision variables of the basic setting model the selection of services through binary variables $y_\sigma \in \{0,1\}, \sigma \in \Sigma$. The flow decision variables are generally defined on the arcs of the service network $\mathscr{G}$, taking the form $x^k_a \geq 0$, $a \in \mathscr{A}, k \in \mathscr{K}$, prescribing the amount of commodity k that travels on arc a, i.e., on service $\sigma(a) \in \Sigma$.

Formally, then, the basic arc-based SND formulation seeks to

$$\min \sum_{\sigma\in\Sigma} f_\sigma y_\sigma + \sum_{k\in\mathscr{K}}\sum_{a\in\mathscr{A}} c_a^k x_a^k \tag{1}$$

$$\text{s.t.} \sum_{a\in\mathscr{A}_\eta^+} x_a^k - \sum_{a\in\mathscr{A}_\eta^-} x_a^k = d_k \qquad \eta\in\mathscr{N}, k\in\mathscr{K}, \tag{2}$$

$$\sum_{k\in\mathscr{K}} x_a^k \le u_a y_{\sigma(a)}, \qquad a\in\mathscr{A}, \tag{3}$$

$$x_a^k \le u_a^k y_{\sigma(a)}, \qquad a\in\mathscr{A}, k\in\mathscr{K}, \tag{4}$$

$$y_\sigma \in \{0,1\}, \qquad \sigma\in\Sigma, \tag{5}$$

$$x_a^k \ge 0, \qquad a\in\mathscr{A}, k\in\mathscr{K}, \tag{6}$$

where $\mathscr{A}_\eta^+ = \{(\eta,\eta')\in\mathscr{A}\}$ and $\mathscr{A}_\eta^- = \{(\eta',\eta)\in\mathscr{A}\}$ define the sets of outgoing and incoming arcs for node $\eta\in\mathscr{N}$, respectively, while $d_k = vol_k$ at the demand origin $\eta = O(k)$, $-vol_k$ at the demand destination $\eta = D(k)$, and zero at all other nodes.

The objective of the SND minimizes the total cost of operating the system, computed as the sum of the fixed costs associated with selecting the service network and the variable cost of transporting commodities using the selected services. Equations (2) are often referred to as *flow-balance* constraints and ensure that all of a commodity's demand departs from its origin, arrives at its destination, and departs from any other locations at which it arrives. The expression on the left-hand side of the *linking* constraints (3) computes the total flow traveling on arc $a\in\mathscr{A}$, whereas the expression on the right-hand side gives the global arc capacity provided by the corresponding service (selected or not). The commodity-disaggregated linking constrains are given by (4). Constraints (5) and (6) define the variable domains.

The arc-flow variables x describe how the demand loads are moving through the selected service network. Clearly, the flow of each demand may follow one or several paths from its origin to its destination. An equivalent formulation explicitly identifies these paths.

Let $\Pi^k, k\in\mathscr{K}$, identify the set of possible *itineraries* (paths) of commodity k on the potential service network (all potential services and flow-terminal operations). Following classic network notation, let set $\mathscr{A}_\pi^k$ hold the sequence of arcs $a\in\mathscr{A}$ making up the itinerary $\pi\in\Pi^k$, and let the Kronecker delta coefficients $\delta_a^{\pi k}$ define path π of k, i.e., $\delta_a^{\pi k} = 1$ when $a\in\mathscr{A}_\pi^k$, 0, otherwise. The unit itinerary flow cost is then defined as $c_\pi^k = \sum_{a\in\mathscr{A}_\pi^k} c_a^k, k\in\mathscr{K}$.

Let the *itinerary-flow* decision variable h_π^k be the amount of commodity $k\in\mathscr{K}$ moved on its itinerary $\pi\in\Pi^k$. Then, $x_a^k = \sum_{\pi\in\Pi^k} \delta_a^{\pi k} h_\pi^k, a\in\mathscr{A}, k\in\mathscr{K}$, and the basic path-based SND formulation becomes

$$\min \sum_{\sigma\in\Sigma} f_\sigma y_\sigma + \sum_{k\in\mathscr{K}}\sum_{\pi\in\Pi^k} c_\pi^k h_\pi^k \tag{7}$$

$$\text{s.t.} \sum_{\pi\in\Pi^k} h_\pi^k = d_k, \qquad k\in\mathscr{K}, \tag{8}$$

$$\sum_{k\in\mathscr{K}} \delta_a^{\pi k} h_\pi^k \le u_a y_{\sigma(a)}, \qquad a\in\mathscr{A}, \tag{9}$$

$$\delta_a^{\pi k} h_\pi^k \leq u_a^k y_{\sigma(a)}, \qquad a \in \mathscr{A}, k \in \mathscr{K}, \tag{10}$$

$$y_\sigma \in \{0,1\}, \qquad \sigma \in \Sigma, \tag{11}$$

$$h_\pi^k \geq 0, \qquad \pi \in \Pi^k, k \in \mathscr{K}. \tag{12}$$

Recall that, the linear-cost network design arc and path-based formulations are equivalent, that is, they yield the same service network and objective-function value (Crainic et al., 2021a). This holds for the SND formlations above, as well as for the linear-cost settings that follow.

3.2 Problem and Formulation Extensions

The particular characteristics of the various carriers, modes, and operation policies yield a number of extensions to the basic formulations.

Multi-leg services. A service $\sigma \in \Sigma$ follows a path in the physical network from its origin $O(\sigma)$ to its destination $D(\sigma)$ (Figure 1). Several other terminals may be located along this path. A *direct* service passes by these terminals without stopping. The service route is then represented as a single arc $a \in \mathscr{A}$ of the potential service network, as defined above.

