ELSEVIER

European Journal of Operational Research 122 (2000) 272–288

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/orms

Service network design in freight transportation

Teodor Gabriel Crainic [a,b,*]

[a] *Département Management et Technologie, Université du Québec à Montréal, Montreal, QC, Canada*
[b] *Centre de Recherche sur les Transports, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, QC, Canada H3C 3J7*

Received 1 October 1998; accepted 1 April 1999

Abstract

Tactical planning of operations is comprised of a set of interrelated decisions that aim to ensure an optimal allocation and utilization of resources to achieve the economic and customer service goals of the company. Tactical planning is particularly vital for intercity freight carriers that make intensive use of consolidation operations. Railways and less-than-truckload motor carriers are typical examples of such systems. Service Network Design is increasingly used to designate the main tactical issues for this type of carriers: selection and scheduling of services, specification of terminal operations, routing of freight. The corresponding models usually take the form of network design formulations that are difficult to solve, except in the simplest of cases. The paper presents a state-of-the-art review of service network design modelling efforts and mathematical programming developments for network design. A new classification of service network design problems and formulations is also introduced.

Keywords: Transportation; Service network design; Freight transportation; Tactical planning; Modelling

1. Introduction

Freight transportation is a vital component of the economy. It supports production, trade, and consumption activities by ensuring the efficient movement and timely availability of raw materials and finished goods. In consequence, freight transportation represents a significant part of the cost of a product, as well as of the national expenditures of any country [16]. This translates in a highly competitive environment for freight transportation firms. Carriers have to rapidly adjust to changing economic and regulatory conditions, offer reliable, high quality, low cost services to their customers and, obviously, make a profit. All the planning levels and operational units of the firm have to work together, smoothly and seamlessly, toward the accomplishment of these goals.

Tactical planning of operations refers to a set of interrelated decisions that aim to ensure an optimal allocation and utilization of resources to achieve the economic and customer service goals of the company. Tactical planning is particularly relevant for intercity to firms and organizations that supply or regulate transportation services, control, at least partially, the routing of goods

* Tel.: +1-514-343-7143; fax: +1-514-343-7121.
E-mail address: theo@crt.umontreal.ca (T.G. Crainic).

0377-2217/00/$ - see front matter
PII: S0377-2217(99)00233-7

through the service network, and make intensive use of consolidation operations. Railways, Less-Than-Truckload (LTL) motor carriers, express package services, intermodal container shipping lines are typical examples of such systems. Freight transportation in some countries where a central authority more or less controls a large part of the transportation system also belongs to this category. The tactical planning process is particularly difficult in these cases due to the network-wide scale of the decisions involved, the complexity of each type of operation, and the relationships and tradeoffs between these operations, on the one hand, and the associated economic and service productivity measures and decision criteria, on the other hand. Operations research-based models and tools may assist and significantly enhance these analyses and decision making processes.

Service network design is increasingly used to designate the set of main tactical issues and decisions relevant for this type of carriers: the selection and scheduling of the services to operate, the specification of the terminal operations, and the routing of freight. The corresponding models usually take the form of *network design* formulations, a class of mixed-integer network optimization problems for which no efficient, exact solution method exists, except for special variants. Heuristics are therefore proposed in most cases. However, progress is being made. Simultaneously, one witnesses an increased interest in service network design issues, models, and efficient solution methods.

The objective of this paper is to bring together these two threads and present a joint state-of-the-art review of service network design modelling efforts and mathematical programming developments for network design. A new classification of service network design problems and formulations is also introduced. The taxonomy emphasizes the functionality of the formulation rather than the transportation mode to which it is applied. This allows to better analyze the object and scope of the various models, to identify similitudes and differences across transportation modes and their impact on formulations, and to emphasize modelling challenges.

The paper is organized as follows. Section 2 situates tactical planning within the larger context of transportation systems and planning issues. Section 3 reviews network design formulations which are often associated to tactical planning. Service network design models are reviewed in Section 4.

2. Freight transportation with consolidation

The transportation of goods is a large and complex industry. See Crainic [11] for a general presentation of freight transportation players, issues, and problem classes. The focus of the paper is on *long-haul*, *intercity transportation*, that is, on transportation operations that are mainly concerned with the movements of goods over relatively long distances, between terminals or cities. Goods may be moved by rail, truck, ship, etc., or any combination of modes.

We are particularly interested in transportation systems where service cannot be tailored for each customer individually and one vehicle or convoy usually moves freight of different customers with possibly different initial origins and final destinations. Carriers then establish regular service routes and adjust their characteristics to satisfy the expectations of the largest number of customers possible. Externally, the carrier then proposes a series of *services*, often grouped in a *schedule* that indicates departure and arrival times at the stops of the route. Internally, the carrier builds a series of rules and policies that affect the whole system and are often collected in an *operational* (also referred to as a *load* or *transportation*) *plan*. The aim is to ensure that the proposed services are performed as stated (or as close as possible) and demand is satisfied, while operating in a rational, efficient, and profitable way. Building these services and schedules is the object of tactical planning and service network design. The presence of terminals where cargo and vehicles are consolidated, grouped or simply moved from one service to another strongly characterizes this type of transportation. Such *consolidation* operations are central to these systems and an important difference to "door-to-door" transportation operations

performed, for example, by truckload motor carriers.

The underlying structure of any large consolidation transportation system consists of a rather complex network of terminals (e.g. rail yards, ports, LTL breakbulks and end-of-lines) connected by physical (e.g. rail tracks) or conceptual (e.g. sea or truck lines) links. Freight has to be moved between given points of this network. Other than its specific origin, destination, and commodity-related physical characteristics (weight, volume, etc.), each individual shipment may present any number of particular service requirements in terms of delivery conditions, type of vehicle, etc. A profit or cost also usually accompanies a specific demand. The carrier moves the freight by services performed by a large number of vehicles that move, usually on specified routes and sometimes following a schedule, either individually or grouped in convoys such as rail trains or assemblies of several trucks or barges. Convoys are formed and dismantled in terminals, while vehicles may be moved from one convoy to another. Also in terminals, freight may be loaded into and unloaded from vehicles, or it may be sorted and consolidated for the next portion of the journey.

A constant characteristic of any freight transportation system is the need to move empty vehicles. This follows from the imbalances that exist in the trade flows which result in discrepancies between the supply and demand of vehicles at the terminals of the system. To correct these differences, vehicles have to be *repositioned*, such that they are ready to answer the demand of the next period. In most cases, the decision of *how many* and *where* to send vehicles is complicated by the numerous possibilities for movement and the uncertainties of future supplies and demands. The search for the most economic strategy is a significant problem in itself [20,8,11]. At a somewhat aggregated level, *empty balancing* issues are also part of service network design.