A *multi-leg* service halts at intermediary terminals on its route to load and unload cargo, not only at the origin and destination terminals of the commodity, but also for consolidation and transfer purposes. When convoys are involved (e.g., rail, road, and barge trains), the service may also stop to pick up or drop off individual or groups of vehicles (e.g., car or blocks for railroads and trailers for LTL motor carriers operating multi-trailer road trains). The service route is then described by the sequences of $n(\sigma)$ terminal stops and $n(\sigma)-1$ service legs connecting them. The single-leg, direct, service case has $n(\sigma) = 2$. Let $\mathscr{N}^{\text{PH}}(\sigma) = \{\eta_i(\sigma) \mid i = 1,\ldots,n(\sigma), O(\sigma) = \eta_1, D(\sigma) = \eta_{n(\sigma)}\}$ be the stop sequence of service $\sigma \in \Sigma$. Then, the *service leg* $l_i^\sigma = (\eta_i, \eta_{i+1})$ is defined as the sub-path connecting the consecutive terminals η_i, $\eta_{i+1} \in \mathscr{N}^{\text{PH}}(\sigma)$ of the route of service σ, with $\mathscr{L}(\sigma) = \{l_i^\sigma, i = 1,\ldots,n(\sigma)-1\}, \sigma \in \Sigma$.

Multi-leg services yield several arcs in the potential service network G. Each service leg makes up an arc, yielding $\mathscr{A} = \mathscr{L} = \cup_{\sigma\in\Sigma}\mathscr{L}(\sigma)$. Let $l_i^{\sigma(a)}$ be the leg i of service σ defining arc a. Then $u_a = u_{l_i^{\sigma(a)}}$ and $u_a^k = u^k_{l_i^{\sigma(a)}}$ in constraints (3) - (4).

Service frequency. Independently of operating according to fixed schedules or not, a service may have several departures during a given time interval. Such cases are modeled by defining non-negative integer service-selection decision variables $y_\sigma \in \mathbb{Z}_+, \sigma \in \Sigma$.

Service capacity feasibility. OR models may raise feasibility issues when including capacity limitations explicitly. These are disturbing, both from a computational point of view and because, in practice, there is "always" a feasible solution, even if quite costly, e.g., by calling on

ad-hoc capacity provided by additional vehicles or outsourcing part of the demand transportation, and paying the additional costs. The simplest approach to address this issue is to include *dummy arcs* $a^k = (O(k), D(k)), k \in \mathscr{K}$, between the origin and destination of each commodity with no capacity restrictions and appropriately high unit cost c_{a^k}. The associated *slack-flow variable* ς^k then captures the volume of demand unfulfilled by the capacity of the selected services, and takes care of the feasibility issue. The term $\sum_{k\in\mathscr{K}} c_{a^k}\varsigma^k$ is added to the objective function, and the flow conservation constraints (2) become

$$\sum_{a\in\mathscr{A}^+_\eta} x^k_a - \sum_{a\in\mathscr{A}^-_\eta} x^k_a = \begin{cases} vol_k - \varsigma^k, & \text{if } \eta = O(k), \\ -vol_k + \varsigma^k, & \text{if } \eta = D(k), \\ 0, & \text{otherwise,} \end{cases} \quad \eta \in \mathscr{N}, k \in \mathscr{K}. \tag{13}$$

These modifications are implicit when the set of dummy arcs a^k is included in $\mathscr{A}$.

Demand distribution. The previous arc and path basic formulations address the case when the volume of any particular OD demand may be *split* between several itineraries. While this corresponds to a very large range of applications, one equally finds many situations where the freight of an OD demand must travel together, following a single path. This is achieved by modifying the definitions of the flow variables:

- $x^k_a = 1$ if commodity $k \in \mathscr{K}$ travels on the arc $a \in \mathscr{A}$ (on service $\sigma(a)$), and 0, otherwise;
- $h^k_\pi = 1$ if commodity $k \in \mathscr{K}$ is moved on its itinerary $\pi \in \Pi^k$, and 0, otherwise .

The arc (path) formulation is then modified by multiplying x^k_a (h^k_π, respectively) by d_k in the objective function (1) ((7)) and constraints (2) - (4) ((8) - (10)), and by changing the domain restrictions of the flow variables (6) to $x^k_a \in \{0,1\}$ ((12) to $h^k_\pi \in \{0,1\}$).

Objective function and particular system features. The linear-cost formulation (1) represents a very broad set of issues and problem settings. Other cases require more elaborate objective-function formulations, however, e.g., modeling of the total cost associated to several more-or-less simultaneous service departures, and the explicit representation of congestion phenomena in terminals or on the infrastructure. Moreover, a number of restrictions and requirements may characterize the carrier's system, e.g., limited terminal capacity to handle vehicles and freight, freight-vehicle compatibility restrictions, and budgetary limits.

We discuss these issues and more in Section 4.3. For now, we represent them through a general notation framework:

- $\phi_\sigma(y), \sigma \in \Sigma$: Fixed cost of selecting service σ given the selected services y;
- $\varphi_{ak}(y,x), k \in \mathscr{K}$: Unit-transportation cost of commodity $k \in \mathscr{K}$ on arc $a \in \mathscr{A}$, given the selected services y and flow distribution x;
- Ψ: Set of special features constraints linking the decision variables (y,x) and restricting their domains.

The general basic SND formulation then takes the form

$$\min \sum_{\sigma\in\Sigma} \phi_\sigma(y) y_\sigma + \sum_{k\in\mathscr{K}} \sum_{a\in\mathscr{A}} \varphi_{ak}(y,x) x_a^k \tag{14}$$

subject to constraints (2) - (4) and (6), plus

$$y_\sigma \in \mathbb{Z}_+, \quad \sigma \in \Sigma, \tag{15}$$

$$(y,x) \in \Psi. \tag{16}$$

The basic formulations presented in this section are found in many contributions in the literature targeting freight-transportation planning issues, as synthesized in the papers and chapters referred to in the chapter. They also emphasize the network-design nature of SND. Indeed, as stated in the Introduction, any SND model may be cast as a ND formulation on an appropriately-defined network. Thus, model (1) - (6) corresponds to the multi-commodity, fixed-cost, capacitated network design problem defined on a network with $\mathscr{A} = \Sigma$ (Crainic et al., 2021a). Moreover, the (14) - (16) formulation is the same as the general network design model defined in the seminal Magnanti and Wong (1984) paper, where the authors also showed that ND encompasses Minimal Spanning Tree, Shortest Path, Traveling Salesman, Vehicle Routing, and Facility Location problems as special cases.