Another notion often encountered in transportation planning has to do with *schedules* and *scheduled services.* In the general sense, a schedule specifies timing information for each possible occurrence of a service during a given time period (a week, typically): departure time at the origin, arrival/departure time at intermediary stops, and arrival time at the final destination. The schedule may also include indications on the *cut-off* time: the latest moment freight may be given to the carrier and still meet the scheduled departure of the service. Many LTL carriers operate on a "go when full" policy. Alternatively, earliest and latest departures may be planned, as well as the distribution of departures during the evening, which usually is the busiest period. The tradition in most rail systems around the world is to follow some variant of the "go when full" rule. The high volume of passenger trains already in the system, as well as the desire to decrease total transit time and improve connections, has pushed European rail companies toward more stringent schedules for their freight trains; some companies operate according to fixed schedules and bookings similar to the ones used for passenger transportation. In recent years, North American companies have also migrated toward scheduled service operations, at least for part of their traffic, with various degrees of rapidity and success.

Transportation systems are thus complex organizations which involve many human and material resources and which display intricate relationships and tradeoffs among the various decisions and management policies affecting their different components. It is then convenient to classify these policies according to the following three *planning levels*:

1. *Strategic* (long term): Such decisions determine general development policies and broadly shape the operating strategies of the system including: the design and evolution of the physical network; the acquisition of major resources (e.g. motive power units); the definition of broad service and tariff policies.
2. *Tactical* (medium term): Typical tactical decisions concern the *design of the service network* and are further discussed in the following.
3. *Operational* (short-term): Performed by local management, yard masters and dispatchers, for example, in a highly dynamic environment where the time factor plays an important role and detailed representations of vehicles, facilities and activities are essential, it includes: the implementation and adjustment of schedules

for services, crews, and maintenance activities; the routing and dispatching of vehicles and crews; the allocation of scarce resources.

This classification highlights the data and decision flows. The strategic level sets general policies and guidelines for tactical decisions that determine goals, rules, and limits for the operational decision level regulating the transportation system. The data flow follows the reverse route, each level of planning supplying information essential for decision-making processes at higher levels. This hierarchical relationships emphasizes the need for formulations that address specific problems at particular levels of decision making. It also stresses the importance of tactical planning activities and associated service network design models.

To illustrate the complexity of decisions and tradeoffs characteristic of tactical planning, consider the routing of a shipment between two terminals of an intermodal service network operated, for example, by a railway or LTL motor carrier. The shipment is sorted (classified) at the origin terminal and may be routed according to a number of strategies, including:

1. Consolidate it with other shipments going directly to its destination terminal and move it by using one of the available direct services of possibly different types.
2. Same consolidation, but move it by using a service that stops at one or several other terminals to drop and pickup traffic.
3. Consolidate the shipment into a load for an intermediate terminal where it will be reclassified and consolidated together with traffic originating at various other terminals into a load for its final destination.
4. Put it on a dedicated service, truck or direct train, if the freight volume is sufficiently high and the customer contract allows it.

Which alternative is "best"? Each has its own cost, delay, and reliability measures that follow from the service characteristics of each terminal and service. Thus, strategies based on routing through intermediate terminals and reconsolidation may be more efficient when direct services between the origin and destination terminals of the shipment are offered only rarely, due to generally low level of traffic demand, for example. This would probably result in higher equipment utilization and a decrease of waiting time at the original terminal; hence in a more rapid service for the customer. The same decision would also result, however, in additional unloading, consolidation, and loading operations, creating heavier delays and higher congestion levels at intermediary terminals, as well as a decrease in the total reliability of the shipment. On the other hand, to increase the frequency of a direct service between the origin and destination terminals of the shipment would imply a faster and more reliable service for the corresponding traffic, as well as a decrease in the level of congestion at the intermediate terminals at the expense of additional resources, thus increasing the direct costs of the system. Thus, to select the "best" solution for the customer and the company, one has to simultaneously consider the routing of all traffic, the level of service on each route, as well as the costs and service characteristics of each terminal. More formally, main decisions made at the tactical level concern the following issues:

1. *Service selection*: The routes on which services will be offered and the characteristics of each service. *Frequency* or *scheduling* decisions are part of this process.
2. *Traffic distribution*: The *itineraries* (routes) used to move the traffic of each demand: services used, terminals passed through, operations performed in these terminals.
3. *Terminal policies*: General rules that specify for each terminal the consolidation activities to perform. For rail applications, these rules would specify, for example, the blocks into which cars should be classified (the *blocking* policies), and the trains that are to be formed and the blocks that should be put on each train (the *make up* rules). An efficient allocation of work among terminals is an important policy objective.
4. General *empty balancing* strategies, indicating how to reposition empty vehicles to meet the forecast needs of the next planning period.

Several efforts have been directed toward the formulation of tactical models [1,10,21,8,11]. Network models, that take advantage of the structure of the system and integrate policies

affecting several terminal and line operations, are the most widely developed. Simulation models have been proposed and used by transportation firms to evaluate scenarios and select policies. Network optimization formulations, on the other hand, may efficiently generate, evaluate, and select integrated network-wide operating strategies, transportation plans, and schedules. These models are the object of this paper.

Most service network design and related issues yield *fixed cost*, *capacitated*, *multicommodity network design* formulations. These formulations may be static or dynamic but are always deterministic. The next section briefly reviews network design formulations and solution methods.

3. Network design

Network design models are extensively used to represent a wide range of planning and operations issues in transportation, telecommunications, logistics, and production–distribution systems. For freight transportation systems, such representations may be used, for example, to assist the decision processes concerning the construction or improvement of infrastructure and facilities, the selection of transportation services, their frequencies and schedules, as well as the allocation of human and material resources to tasks.

Network design formulations are defined on graphs containing *nodes* or *vertices*, connected by *links*. Typically, links are directed and are represented by *arcs* in a network. Some of the vertices represent *origins* of some transportation demand for one or several commodities or products, while others (possibly the same) stand for the *destinations* of this traffic. Links may have various characteristics, such as length, capacity, and cost. In particular, *fixed costs* may be associated to some or all links, indicating that as soon as one chooses to use that particular arc, one incurs the fixed cost, in excess of the utilization cost which is in most cases related to the volume of traffic on the link. The objective is to select links in a network, along with capacities, eventually, in order to satisfy the demand for transportation at the lowest possible system cost computed as the total fixed cost of the selected links, plus the total variable cost of using the network.

In the following, we present a general formulation and a few extensions that are used in the context of freight transportation planning. We also indicate the main algorithmic approaches. We do not attempt, however, to present an exhaustive state-of-the-art survey of network design formulations and solution methods. This is a considerable endeavour by itself and clearly well beyond the scope of this paper. The interested reader should consult the surveys by Magnanti and Wong [46] and Minoux [47], and the annotated bibliography of Balakrishnan et al. [2].