4 Modeling Systems & Components

This section is dedicated to the modeling of the three main components of SND formulations for the tactical planning of consolidation-based freight carriers, namely, the physical system, Section 4.1, the demand, Section 4.2, and the potential service network, Section 4.3.

4.1 Physical system

Planning is performed and SND models are formulated on a network representation of the physical infrastructure on which the carrier operates. Modeling this *physical network* $\mathscr{G}^{\text{PH}} = (\mathscr{N}^{\text{PH}}, \mathscr{A}^{\text{PH}})$ means representing the infrastructure, activity, and performance of the terminals and inter-terminal connections at a level relevant for tactical planning.

Terminals = Network nodes $\eta \in \mathscr{N}^{\text{PH}}$ stand for major hubs and regional terminals equipped and maned to perform many types of handling operations on freight and vehicles. The other terminals are represented by aggregating the freight originating and terminating there and assigning it to the regional terminal or hub to which the feeder line is attached.

Several node attributes may thus be defined to represent the characteristics and capabilities of the corresponding terminals to handle freight, vehicles, and convoys (when relevant). Three

types of attributes are generally encountered, economic, capacity, and efficiency, measured for a time duration appropriate for tactical-planning purposes, e.g., a work shift or day.

Economic node measures mean primarily *unit operating costs*. *Holding* costs may also be defined to represent the cost of time spent waiting in the terminal by commodities or vehicles (e.g., the time-related port fees for ships). Depending on the data available and the desired degree of detail, these costs may be commodity- or mode / vehicle type-specific or a general figure for the terminal. One or several *capacity* measures are generally defined to represent the terminal operational capability in terms of the maximum volume of freight, commodity-specific or no, and number of vehicles or convoys the terminal may service. Within these capacity limits, *efficiency* measures generally stand either for the throughput, e.g., volumes processed or vehicles and convoys services, or the duration / delay of the activity.

Often, a single measure is defined for the terminal attributes, aggregating into an unique figure the average (sometimes the variance is given as well) terminal performance. Operation-specific measures may however be defined for particular activities with particular impacts. Thus, for example, several measures may be defined for classification rail yards, including 1) aggregated measures of the reception, inspection, and departure activities (including loading/unloading when appropriate); 2) transfer of cars or blocks from one train to another; 3) railcar classification/consolidation into blocks; and 4) waiting for the block to be completed or the next train to be available (e.g., Crainic et al., 1984).

The physical networks of most SND models include a single node for each terminal and single economic and capacity values. More complex representations may be used to account for particular operations or terminal characteristics, e.g., different traffic directions and capacities, usually taking the form of mini networks. The simplest mini network terminal representation involves two nodes, capturing the incoming and outgoing traffic, respectively, the terminal attributes being assigned to the arc connecting them (e.g., Zhu et al., 2014).

Larger networks may capture a richer set of activities and measures. Figure 2 illustrates such structures through the representation of an intermodal terminal connecting maritime and land networks (inspired by Andersen et al., 2009). Several types of nodes make up this network. A blue square represents the navigation network and two squares stand for the rail (in red) and road (black) modes of the land network, respectively (obviously, more detailed networks appear in actual applications). Two blue disks represent the unloading and loading of ships at quay, two red ones stand for the dismantling and forming of trains and blocks in the rail yard of the port, and one black disk modeling the port gate that controls the passage of incoming and outgoing motor vehicles. These last five nodes, together with the links connecting them, make up the actual intermodal terminal. The solid blue arrows between the maritime and the port-quay networks represent the links of the mini networks capturing the arrival and departures of ships. The same role is played by the black and red solid arrows for the road and rail modes of the land network. The dot-dashed red arrow connecting the two port-rail yard nodes stands for the situation when railcars are brought into the port and taken out of it by the railroad's own engines, in-bounding ones needing to turn around and be assigned to outgoing movements. Finally, the red and black dashed arrows stand for the transfers between the maritime network and the rail and road components, respectively, of the land network.

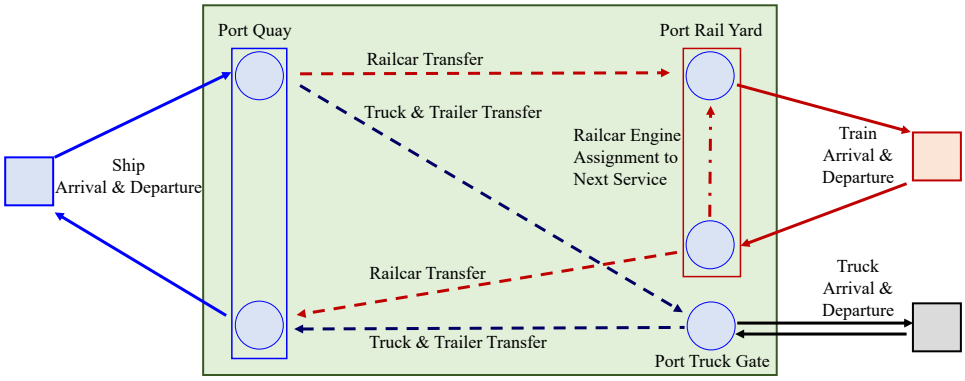

Figure 2: Mini-network terminal representation

Moving among terminals = Network arcs. Each arc $a \in \mathscr{A}^{\text{PH}}$ stands for the possibility to move directly between the terminals represented by the two nodes defining it (stops for crew rest or refueling may be accounted for through the arc attributes, but are not included explicitly). The arc thus represents a path on the infrastructure network on which the carrier operates. Arcs in mini networks modeling particular terminals follow the same rule, but may also stand for specific operations, e.g., ship unloading of containers, rather than actual movements.

The term “arc" indicates that the connection between its defining nodes is *directed*, providing the means to represent direction-specific characteristics, measures, and flows. Arcs are also *modal*, *unimodal*, in fact, each being characterized by a specific transportation mode. Parallel arcs may be defined between two adjacent nodes, when more than one modal connection exists between the two corresponding terminals.

Mode is a very general term in transportation and logistics, referring to different concepts and definitions according to the topic at hand. Thus, in a very fundamental way, it may refer to the nature of the support of transportation, e.g., land, water, air, and space, while it may also describe in a very detailed way to the specifics of a transportation system, discriminating, for example, by the type of infrastructure (highways, national roads, local roads), traction type (e.g., diesel, hydrogen, or electric), speed (express or regular), etc. Most SND settings in the literature refer to a “classical" mode definition, based on a high-level combination of such elements:

***Road*,** more particularly LTL motor inter-urban transportation, with an increasingly larger array of road-based modes in cities, e.g., people-driven or autonomous electric or hydrogen powered vans, cargo bikes, and robots;

***Rail*,** representing the *general* train services performing consolidation-based rail transportation for mostly all categories of goods, and *intermodal* services dedicated to moving containerized cargo;

***Navigation*,** with the largest part of the literature dedicated to maritime transport of containers performed by liner ships; One notices an increased interest in intermodal barge river & canal navigation;

***Air*,** with relatively few contributions related to SND, most of which target the design of mul-

timodal express-courier networks (contributions addressing the air-cargo industry are mostly directed toward revenue-management issues).

Several other attributes may be associated to each arc of $\mathscr{G}^{\text{PH}}$. The nature of each attribute determines how it is computed. *Length* is thus equal to the sum of the lengths of the links of the infrastructure path, while the *capacity* is computed as the minimum of the corresponding capacities. When defined, *time* and *cost* are also computed as sums, including, when relevant, the time or cost required by the technical activities (e.g., inspection or refueling) at facilities which are part of the arc definition (as indicated above). It is noteworthy, however, that traveling cost and time are dependent upon the characteristics of each service, and are thus mostly associated to the service network rather than the physical one.

The definition of arc *capacity* follows the general rules indicated for nodes. It is generally measured in numbers of vehicles, convoys, or loading units (e.g., containers) that can pass through, be processed by, or wait at the relevant infrastructure during a given time length. Several other capacity measures may be relevant, including

- *Length* limiting, e.g., how long a rail train may be given the characteristics of the tracks and signaling system, and the total length of the ships that can moor simultaneously at a given port quay.
- Maximum *weight* of vehicles or convoys moving on the arc.
- *Volume* of freight which may be processed or stored at the terminal (mini-network arc).
- *Water levels* in port terminals and the various segments of river, canal, sea, and ocean routes may significantly impact the capacity of the associated physical arcs in terms of the vessels that may be accommodated for berthing or navigation, and the weight of the cargo those vessels may carry (bridge heights may also limit these measures).

Several attribute measures may be simultaneously defined for a physical-network arc to model the set of infrastructure and operation limits. Consider, to illustrate, the case of services of different types and priorities (e.g., passenger and freight trains) operating on the same infrastructure. The travel time of each service type is then given by a multivaried function of the total number of services of each type operating simultaneously on the arc (e.g. Crainic et al., 1984).

Notice, finally, that the geographical, topological, and infrastructure-quality characteristics of the regions concerned by the transportation system under study may have an impact on the network characteristics, particularly in terms of capacity and travel time. Thus, for example, speed and maximum weight generally decrease in mountainous terrain (unless additional power units are assigned to services) or on sinuous physical paths. Badly maintained roads, rail tracks, and bridges also significantly impact the performance of the services using them. Such impacts may be modeled at the level of the physical or service arcs. In the former case, one may define adjustment coefficients, possibly by service type, to reduce the speed and capacity of the services using the arc. In the latter case, the loss in speed and capacity has to be computed and applied to each service leg in the potential service network.

Climate has also a strong impact. The spring thawing conditions in northern regions, for example, weaken the road infrastructure, which prompted authorities to reduce the maximum

permitted truck weight during those periods. Climate change brings additional concerns. Raising temperatures cause the permafrost thawing in northern regions, which not only releases additional carbon but also weakens the transport infrastructure. In other world regions, increasingly long and strong droughts decrease the water levels of rivers and canals, jeopardizing plans and traffic. Enhancing SND models and methods to address these situations is a research challenge for the OR and transportation communities.

4.2 Demand

Recall that carrier tactical planning and the associated SND methodology target the organization of the supply side of the system to answer the estimated/forecast regular demand for transportation of its shipper customers. In all generality and from a carrier-shipper relation point of view, each component of the demand corresponds to a request to move a specific quantity of a given type of freight between two locations serviced by the carrier, at, possibly, given time moments. The demand considered at tactical-planning level is a significantly less detailed estimation of what will bew moved on a regular basis. A few modeling clarifications are in order.

First, *aggregation* is a core element in defining the demand for tactical planning. Customer locations are assigned to a regional terminal or hub, their associated outgoing and incoming demands being aggregated accordingly. Hence, the estimated demands of shippers with similar characteristics in terms of origin, destination, type, handling cost, and fare are aggregated into a so-called *commodity*. (Note that, very important customers, with respect to freight volumes, revenues, personal relations, etc., may be individualized as a separate commodity.) This yields the OD-specific multi-commodity demand of SND models. When some different commodities share the same origin or destination terminal, or both, one creates the appropriate number of nodes in the physical network to differentiate the terminal representation for each commodity.