Consider the graph $\mathscr{G} = (\mathscr{N}, \mathscr{A})$, where $\mathscr{N}$ is a vertex set and $\mathscr{A}$ is a link set. $\mathscr{P}$ represents the set of commodities to be transported through the network. A *fixed cost network design* formulation may then take the following form:

$$\text{Minimize} \quad \sum_{(ij)\in\mathscr{A}} f_{ij}y_{ij} + \sum_{(ij)\in\mathscr{A}}\sum_{p\in\mathscr{P}} c_{ij}^{p}x_{ij}^{p}, \tag{1}$$

subject to

$$\sum_{j\in\mathscr{N}} x_{ij}^{p} - \sum_{j\in\mathscr{N}} x_{ji}^{p} = d_i^{p}, \quad i \in \mathscr{N}, \quad p \in \mathscr{P}, \tag{2}$$

$$\sum_{p\in\mathscr{P}} x_{ij}^{p} \leqslant u_{ij}y_{ij}, \quad (i,j) \in \mathscr{A}, \tag{3}$$

$$(y,x) \in \mathscr{S}, \quad (i,j) \in \mathscr{A}, \ p \in \mathscr{P}, \tag{4}$$

$$y \in \mathscr{Y}, \quad (i,j) \in \mathscr{A}, \tag{5}$$

$$x_{ij}^{p} \geqslant 0, \quad (i,j) \in \mathscr{A}, \ p \in \mathscr{P}, \tag{6}$$

where y_{ij} are integer variables modelling discrete choice design decisions. When $\mathscr{Y} = \{0,1\}^{|\mathscr{A}|}$, $y_{ij} = 1$ only if link (i,j) is *open*, i.e. selected for inclusion in the final network or for capacity expansion; the link is *closed* when $y_{ij} = 0$. When $\mathscr{Y} = \mathbf{N}_{+}^{|\mathscr{A}|}$, the y_{ij} variables are not restricted to $\{0,1\}$ values and usually represent the number of facilities or units of capacity installed, or the level of service offered; x_{ij}^{p} are continuous flow decision variables indicating the amount of flow of commodity p using link (i,j); f_{ij} the fixed cost of opening link (i,j) and when $\mathscr{Y} = \mathbf{N}_{+}^{|\mathscr{A}|}$, the hypothesis is that a f_{ij} cost is incurred for each unit of facility installed or service offered; c_{ij}^{p} the transportation cost per unit of flow of product p on link

(i,j); u_{ij} the capacity of link (i,j); d_i^p the demand of product p at node i.

This is the *linear cost*, *capacitated*, *multicommodity* (*MCND*) version of the formulation. Most applications and methodological developments target the formulations where the design variables are restricted to 0 or 1 values. A number of important applications require nonlinear formulations (see Section 4.1), while others require that flow variables be restricted to integer values, increasing the difficulty to solve these problems. Most methodological developments, however, have been dedicated to fixed cost linear formulations with continuous flow variables similar to formulation (1)–(6).

The objective function (1) measures the total cost of the system. An interesting point of view, characteristic of service network design formulations, is to consider this objective as also capturing the tradeoffs between the costs of offering the transportation infrastructure or services and those of operating the system to flow the traffic. Eq. (2) expresses the usual flow conservation and *demand satisfaction* requirements. Several demand patterns may be defined, resulting in different models. In some cases, a product may be shipped from (one or) several origins to satisfy the demand of (one or) several destinations. These are models where the supply from several origins may be substituted to satisfy a given demand and are often used in the study of the distribution of raw materials. Variants with single product origin (or destination) may also be encountered. In most transportation applications, however, each demand is defined between pairs of origin and destination points. In this case, and irrespective of the number of true commodities, a product may be associated with each origin–destination pair (by a modification of the graph that makes multiple copies of the nodes where several products originate or terminate their journeys). Let w^p be the total demand of product p. Then,

$$d_i^p = \begin{cases} w^p & \text{if vertex } i \text{ is the origin of commodity } p, \\ -w^p & \text{if } i \text{ is the destination of commodity } p, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Constraint (3), often identified as a *bundle* or *forcing* constraint, states that the total flow on link (i,j) cannot exceed its capacity u_{ij} if the link is chosen in the design of the network (i.e. $y_{ij} = 1$) and must be 0 if (i,j) is not part of the selected network (i.e. $y_{ij} = 0$). When the capacity is so large that it is never binding (i.e. u_{ij} is at least the largest possible flow on the link $\sum_p w^p$), demands may be normalized to 1 and u_{ij} may be set to $|\mathscr{P}|$. This corresponds to the *uncapacitated* formulation.

Relation (4) captures additional constraints related to the design of the network or relationships among the flow variables. Together, they may be used to model a wide variety of practical situations, and this is what makes network design problems so interesting. An important type of additional constraint reflects the usually limited nature of available financial resources:

$$\sum_{(i,j)\in\mathscr{A}} f_{ij} y_{ij} \leqslant B. \tag{8}$$

These *budget* constraints illustrate a relatively general class of restrictions imposed upon resources shared by several (or all) links. Note that, quite often, budget constraints replace the fixed cost term in the objective function. *Partial capacity* constraints also belong to this group:

$$x_{ij}^p \leqslant u_{ij}^p, \qquad (i,j) \in \mathscr{A}, \quad p \in \mathscr{P}, \tag{9}$$

and reflect restrictions imposed on the use of some facilities by individual commodities. Such conditions may be used to model, for example, the maximum quantity of some hazardous goods moved by a train or a ship. Note that including constraints (9) yields a *tighter* formulation, where the feasible domain of the decision variables is more restricted compared to the initial formulation, without loosing the optimal solution. Tighter formulations are sought after because they usually make for more efficient solution methods.

An equivalent model is the *path*-based multicommodity capacitated network design formulation PMCND:

$$\text{Minimize} \sum_{(ij)\in\mathscr{A}} f_{ij} y_{ij} + \sum_{p\in\mathscr{P}} \sum_{l\in\mathscr{L}} k_l^p h_l^p, \tag{10}$$

subject to

$$\sum_{l\in\mathscr{L}^p} h_l^p = w^p, \quad p \in \mathscr{P}, \tag{11}$$

$$\sum_{p\in\mathscr{P}}\sum_{l\in\mathscr{L}^p} h_l^p \delta_{ij}^{lp} \leqslant u_{ij} y_{ij}, \quad (i,j) \in \mathscr{A}, \tag{12}$$

$$y_{ij} \in \mathscr{Y}, \quad (i,j) \in \mathscr{A}, \tag{13}$$

$$h_l^p \geqslant 0, \quad p \in \mathscr{P}, \quad l \in \mathscr{L}^p, \tag{14}$$

where $\mathscr{L}^p$ is the set of paths for commodity p; h_l^p the flow of commodity p on path l; $\delta_{ij}^{lp} = 1$, if arc (i,j) belongs to path $l \in \mathscr{L}^p$ for product p (0, otherwise); k_l^p the transportation cost of commodity p on path l, $k_l^p = \sum_{(ij)\in\mathscr{A}} c_{ij}^p \delta_{ij}^{lp}$; and $x_{ij}^p = \sum_{l\in\mathscr{L}^p} h_l^p \delta_{ij}^{lp}$. Constraint (4) of MCND does not appear in this formulation; it is usually addressed when paths are build. The same mechanisms may also handle some nonlinear route costs. Furthermore, the explicit consideration of path flows may open interesting algorithmic perspectives [15].

Note that for any setting of the design variables, these models yield capacitated multicommodity minimum cost network flow problems in arc and path formulations, respectively. For uncapacitated design formulations, the subproblem obtained by fixing the design variables becomes an uncapacitated multicommodity flow problem that decomposes into $|P|$ shortest path problems.

Many problem classes may be derived from these general formulations by an appropriate definition of the network $\mathscr{G}$ and, eventually, of constraints in $\mathscr{S}$ [46]. Thus, when fixed costs are associated to nodes, one obtains *location* formulations. Constraints that impose the final design to be a Hamiltonian circuit yield the *Traveling Salesman* problem. Different such constraints on the form of the final network yield the *Steiner* and the *Spanning Tree* problems [45]. The capacitated *Vehicle Routing* problem may be viewed as a special case of the capacitated spanning tree formulation. This illustrates the richness of the network design class of models and explains its wide range of applications.

The previous models are mixed-integer formulations that may be approached by any of the methodologies available for this class of problems [48,56]. A widely used methodology is to relax one or several groups of constraints in a Lagrangian fashion to obtain a simpler problem. A sequence of multiplier adjustments, by using nondifferentiable optimization techniques, subgradient or bundle, for example, and resolutions of the relaxation subproblem yields a lower bound on the optimal value of the original formulation. *Dual ascent* is another often-used approach to obtain this lower bound. In this case, the dual formulation of the linear relaxation of the problem is the starting point. Dual variables are then iteratively and systematically increased, while conforming to the complementary slackness conditions. An upper bound is then obtained as the objective value of a feasible solution heuristically derived from the solution to the relaxed problem. The lower and upper bounds are then usually integrated into an implicit enumeration scheme such as the *branch-and-bound* algorithm.

As for other mixed-integer programming formulations, the polyhedral structure of the model may be studied to derive *valid inequalities* (or *cuts*) to be added to the formulation. Briefly, the objective is to construct, or approximate, the convex hull of the mixed-integer programming formulation by adding valid inequalities. If one succeeds and the convex hull is found, the original problem may be solved by linear programming methods. The *cutting plane* method is based on this idea and proceeds via a succession of resolutions of the linear relaxation of the problem and cut generations. If the convex hull can only be approximated, the bounds may be strengthened, yielding more efficient *branch-and-bound* algorithms.

In many cases, the additional complexity introduced to account for the particularities of the application at hand and the large size of the problem instance make the exact resolution of the problem impractical. Heuristics are then used to obtain solutions of, hopefully, good quality. A number of heuristics – e.g., greedly adding or dropping arcs – aim to avoid mathematical programming techniques for the mixed-integer formulation altogether but are not very successful for capacitated models. The dual-ascent type of methods and relaxations presented above are also

often used as heuristics with interesting results. Metaheuristics, principally *Tabu Search* [29], *Simulated Annealing* [41], and *Genetic Algorithms* [30], are also increasingly being applied.

Network design formulations are generally difficult to solve, however. From a theoretical point of view, most design formulations are $\mathcal{NP}$-hard. It has also been observed that for capacitated models, linear relaxation yields a poor approximation of the mixed-integer polytope. In particular, the interplay between link capacities and fixed costs is not adequately reflected by these relaxations. Moreover, the network flow subproblems are often highly degenerate, increasingly so when the number of commodities increases. Additional algorithmic challenges stem from the very large problem dimensions characteristic of most applications. Important results have been obtained for some problem classes, uncapacitated and tree-based formulations, for example. Much work is still needed for more general problem settings, however. In the remainder of this section, we point to some of these results and research challenges. The authors mentioned at the beginning of the section, and the references indicated within, offer a more in-depth treatment of the subject.

Much effort has been dedicated to uncapacitated versions of the problem and significant results have been obtained. In particular, Balakrishnan et al. [3] present a dual ascent procedure that very quickly achieves lower bounds within 1–4% of optimality. Used in conjunction with an add–drop heuristic, the method was able to efficiently address realistically sized instances of LTL consolidation problems. The very interesting performances of the dual ascent procedure have incited the development of extensions to other design formulations and applications; see, for example, the work of Barnhart et al. [4] on railroad blocking. An exact solution method for the uncapacitated formulation has been proposed by Holmberg and Hellstrand [35]. The Lagrangian-based branch-and-bound scheme solved problems with up to 1000 design arcs and 600 commodities, and outperformed a state-of-the-art mixed-integer code with respect to problem size and computation time.

Interesting results have also been obtained for the *Network Loading* problem. In this particular version of capacitated formulations, the objective is to install (load) on each design arc an integer number of different capacitated facilities, such as different transportation services. The facility capacities are usually modular, that is, if capacities are $C_i < C_{i+1}, i = 1, \ldots, l-1$, then C_{i+1} is a multiple of C_i. Efforts have been directed mainly toward the polyhedral study of the problem in order to determine valid inequalities and facets to strengthen the formulation [43,44,2]. Berger et al. [6] present an efficient Tabu Search procedure for problems with multiple facilities where the modular restriction is relaxed and flows for each origin–destination pair must follow a single path.

Less effort has been directed towards capacitated problems on general networks which are more difficult to solve and pose considerable algorithmic challenges. The possibility to efficiently compute good bounds on the optimal value of the design problem is a prerequisite to the development of solution methods that perform on large-scale problem instances with large numbers of commodities. Lagrangian relaxation approaches have recently been shown appropriate to address this issue [26,27,36,28]. Several Lagrangian relaxations are possible, however, and many offer the same theoretical bound, which is also the bound one obtains from the strong linear relaxation of the formulation [26]. Yet, the quality of the solution one may actually attain, as well as the computing efficiency and the convergence properties of the bounding procedures are strongly dependent upon the choice of the nondifferentiable optimization technique used to solve the Lagrangian duals. It also strongly depends upon the implementation and calibration of these methods. Crainic et al. [13] calibrate and compare subgradient and bundle-based methods for various relaxations of the MCND, and show, in particular, that the latter converge faster toward the optimal value of the Lagrangian dual and are more robust relative to parameter calibration.

The lower bounds obtained in [13] were reported within 9% of the optimum on average. Feasible solutions were deduced by using

resource-based decomposition methods and yielded significantly poorer bounds. Crainic et al. [15] propose a tabu search metaheuristic for the PMCND that combines simplex pivot moves and column generation. Extensive experiments, on the same set of problems used in [13], have shown that the tabu search method yields high quality solutions, dramatically improving the solutions found by the resource decomposition method. The average optimality gap was of the order of 4.5% when optimal solutions were known. The method also permitted to find solutions to problems too hard for the standard branch-and-bound in terms of CPU time or memory limitations. A cooperative parallel implementation further improved this performance [14].

Very few, if any, polyhedral results exist for the general MCND. When actually used, inequalities derived for "simpler" formulations are adapted to the more general model. Thus, *cutset* inequalities, derived for the network loading problem and stating that *the total capacity of any cut must support the total demand with end points in the two sides of the cut*, have been used to design service networks for express package delivery firms [40]. These inequalities are certainly valid for the formulation. We do not know, however, if they define facets or how efficient they are. We certainly do not know how to generate them efficiently, which is a significant issue given their extremely high number. The situation and needs are similar concerning methods to identify the optimal solution of general MCND formulations. Holmberg and Yuan [36] propose a branch-and-bound algorithm based on the Lagrangian relaxation of the flow constraints and subgradient optimization. The results appear promising, but not conclusive, especially when the dimensions of the network and the number of commodities increase. For a larger problem, Kim, Barnhart, and Ware [40,39] apply heuristics to reduce the size of the problem, followed by branch-and-bound with column and constraint generation (the so-called *branch-and-price-and-cut* – cuts are added to the root problem only). This constitutes a very interesting overture to a promising algorithmic avenue, but we are still far from a general method to optimally solve the MCND.