Second, we refer to the *freight type* as the particular product concerned by a given demand, together with it's specific physical and transportation characteristics and requirements, including weight, size, packaging, handling, and product-vehicle adequacy rules (e.g., fresh or frozen food requiring refrigerated vehicles) The restrictions imposed by most of these characteristics may be addressed by an adequate definition of the decision variables of the SND model. Thus, for example, enforcing the modes and vehicles that may be used for a given product type may be performed by restricting the definition of the flow utilization variables x_a^k to the appropriate modal arcs. (Notice that the logistics of dangerous goods is governed by particular laws and rules, such materials not being pat of the usual consolidation processes.)

Most applications in the literature and this chapter assume that all demand types may be loaded together in the vehicles of the modes considered. This approach is particularly appropriate when freight is packaged into containers or vehicles, at a shipper or carrier location, before those loaded units are brought to the origin terminal. Demand is then measured in number of loaded units of specific types, e.g., 20- and 40-feet long containers for intermodal transportation, modular containers in City Logistics, and rail boxcars for general cargo.

The economic elements of the demand characterization reflect their impact on the carrier profit. Revenues come from the *fares* (*tariffs*) shippers pay for the transportation of their goods, which are conditioned by a combination of freight type, distance, service requirements, and commercial understandings (e.g., long-term contracts offering discounts on regularity and volume). Revenues are not included in most tactical-planning applications, which take the estimated regular demand as input. The companion chapter Crainic (2024) discusses the issue.

Transportation and terminal-related costs are included in tactical planning and SND models. Defined as *unit* costs, they reflect the carrier's estimation of the distribution to each unit of freight moved or handled (eventually, stored) in terminals of the respective service (irrespective of the fixed activation cost) or terminal-operation cost. We discuss this issue together with the evaluations measures of the service network (Section 4.3).

A final remark with respect to demand modeling. Trade is unbalanced and, consequently, so is the demand and the needs for particular resource, vehicle, traction-unit, container, types. Hence, even though moving empty vehicles is costly and carriers aim to minimize them, one needs to balance the resource flows. These decisions have to be echoed in the tactical-planning methods, to correctly represent the problem setting and avoid significant underestimation of the resource and infrastructure utilization and costs. This is increasingly performed by including resource-management concerns into SND models (Crainic, 2024). When this is not the case, one may still account for empty flows by estimating OD "empty-vehicle" volumes and including them in the demand definition (e.g., Crainic et al., 1984).

4.3 Service Network

SND models optimize the tactical plan on a potential service network $\mathscr{G} = (\mathscr{N}, \mathscr{A})$. In all generality, all possible services connecting the terminals in $\mathscr{N}$ are included in $\mathscr{G}$. In practice, one considers the set of services operated during the last appropriate season (e.g., last summer if the plan is built for summer operations), enriched by a more or less large set of extra services that could take care of additional and modified demand compared to the previous plan.

As described previously, the service routes are defined as paths on the physical network. Services may be either direct between their origin and destination terminal nodes, or multi-leg and include additional stops at terminals along their routes. In all cases, a service is characterized by its mode, type, capacity, and economic attributes.

The service *mode* is strongly related to the mode of the infrastructure supporting its operation. The definition may be extended to include operational characteristics. One finds, for example, modes indicating the size class of the service, e.g., large versus medium-size container-liner ships, the targeted product class, e.g., intermodal versus general rail trains, and the organization of the service, e.g., single-trailer LTL trucks versus multi-trailer LTL road trains.

Worth noticing are the efforts to introduce more environment-friendly modes and vehicles. Many such developments occur in the urban and City Logistics, and include cargo bikes, elec-

tric and hydrogen vans, drones and robots, and more or less automated vehicles, to name but a few of the best known initiatives that found their way into practice. Research is also under way regarding inter-urban transport, but large-scale deployment is still to come. An interesting emerging research field addresses the grouping of motor vehicles into platoons for part of their journeys, with promises to reduce energy consumption, environmental negative impacts, and crew costs. The SND research to integrate these new modes is emerging and challenging (see, e.g., Scherr et al., 2022; Ammann et al., 2024).

The service *type*, or *target efficiency*, is mainly related to the duration of the service from origin to the destination, as well as to the leg-specific travel times, when relevant. One may define, for example, two services on the same physical route, one regular and one express, the latter moving freight at a higher pace and, generally, higher cost. When relevant, e.g., rail services captive of their infrastructure, the *priority* when meeting or overtaking other services is part of the service type. *Service-quality targets* are also part of the type set of attributes, indicating the level of predictability/regularity of the service in achieving the stated duration and delivery dates over the planning horizon. This is often measured as the percentage of the number of on-time travel and delivery achievements with respect to the total number performed.

The *service duration* is the total time required to reach the destination once the service leaves its origin terminal, and is the sum of the travel times of the service legs and the stopping times at intermediary terminals. The leg travel time equals the sum of the times required to cross each physical arc on its route, at the speed indicated by the service type, possibly adjusted by the infrastructure-quality coefficient of the arc.

Notice that, even though the time dimension is not explicitly accounted for in static formulations, this does not mean its impact is ignored. The implicit representations of time in tactical planning and SND models is illustrated by the duration and service-quality targets defining the service type, and the *frequency*-based SND formulations.

Frequency-based operation implements a somewhat flexible service-dispatching policy combining "leave when full as much as possible" and "dispatch another vehicle on the same service, if needed" strategies. The "same" service may thus be operated several times during the schedule length, e.g., several trucks may be dispatched during the day between the same two cities, or several ships (or the same ship performing back-and-forth services) may sail during the week or the month. This policy is captured through integer-defined service selection decision variables (constraints (15) of the general SND formulation), assuming that departures are uniformly distributed over the activity period.

A global *capacity* u_σ characterizes each service $\sigma \in \Sigma$. Detailed commodity and leg-specific capacities, u^k_σ and $u_{l^\sigma_i}$ ($u^k_{l^\sigma_i}$), $l^\sigma_i \in \mathscr{L}(\sigma)$, respectively, may be also present, when relevant to capture differences among the attributes of physical arcs or commodities. Capacity stands for the total volume of freight, possibly commodity specific, the service may load and haul on the complete physical route of the leg, given its particular type. Clearly, $u_\sigma = \min\{u_{l^\sigma_i}, l^\sigma_i \in \mathscr{L}(\sigma)\}$ (same relation for commodity-specific capacities).

The economic attributes of a service are modeled through the *fixed design* (*selection*) cost payed to set up the service, and the *utilization* (*commodity*, *flow*, and *transportation* are also

used) cost incurred to transport the demand assigned to the service. Notice that, “cost" is a generic term used to capture the monetary / economic measures associated to operating the system and executing the tactical plan and, thus, to evaluate alternatives and make choices.

In network-design vocabulary, the “fixed" term aims to indicate that those costs are to be paid if the service is selected, independently of the flow assigned to it. It captures the economics of operating each occurrence of the service. It generally accounts both for the office cost to set up the service (e.g., management, marketing, etc.) and the in-the-field cost of operations. The latter depends upon the application. It may involve, for example, the acquisition-depreciation-maintenance cost of the power units or vehicles, particularly when resources and time are not explicitly included in the formulation. Energy costs to haul the nominal service load over the route distance at the planned speed, are generally included, as are crew-related costs.

The unit arc flow term reflects the part of the total estimated service-operating cost that relates to hauling an unit of commodity. It may be general or leg and commodity specific, accounting for the length of the arc, the weight of an unit of commodity, and the particular requirements in terms of vehicle used or handling procedures.

Most SND applications and contributions include constant fixed and unit costs, e.g., $f_\sigma, \sigma \in \Sigma$, $c_a^k, k \in \mathscr{K}, a \in \mathscr{A}$, and $c_\pi^k, k \in \mathscr{K}, \pi \in \Pi$, in linear total cost functions, illustrated in objective functions (1) and (7) of the basic models.

The definition of “cost" is becoming broader, however, as shippers expect not only low fares, but also high-quality service, while the industry is increasingly sensitive to social and legal pressure regarding energy consumption and environmental impacts. Service performance measures reflecting these expectations and concerns are modeled, in most cases, by the *time* or *delay* incurred by freight and vehicles or by the respect of predefined performance targets.

These measures are then translated in monetary values The resulting time-related service and unit-commodity costs may be then either added to the unit flow costs or kept as separate terms of the objective function. Choosing one or the other approach does not impact the optimization model and results. The second one may, however, provide more flexibility in undertaking managerial studies to evaluate, e.g., the impact of modifying the relative value of time and delays relative to the cost of operations. In all cases, one obtains a *generalized SND objective functio*n that captures the trade-offs between operating costs and service quality.

In its basic form, and the one most widely encountered in the literature, stopping and working times in terminals, as well as travel times between terminals are assumed known and fixed. Hence, the generalized SND objective function is still linear. While the linear-cost formulations represent a very broad set of issues and problem settings, other cases require more elaborate models. We illustrate the case with the explicit representation of two phenomena through non-linear functions: the impact of congestion in terminals or the physical arcs, and the penalties for violating capacity restrictions or service-quality targets (e.g., Crainic et al., 1984).

As widely used in road traffic modeling and analyses, congestion functions approximate the impact of traffic conditions on the *average* time required to pass through (be processed by) án infrastructure of limited capacity. Starting generally from queuing models (engineering

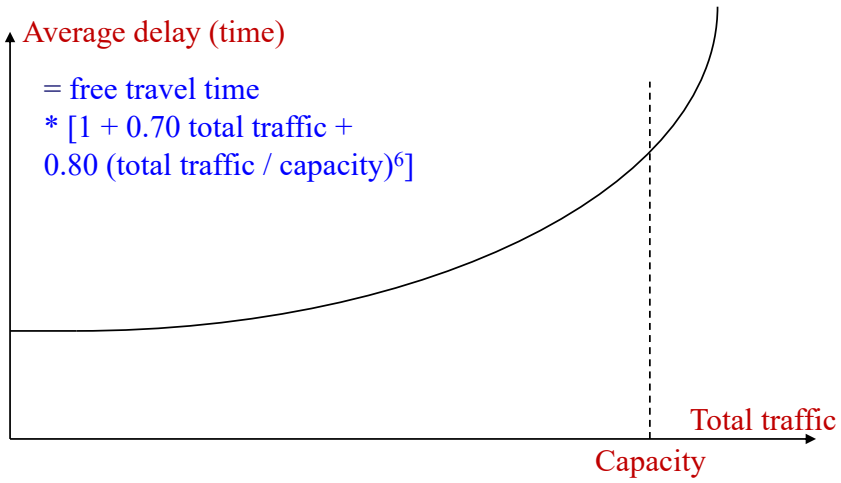

Figure 3: Average delay under congestion conditions

procedures is some cases), continuous and convex *volume-delay functions* are approximated to represent the behavior of the infrastructure under study (e.g., Crainic and Gendreau, 1986; Powell and Humblet, 1986). Figure 3 illustrates such an idealized congestion function. The traffic may be in terms of vehicles, convoys, or freight, e.g., railcars in classification terminals, ships in ports, trains on rail tracks, and load classification and consolidation in LTL terminals. Notice that, the "capacity" measure used ín such function is rarely the actual physical capacity of the corresponding infrastructure. It is rather an "ideal" limit, in terms of managerial acceptance of delays and costs, of the maximum traffic handled under normal operating conditions. One identifies three main parts to a volume-delay function. In the first part, the function reflects the fact that there is a non-compressible duration required to perform an activity and that this duration is fairly constant when the traffic volume is low. The delay than gradually increases with traffic, more or less linearly in the second part, and then non-linearly in the third. This last part, where the function increases more or less sharply with traffic, plays several roles: 1) reflects the non-linear increase of delays with traffic and models the actual behavior of the system before hitting the ideal capacity; 2) models the planning reality that, the capacity of a vehicle or facility may be exceeded at a cost, increasingly higher as the overflow grows, representing the recourse to additional capacity, own, rented, or outsourced; 3) guides the optimization model, through the sharpness of the curve near and passed the capacity, to direct flows to other paths and facilities to avoid congestion; it thus captures the managerial objective of diverting traffic to other paths before the system becomes too congested (i.e., traffic is larger than the ideal capacity).