Parallel computation may help solve realistically dimensioned problem instances in reasonable times. Parallelism may be applied to solve the subproblem at each node of the tree [25] and to explore the tree in parallel [24]. Many issues still remain to be addressed in this area, however. For example, the addition of cuts often destroys the "nice" structures (network, knapsack, etc.) obtained by relaxing some constraints. The relaxation of the cut constraints could then be contemplated. The issue might become even more challenging when constraints are to be generated at nodes other than the root. It is generally believed, however, that the combination of relaxations, polyhedral results, and heuristics within a parallel computation framework constitutes a promising avenue towards a comprehensive solver for capacitated, multicommodity network design.

4. Service network design

Service network design formulations are typically developed to assist the tactical planning of operations, although this planning level may be referred to as strategic/tactical or tactical/operational according to the planning traditions and horizons of the firm. The goal of such formulations is to plan services and operations to answer demand and ensure the profitability of the firm. The "supply" side of this equation implies a system-wide, network view of operations that simultaneously addresses various operations performed in different facilities and globally reconciliates their often conflicting objectives and requirements. On the "demand" side, the routing of freight through the network has to be planned to ensure timely and reliable delivery according to the customer specifications and the carrier's own targets.

The objectives of service network design formulations are complex as well. The customers' expectations have traditionally been expressed in terms of "getting there" at the lowest cost possible. This, combined with the usual cost consciousness of any firm, has implied that the primary objective of a freight carrier was, and still is for many car-

riers, to operate at the lowest possible cost. Increasingly, however, customers not only expect low tariffs, but also require high quality service, mostly in terms of speed, flexibility, and reliability. The significant increase in the market share achieved by motor carriers, mainly at the expense of railway transportation, is due to a large extent to this phenomenon. Consequently, how to achieve the best tradeoffs between operating costs (firm profitability) and service performance (measured in most cases by delays incurred by freight and rolling-stock, or by the respect of predefined performance targets) constitutes one of the major objectives of tactical planning and have to be reflected in the objective functions of the associated models.

For a clearer view of tactical planning issues and service network design formulations, we distinguish between *frequency* and *dynamic* service network design models. The former typically addresses strategic/tactical planning issues. The study and representation of interactions and tradeoffs among subsystems and decisions form a central part of this class of approaches. Typical issues addressed by such models concern questions such as: *What* type of service to offer? *How often* over the planning horizon to offer it? What traffic *itineraries* to operate? What are the appropriate *terminal workloads* and *policies*? Frequency service network design models may be further classified according to the role service levels play in the formulations: *decision* or *output*. In a nutshell, service frequencies are explicit integer decision variables in the first class of models. Formulations that belong to the second class include "operate or not" $(0,1)$ variables and derive frequencies from traffic flows subject to lower bound restrictions that represent minimum service levels. The output of frequency service network design models, the *transportation* or *load plan*, is used to determine the day-to-day policies that guide the operations of the system and is also a privileged evaluation tool for "what-if" questions raised during scenario analysis in strategic planning. Dynamic formulations are closer to the operational side of things. They usually target the planning of *schedules* and support decisions related to *If* and *When* services depart.

4.1. Service frequencies as decision variables

The network optimization modelling framework proposed by Crainic and Rousseau [18] constitutes a prototypical frequency service network design formulation where explicit decision variables are used to determine how often each selected service will be run during the planning period. It is a multimode, multicommodity model that integrates the service selection and traffic distribution problems with general terminal and blocking policies. Its goal is the generation of global strategies to improve the cost and service performance of the system. It is a modelling framework in the sense that while it may represent a large variety of real situations, it has to be adapted to each application. Rail applications are to be found in [9,12,17], while [19,55,54] present applications of this framework to LTL trucking. The model has also been used to study the logistic structure – the selection of classification terminals and their interconnections – of the express parcel service of Canada Post. In the following, we present a simplified model in order to emphasize the main modelling issues and challenges.

Let $\mathscr{G} = (\mathscr{N}, \mathscr{A})$ represent the "physical network" over which the carrier operates. Vertices in $\mathscr{T} \subseteq \mathscr{N}$ correspond to nodes where the terminals selected for the particular application are situated. For simplicity, assume that all terminals can perform all operations. The *service network* specifies the transportation services that could be offered to satisfy this demand. Each *service* $s \in \mathscr{S}$ is defined by its route r_s through the physical network and a number of service characteristics:

- $\mathscr{T}_s$: terminal set comprising the origin, destination, and intermediary terminals where the service stops and work may be performed on its vehicles and cargo; $\mathscr{T}_s \subseteq \mathscr{T}$;
- Π_s: set of *service legs* of service s; a service leg π_{sk} corresponds to a subpath in r_s between two consecutive terminals in $\mathscr{T}_s$; the service network may therefore be represented as $(\mathscr{T}, \{\Pi_s\})$;
- θ_s: *service class* that indicates characteristics such as the mode, preferred traffic or restrictions, speed and priority of the service, etc.;
- u_{sk}: capacity of the service on its service leg $\pi_{sk} \in \Pi_s$; $u_{sk} = min_{(i,j)\in\pi_{sk}}\{u_s, u_{ij}^{\theta_s}\}$, where u_s

indicates the theoretical capacity of service s and $u_{ij}^{\theta_s}$ denotes the maximum load a service of type θ_s may haul over link (i,j).

The *transportation demand* is defined in terms of volume (e.g., number of vehicles) of a certain commodity to be moved between two terminals in $\mathscr{T}$. To simplify, we refer to product $p=$(*commodity type*, *origin*, *destination*) with a positive demand w^p; in the literature one also finds the terms *market* and *traffic-class* with a similar meaning. Empty vehicles may be included as commodities to be moved between given origin–destination pairs. Traffic moves according to *itineraries*. An itinerary $l \in \mathscr{L}^p$ for product p specifies the physical route r_{lp} and the service path used to move (part of) the corresponding demand:

- $\mathscr{T}_{lp}$: terminal set comprising the origin, destination, and intermediary terminals where operations are to be performed;
- S_{lp}: service route – sequence of services in $\mathscr{S}$ (one service between each pair of consecutive terminals in T_{lp}).

Service frequencies $y_s, s \in \mathscr{S}$, define the level of service offered: how often each service is run during the planning period. To design the service network thus means to decide the frequency of each service contemplated in the planning process such that the demand is satisfied. Many itineraries may be defined for each product and more than one may be actually used, according to the level of congestion in the system and the service and cost criteria of the particular application. Flow distribution decision are therefore represented by variables h_l^p indicating the volume of product $p \in \mathscr{P}$ moved by using its itinerary $l \in \mathscr{L}^p$. Workloads and general consolidation strategies for each terminal in the system may be derived from these decision variables.

Let $y = \{y_s\}$ and $h = \{h_l^p\}$ be the vectors containing the decision variables. The model states that the total generalized system cost has to be minimized, while satisfying the demand for tranportation and the service standards:

$$\text{Minimize} \sum_{s\in\mathscr{S}} \Psi_s(y) + \sum_{p\in\mathscr{P}} \sum_{l\in\mathscr{L}^p} \Phi_l^p(y,h) + \Theta(y,h) \tag{15}$$

subject to

$$\sum_{l\in\mathscr{L}^p} h_l^p = w^p, \quad p \in \mathscr{P}, \tag{16}$$

$$y_s \geqslant 0 \text{ and integer}, \quad s \in \mathscr{S}, \tag{17}$$

$$h_l^p \geqslant 0, \quad l \in \mathscr{L}, \ p \in \mathscr{P}, \tag{18}$$

where $\Psi_s(y)$ is the total cost of operating service s; $\Phi_l^p(y,h)$ the total cost of moving the freight of product p by using its itinerary l; $\Theta(y,h)$ the penalty terms capturing various relations and restrictions, such as the limited service capacity.

This model is similar to the PMCND formulation introduced in Section 3, except that the linear cost functions have been replaced by a notation that indicates more general functional forms. The objective function includes the total cost of operating a given service network at level y, the total cost of moving freight by using the selected itineraries for each product, as well as a number of terms translating operational and service restrictions into monetary values. $\Psi_s(y)$ and $\Phi_l^p(y,h)$ thus correspond to the *fixed* and *variable* costs, respectively, of the network formulation given the general level of service of the firm and the corresponding traffic pattern. The objective function computes a *generalized* cost, in the sense that it may include various productivity measures related to terminal and transportation operations. Other than the actual costs of performing the operations, one may thus explicitly consider the costs, delays, and other performance measures related to the quality and reliability of the service offered, to evaluate alternatives and determine the most advantageous tradeoffs.

The delays incurred by vehicles, convoys, and freight due to congestion and operational policies in terminals and on the road are generally used as a measure of service quality. Define $T_s(y)$ and $T_l^p(y,h)$ as the total durations of service s and itinerary l for product p, respectively. Eqs. (19) and (20) illustrate one approach to use delays to integrate service considerations into the total generalized system cost. C_s^O and C_{lp}^O stand for unit operating costs for each service and product itinerary, respectively. The corresponding total expected service, $E[T_s(y)]$, and itinerary, $E[T_l^p(y,h)]$,

times are then converted into measures compatible with the operating costs via unit time costs for each traffic (C^D_{lp}) and service (C^D_s) class. These costs are usually based on equipment depreciation values, product inventory costs, and time-related characteristics, such as priority or different degrees of time sensitivity for specific traffic classes.

$$\Psi_s(y) = (C^O_s + C^D_s E[T_s(y)])y_s, \tag{19}$$

$$\Phi^p_l(y,h) = C^O_{lp} + C^D_{lp} E[T^p_l(y,h)])h^p_l, \tag{20}$$

$$\Phi^p_l(y,h) = (C^O_{lp}h^p_l + C^D_{lp}(\min\{0, H_p - E[T^p_l(y,h)] - n\sigma[T^p_l(y,h)]\})^2 h^p_l. \tag{21}$$

Average transportation delays do not tell the whole story, however. Often, the goal is not only rapid delivery but also consistent, reliable service. The variance of the total service or itinerary time may then be used to penalize unreliable operations. Eq. (21) illustrates this approach for the case when service quality targets are announced. Here, each traffic-class has a delivery objective (e.g. 24 hours) and reliability requirements (e.g. target must be achieved for 90% of deliveries), noted H_p and n, respectively. A penalty C^D_{lp} is then imposed when the expected itinerary time, adjusted for its standard deviation $\sigma[T^p_l(y,h)]$, does not comply with the service objective.

Although nonlinear functions could be used, unit operation costs C^O_s and C^D_s are usually computed as the sum of the unit costs of all terminal and transportation activities described in the service routes and freight itineraries. The expected total delays $E[T_s(y)]$ and $E[T^p_l(y,h)]$ are also computed by summing up the expected delays associated with these operations. Although some durations are simply assumed proportional to the volume of vehicle or traffic handled, most time-related functions are built to reflect the increasingly larger delays that result when facilities of limited capacity must serve a growing volume of traffic. Such *congestion* functions are typically derived from engineering procedures and queuing models and are built to represent average classification and consolidation delays in terminals, mean delays incurred by trains when meeting, overtaking, or being overtaken by other trains on the lines of the network, expected departure or connection delays in rail yards, LTL terminals, and maritime ports representing the waiting time for the designated service to be available, and so on.

Finally, Eq. (22) illustrates the use of penalty terms to capture various additional restrictions and conditions. Here, x_{sk} stands for the total volume of freight hauled by service s over its service leg k, $x_{sk} = \sum_{p\in\mathcal{P}} \sum_{l\in\mathcal{L}^p} h^p_l \delta^{lp}_{sk}$, where $\delta^{lp}_{sk} = 1$ if service leg k of service s is used by itinerary l of product p, and 0 otherwise. The service capacity restrictions are then considered as utilization targets and over-assignment of traffic is permitted at the expense of additional costs and delays. Thus, tradeoffs between the cost of increasing the level of service and the extra costs of insufficient capacity may be addressed while the associated mathematical programming problem is solved.

$$\Theta(y,h) = \sum_{s\in\mathcal{S}} C^p_s \sum_{k\in\Pi_s} (\min\{0,\ u_{sk}y_s - x_{sk}\})^2. \tag{22}$$

The model has the structure of a nonlinear, mixed integer, multimodal, multicommodity network flow problem. No exact solution method has yet been proposed for this model. The original method described by Crainic and Rousseau combines a heuristic (based on finite differences in the objective function) that iteratively decreases frequencies from initial high values, and a convex network optimization procedure to distribute the freight. The latter makes use of column generation to create itineraries and descent procedures to optimize the flow distribution. The procedure appeared efficient for the rail and LTL applications considered.

4.2. Service frequencies as derived output

The load planning model for LTL motor carriers introduced by Powell and Sheffi [52,50,53,42] constitutes a major example of frequency service network design formulations that yield service levels as one of their outputs. Here follows a condensed version of this model.