These ideas may be refined to model many restrictions and conditions in a more realistic way than most contributions in the literature. The fundamental concept is that, while limits are hard conditions in actual operations, one does not necessarily need to model this hardness straightforwardly for an optimization formulation. One does not desire the method to crash because a capacity is attained or exceeded. One rather prefers a model that searches for more balanced alternatives, in which vehicle or freight flows are directed toward other terminals or

services. The same situation concerns the freight delivery dues dates, for which one may often deliver late paying a more or less high penalty.

Formula (17) illustrates the *penalty* modeling of capacity restrictions for the basic SND formulation (1) - (6). Let $p_{\sigma(a)}$ be the unit penalty cost for exceeding the capacity of service $\sigma \in \Sigma$ defining arc $a \in \mathscr{A}$. Then, (17) computes the total penalty cost of the system and is to be added to the objective function (1), while the linking-capacity constraints (3) are to be dropped. Trade offs between the cost of increasing the level of service and the extra costs of insufficient capacity may then be addressed while the associated mathematical programming problem is solved.

$$\sum_{a \in \mathscr{A}} p_{\sigma(a)} \left(\min \left\{ 0, u_a y_{\sigma(a)} - \sum_{k \in \mathscr{K}} x_a^k \right\} \right)^2 \tag{17}$$

We complete this section with two observations. The first concerns the still limited representation of the social/environmental impact of transportation into freight transportation planning and SND models. This is explained mainly by the difficulty to 1) estimate emissions and consumption of energy of different types by the different types of modes and vehicles, and 2) eval uate impacts in general. Not to be neglected, there is the additional challenge of building appropriate measures for the level of aggregation proper to tactical planning and SND formulation.

A relatively simple representation is proposed in the City Logistics context through a unit *nuisance* factor associated to the presence of vehicles on particular streets and plazas of the city (Crainic et al., 2009). This factor may then be added to the cost of the service. It may also be considered as a unit transportation cost and be included in the generalized objective function as a separate linear term. This linear-cost-function approach is also found in contributions targeting inter-urban transportation, where mode and vehicle type-specific energy consumption and emission factors per weight unit are associated with each arc (see,e.g., Bauer et al., 2010; Zhang et al., 2017; Truden and Hewitt, 2024).

Including linear-cost representations in SND models appears adequate for the level of aggregation and complexity of tactical planning. As emphasized in most of the limited number of contributions in this domain, the major challenges lie in the scarcity of data and the definition of adequate approximations of the very complex models of energy consumption and emissions. Notice that, even when considering the same mode, vehicle, and hauled weight, these measures are impacted in complex ways by the type of the physical arc (compare, e.g., a street of same design in the suburbs or downtown among high buildings). The impact on mode and service selection of the territorial density and distribution of recharging and refueling stations, for electric or hydrogen-powered vehicles, respectively, appears equally important and insufficiently studied. The breadth and complexity of societal impacts of transportation makes up for a timely, challenging, and exciting research field for OR.

The second observation is that averages often do not tell the full story. To illustrate, consider the case with transportation delays and the goal of planning for a consistent / reliable service. The variance of the total service or itinerary duration may then be used to penalize

unreliable operations. Define for each $k \in \mathscr{K}$ its target delivery objective H_k (e.g., 24 hours), the reliability requirement in achieving this target n_k (e.g., 90% of deliveries), and the penalty cost for not complying with this service objective p^k. Let $E^k_\pi(y,h)$ and $SD^k_\pi(y,h)$ stand for the average and standard deviation, respectively, of the duration of path $\pi \in \Pi^k$ of commodity k. Equation (18) computes the total penalty the carrier contemplates when the expected itinerary duration, adjusted for its standard deviation, does not comply with the service objective. This cost is added to the non-linear generalized objective function (14), which reflects the trade offs to be determined among the cost of operations and the fulfillment of stated service-quality levels. (Notice that, (18) assumes that arc travel times are independent, which is verified in the deterministic version; correlations must be accounted for when independence is not verified.)

$$\sum_{k\in\mathscr{K}} p^k \sum_{\pi\in\Pi^k} \left(\min\left\{0, H_k - E^k_\pi(y,h) - n_k SD^k_\pi(y,h)\right\}\right)^2 \tag{18}$$

5 Conclusions

The chapter presented an overview and synthesis of the fundamental Service Network Design problem settings and models aimed at supporting decision-making in planning the activities and managing the resources of consolidation-based freight carriers and systems. The companion chapter (Crainic, 2024) is dedicated to more Advanced topics, notably, the modeling of multiple interrelated sets of design decisions, time dependency of demand and service, and the uncertainty inherent to transportation and decision making. Challenging, but timely and important research perspectives are also discussed.

Acknowledgements

While working on the project, the author held the UQAM Chair on Intelligent Logistics and Transportation Systems Planning, and was Adjunct Professor, Department of Computer Science and Operations Research, Université de Montréal. He gratefully acknowledges the financial support provided by the Natural Sciences and Engineering Council of Canada, through its Discovery grant program, and of the Fonds de recherche du Québec through their infrastructure grants. The author is grateful to Ms. Diane Larin for her invaluable linguistic and editorial assistance, and to Professor Pirmin Fontaine, Dr Sara Khodaparasti, as well as the Editors and two anonymous referees, for their enriching comments and suggestions.