The model is defined on a service network $\mathscr{G} = (\mathscr{T}, \mathscr{S})$ where all nodes are terminals and links represent potential direct services between two terminals. Two types of terminals are considered: *end-of-lines* where freight originates and terminates, and *breakbulk* consolidation terminals. Although not forbidden, direct movements between end-of-line terminals is extremely rare, especially for very large LTL carriers. Consequently, the design decisions concern only services between end-of-lines and breakbulks and between breakbulk terminals. This has the benefit of considerably reducing the size of the problem. The main parameters and decision variables that define the model are:

C_{ij}: unit linehaul cost per trailer, loaded or empty, from terminal i to terminal j;

C_i^B: unit trailer handling cost at terminal i, if terminal i is a breakbulk (0, otherwise);

$C_i^E(\cdot)$: a function that computes the trailer handling cost at end-of-line i according to the total number of direct services operated out of i (0, if i is a breakbulk);

w_{od}: number of LTL trailers originating at terminal o and destined for terminal d;

$\hat{w}_{od}$: volume of truckload (TL) flow originating at terminal o and destined for terminal d; although truckload traffic is not consolidated, it generates empty trailers that have to be repositioned together with those generated by the LTL traffic, and it thus included in the formulation;

$\mathscr{L}$: set of permissible freight routings, i.e. that respect particular constraints with respect to the association of end-of-line terminals to breakbulks (the so-called *clustering constraints*);

r_{ij}^d: auxiliary flow variable (its use simplifies the representation of the clustering constraints);

y_{ij}: service design decisions; $y_{ij} = 1$ if the carrier is offering direct service from terminal i to terminal j, and 0 otherwise;

x_{ij}^d: volume of LTL traffic on link (i, j) with destination terminal d; $x_{ij} = \sum_d x_{ij}^d$;

$\hat{x}_{ij}^d$: volume of truckload freight on link (i, j) with destination terminal d; $\hat{x}_{ij} = \sum_d \hat{x}_{ij}^d$;

v_{ij}: flow of empty trailers moving from i to j;

x_i^B: volume of total LTL traffic handled at breakbulk i, that is, the traffic that originates at i plus the traffic that is transferred at the terminal;

M_{ij}: minimum frequency if a direct service is offered from terminal i to terminal j;

$F_{ij}(x_{ij})$: service frequency – the number of trailers dispatched over the planning period, from terminal i to terminal j, where

$$F_{ij}(x_{ij}) = \begin{cases} \max\{M_{ij},\ x_{ij}\} & \text{if } x_{ij} \geqslant 0, \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

The model may be written as:

$$\text{Minimize} \sum_{(ij)\in\mathscr{S}} C_{ij}[F_{ij}(x_{ij})y_{ij} + v_{ij}] + \sum_{i\in\mathscr{T}} [C_i^B x_i^B + C_i^E(y)w_i], \tag{24}$$

subject to

$$\sum_{j\in\mathscr{T}} r_{ij}^d = 1, \qquad i, d \in \mathscr{T}, \tag{25}$$

$$\{r_{ij}^d\} \in \mathscr{L}, \tag{26}$$

$$r_{ij}^d \leqslant y_{ij}, \qquad i, j, d \in \mathscr{T}, \tag{27}$$

$$x_{ij}^d = \left[w_{id} + \sum_{k\in\mathscr{T}} x_{ki}^d\right] r_{ij}^d, \qquad i, j, d \in \mathscr{T}, \tag{28}$$

$$\sum_{j\in\mathscr{T}} v_{ij} - \sum_{k\in\mathscr{T}} v_{ki} = w_i, \qquad i \in \mathscr{T}, \tag{29}$$

$$w_i = \sum_{k\in\mathscr{T}} F_{ki}(x_{ki}) + \sum_{o\in\mathscr{T}} \hat{w}_{oi} - \sum_{j\in\mathscr{T}} F_{ij}(x_{ij}) - \sum_{d\in\mathscr{T}} \hat{w}_{id}, \quad i \in \mathscr{T}, \tag{30}$$

$$\sum_{j\in\mathscr{T}} \hat{x}_{ij}^d - \sum_{k\in\mathscr{T}} \hat{x}_{ki}^d = \begin{cases} \hat{w}_{id}, & i \neq d, \\ -\sum_{o\in\mathscr{L}} \hat{w}_{oi}, & i = d, \end{cases} \tag{31}$$

$$y_{ij} \in \{0, 1\}, \qquad (i, j) \in \mathscr{S}, \tag{32}$$

$$r_{ij}^d \in \{0,1\}, \qquad (i,j) \in \mathscr{S}, \tag{33}$$

$$x_{ij}^d,\ v_{ij}^d,\ \hat{x}_{ij}^d \geqslant 0, \qquad i,j,d \in \mathscr{T}. \tag{34}$$

The objective function (24) computes the total cost of dispatching trailers according to the determined service level, moving the loaded and empty trailers, and handling freight in terminals. Constraints (25), (26) and (28) ensure that freight itineraries obey routing restrictions and that demand is satisfied. Relation (27) is the usual linking constraint that ensures that only operated services are used. Eqs. (29) and (30) balance the empty flows. Constraint (31) enforces the flow conservation conditions for truckload volumes.The modelling framework is strongly influenced by the LTL context and the considerable challenges associated to the large size of the LTL carriers operating at the national level in the United States. It may be viewed as an extension of the arc-based multicommodity network design formulation ((1)–(6) of Section 3), with no explicit capacities and a number of complicating constraints. The authors implemented a heuristic procedure based on the hierarchial decomposition of the problem into a master problem and several subproblems. The master problem is a simple network design problem where the total system cost (24) is computed for each given configuration of selected services. The design is modified by adding or dropping one arc at a time. Each time the design is modified, the subproblems have to be solved and the objective function must be evaluated. The first subproblem concerns the routing of loaded LTL trailers and it is solved by shortest path-type procedures with tree constraints [51]. The empty balancing subproblem is solved as a minimum cost transshipment formulation with adjusted supply and demand to account for timing conditions not included in the original formulation.

The model and solution method have been implemented into an interactive decision support system, that has been implemented at a major LTL carrier in the United States. Impressive results are reported with respect to the impact of the system both in the context of load planning operations and for strategical studies of potential terminal locations [53]. The same modelling framework was later used as the basis for the development of a more comprehensive load planning system, implemented at one of the largest US LTL carriers [7]. In this version, the issue of running direct services, bypassing breakbulk terminals, was explicity addressed by including such services into the service network. The routing of the freight also acknowledged the geographic and labor structure of the company and considered the relay points where trailers are passed form one driver to the next. The resulting network representation is huge. Heuristics based on company operating rules are then used to prune it, before the optimization routines are called upon. Other than the optimization model and procedures, the planning system includes demand forecasting, data-base management, user monitoring and control. The system has been used with great success to build the load plan, to study the location and dimension of breakbulks, to determine the routing of loaded and empty trailers, to study which directs should be added or dropped.

Several other service network design modelling efforts make use of $\{0,1\}$, mixed integer network design formulations to address the railroad blocking problem [37,38,49,4], the design of the multimodal network of express package delivery firms [5,40,39], and the reorganization of postal services [33].

4.3. Deterministic dynamic service network design

When schedules are contemplated, a *time* dimension must be introduced into the formulation. This is usually achieved by representing the operations of the system over a certain number of *time periods* by using a *space-time* network.

The representation of the physical network is replicated in each period. Starting from its origin in a given period, a service arrives (and leaves, in the case of intermediary stops) later at other terminals. Services thus yield temporal service links between different terminals at different time periods. Temporal links that connect two representations of the same terminal at two different time periods may represent the time required by terminal activities or the freight waiting for the next

departure. The costs associated with the arcs of this network are similar to those used in the static formulations of the previous subsections. Additional arcs may be used to capture penalties for arriving too early or too late.