References

P. Ammann, S. Albinski, T.G. Crainic, and R. Kolisch. Joint Truck and Driver Routing and Scheduling for Semi-Autonomous Platoons. Technical Report CIRRELT, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et les transports, Université de Montréal, Montréal, QC, Canada, 2024.

J. Andersen, T.G. Crainic, and M. Christiansen. Service Network Design with Management and Coordination of Multiple Fleets. *European Journal of Operational Research*, 193(2): 377–389, 2009.

I. Bakir, A. Erera, and M.W.F. Savelsbergh. Motor Carrier Service Network Design. In T.G. Crainic, M. Gendreau, and B. Gendron, editors, *Network Design with Applications in Transportation and Logistics*, chapter 16, pages 427–467. Springer, Boston, 2021.

J. Bauer, T. Bektaş, and T.G. Crainic. Minimizing Greenhouse Gas Emissions in Intermodal Freight Transport: An Application to Rail Service Design. *Journal of the Operational Research Society*, 61:530–542, 2010.

M.E. Bruni, T.G. Crainic, and G. Perboli. Bin Packing Methodologies for Capacity Planning in Freight Transportation and Logistics. In T.G. Crainic, M. Gendreau, and A. Frangioni, editors, *Contributions to Combinatorial Optimization and Applications – A Tribute to Bernard Gendron*, chapter 6, pages 115–147. Springer, 2024.

M. Chouman and T.G. Crainic. Freight Railroad Service Network Design. In T.G. Crainic, M. Gendreau, and B. Gendron, editors, *Network Design with Applications in Transportation and Logistics*, chapter 13, pages 383–426. Springer, Boston, 2021.

M. Christiansen, E. Helsen, D. Pisinger, D. Sacramento, and C. Vilhelmsen. Liner Shipping Network Design. In T.G. Crainic, M. Gendreau, and B. Gendron, editors, *Network Design with Applications in Transportation and Logistics*, chapter 15, pages 469–505. Springer, Boston, 2021.

T.G. Crainic. Service Network Design for Consolidation-based Transportation – Advanced Topics. In S. Parragh and T. Van Woensel, editors, *Research Handbook on Transport Modeling*, chapter 4. Edgar Elgar Publishing, 2024. Publication CIRRELT-2024-31.

T.G. Crainic and M. Gendreau. Approximate Formulas for the Computation of Connection Delays under Capacity Restrictions in Rail Freight Transportation. In *Research for Tomorrow's Transport Requirements*, volume 2, pages 1142–1155. Fourth World Conference on Transport Research, Vancouver, Canada, 1986.

T.G. Crainic and M. Gendreau. Heuristics and Metaheuristics for Fixed-Charge Network Design. In T.G. Crainic, M. Gendreau, and B. Gendron, editors, *Network Design with Applications in Transportation and Logistics*, chapter 4, pages 91–138. Springer, Boston, 2021.

T.G. Crainic and B. Gendron. Exact Methods for Fixed-Charge Network Design. In T.G. Crainic, M. Gendreau, and B. Gendron, editors, *Network Design with Applications in Transportation and Logistics*, chapter 3, pages 29–89. Springer, Boston, 2021.

T.G. Crainic and M. Hewitt. Service Network Design. In T.G. Crainic, M. Gendreau, and B. Gendron, editors, *Network Design with Applications in Transportation and Logistics*, chapter 12, pages 347–382. Springer, Boston, 2021.

T.G. Crainic and W. Rei. 50 Years of Operations Research for Planning Consolidation-based Freight Transportation. Publication CIRRELT-2024, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport, Université de Montréal, Montreal, 2024.

T.G. Crainic, J.-A. Ferland, and J.-M. Rousseau. A Tactical Planning Model for Rail Freight Transportation. *Transportation Science*, 18(2):165–184, 1984.

T.G. Crainic, N. Ricciardi, and G. Storchi. Models for Evaluating and Planning City Logistics Transportation Systems. *Transportation Science*, 43(4):432–454, 2009.

T.G. Crainic, M. Gendreau, and B. Gendron. Fixed-Charge Network Design Problems. In T.G. Crainic, M. Gendreau, and B. Gendron, editors, *Network Design with Applications in Transportation and Logistics*, chapter 2, pages 15–28. Springer, Boston, 2021a.

T.G. Crainic, M. Gendreau, and B. Gendron, editors. *Network Design with Applications in Transportation and Logistics*. Springer, Boston, 2021b.

T.G. Crainic, G. Perboli, and N. Ricciardi. City Logistics. In T.G. Crainic, M. Gendreau, and B. Gendron, editors, *Network Design with Applications in Transportation and Logistics*, chapter 16, pages 507–537. Springer, Boston, 2021c.

T.L. Magnanti and R.T. Wong. Network Design and Transportation Planning: Models and Algorithms. *Transportation Science*, 18(1):1–55, 1984.

W.B. Powell and P. Humblet. Queue Length and Waiting Time Transforms for Bulk Arrival, Bulk Service Queues with a General Control Strategy. *Operations Research*, 34:267–275, 1986.

Y.O. Scherr, M. Hewitt, and D.C. Mattfeld. Stochastic Service Network Design for a Platooning Service Provider. *Transportation Research Part C: Emerging Technologies*, 144:103912, 2022.

C. Truden and M. Hewitt. The Service Network Design Problem with Fleet and Emissions Management. *Transportation Research Part C: Emerging Technologies*, 166(104769), 2024.

D. Zhang, R. He, S. Li, and Z. Wang. A Multimodal Logistics Service Network Design with Time Windows and Environmental Concerns. *PLOS ONE*, 12(9):1–19, 2017.

E. Zhu, T.G.. Crainic, and M. Gendreau. Scheduled Service Network Design for Freight Rail Transportation. *Operations Research*, 62(2):383–400, 2014.