There are again two types of decision variables. Integer design variables are associated with each service. Restricted to $\{0,1\}$ values, these variables indicate whether or not the service leaves at the specified time. When several departures may take place in the same time period, general (nonnegative) integer variables must be used. (Note that by making the time periods appropriately small, one can always use $\{0,1\}$ variables only.) Continuous variables are used to represent the distribution of the freight flows through this service network.

The resulting formulations are network design models similar to those presented in Section 3, but on significantly larger graphs due to the time dimension. Actually, any of the two previous formulation frameworks, service network design with frequency variables or derived output, may be used as the basis for a dynamic scheduling model. The sheer size of the dynamic network, as well as the additional constraints usually required by the time dimension, makes this class of problems even harder to solve than the static ones. Heuristic methods have been used so far.

Farvolden et al. [22] present a dynamic service network design model for LTL transportation. The formulation allows for several departures in the same period, but the simpler $\{0,1\}$ version is solved. An efficient primal-partitioning with column generation algorithm [23] is used to solve the freight routing problem for any given service configuration, and to determine the dual variables for service links used by the add–drop heuristic. The approach appears interesting, especially in the evaluation of the add–drop moves. No comprehensive experimental analysis is available, however.

Haghani [34] attempts to combine the empty car distribution with the train make-up and routing problems. The dynamic network includes normal and express modes for each service route for each time period, but traffic on each link is prespecified and access to express links is restricted to given markets. Travel times are fixed. Linear functions are used for costs and delays, except for classification, which makes use of a convex congestion function. The dynamic service network design has continuous empty and loaded car flows and integer engine flows. A heuristic decomposition approach is used to solve somewhat simpler problems and appears efficient for small applications.

Gorman [31] also attempts to integrate the various service network design aspects into a scheduled operating plan that minimizes operating costs, meets the customer's service requirements, and obeys the operation rules of a particular railroad. Model simplifications are introduced, however, in order to achieve a comprehensive mathematical network design formulation. The solution method is innovative. A hybrid metaheuristic, a tabu-enhanced genetic search, is used to generate candidate train schedules, which are evaluated on their economic, service, and operational performances. On relatively small but realistic problems, the metaheuristic performed very well. A major US railroad has successfully used this model for strategic scenario analysis of their operations [32]. This work emphasizes the interesting perspectives offered by modern heuristics in addressing complex service network design problems.

5. Conclusions and perspectives

We have presented a joint state-of-the-art review of service network design modelling efforts and mathematical programming developments for network design. A new classification of service network design problems and formulations has been introduced that emphasizes the functionality of the formulation rather than the transportation mode to which it is applied. This permitted to better analyze the scope of the various models, to emphasize modelling challenges, to identify a number of important research avenues.

We hope that this side-by-side review of two important methodological and application topics will draw more researchers to network design and its applications to freight transportation planning.

Acknowledgements

Financial support for this project was provided by N.S.E.R.C. (Canada) and the Fonds F.C.A.R. (Québec).

References

[1] A.A. Assad, Models for rail transportation, Transportation Research A: Policy and Practice 14 (1980) 205–220.
[2] A. Balakrishnan, T.L. Magnanti, P. Mirchandani, Network design, in: M. Dell'Amico, F. Maffioli, S. Martello (Eds.), Annotated Bibliographies in Combinatorial Optimization, Wiley, New York, 1997, pp. 311–334.
[3] A. Balakrishnan, T.L. Magnanti, R.T. Wong, A dual-ascent procedure for large-scale uncapacitated network design, Operations Research 37 (5) (1989) 716–740.
[4] C. Barnhart, H. Jin, P.H. Vance, Railroad blocking: a network design application, Operations Research, 1997 (forthcoming).
[5] C. Barnhart, R.R. Schneur, Network design for express freight service, Operations Research 44 (6) (1996) 852–863.
[6] D. Berger, B. Gendron, J.-Y. Potvin, S. Raghavan, P. Soriano, Tabu search for a network loading problem with multiple facilities, Publication CRT-98-20, Centre de recherche sur les transports, Université de Montréal, 1998.
[7] J.W. Braklow, W.W. Graham, S.M. Hassler, K.E. Peck, W.B. Powell, Interactive optimization improves service and performance for Yellow Freight System, Interfaces 22 (1) (1992) 147–172.
[8] J.-F. Cordeau, P. Toth, D. Vigo, A survey of optimization models for train routing and scheduling, Transportation Science 32 (4) (1998) 380–404.
[9] T.G. Crainic, A comparison of two methods for tactical planning in rail freight transportation, in: J.P. Brans (Ed.), Operational Research'84, North-Holland, Amsterdam, 1984, pp. 707–720.
[10] T.G. Crainic, Rail tactical planning: Issues, models and tools, in: L. Bianco, A. LaBella (Eds.), Freight Transport Planning and Logistics, Springer, Berlin, 1988, pp. 463–509.
[11] T.G. Crainic, Long haul freight transportation, in: R.W. Hall (Ed.), Handbook of Transportation Science, Kluwer, Norwell, MA, 1999 (forthcoming).
[12] T.G. Crainic, J.-A. Ferland, J.-M. Rousseau, A tactical planning model for rail freight transportation, Transportation Science 18 (2) (1984) 165–184.
[13] T.G. Crainic, A. Frangioni, B. Gendron, Bundle-based relaxation methods for multicommodity capacitated network design, Discrete Applied Mathematics (forthcoming).
[14] T.G. Crainic, M. Gendreau, Cooperative parallel tabu search for capacitated network design, Publication CRT-98-71, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada, 1998.
[15] T.G. Crainic, M. Gendreau, J.M. Farvolden, A simplex-based tabu search method for capacitated network design, INFORMS Journal on Computing, 1999 (forthcoming).
[16] T.G. Crainic, G. Laporte, Planning models for freight transportation, European Journal of Operational Research (1997) 409–438.
[17] T.G. Crainic, M.-C. Nicolle, Planification tactique du transport ferroviaire des marchandises: quelques aspects de modélisation, in: Actes du Premier Congrès International en France de Génie Industriel, CEFI-AFCET-GGI, Paris, 1986, pp. 161–174.
[18] J.T.G. Crainic, J.-M. Rousseau, Multicommodity, multimode freight transportation: A general modeling and algorithmic framework, Transportation Research B: Methodological 20 (1986) 225–242.
[19] T.G. Crainic, J. Roy, O.R. tools for tactical freight transportation planning, European Journal of Operational Research 33 (3) (1988) 290–297.
[20] P.J. Dejax, T.G. Crainic, A review of empty flows and fleet management models in freight transportation, Transportation Science 21 (4) (1987) 227–247.
[21] L. Delorme, J. Roy, J.-M. Rousseau, Motor-carrier operation planning models: a state of the art, in: L. Bianco, A. LaBella (Eds.), Freight Transport Planning and Logistics, Springer, Berlin, 1988, pp. 510–545.
[22] J.M. Farvolden, W.B. Powell, Subgradient methods for the service network design problem, Transportation Science 28 (3) (1994) 256–272.
[23] J.M. Farvolden, W.B. Powell, I.J. Lustig, A primal partitioning solution for the arc-chain formulation of a multicommodity network flow problem, Operations Research 41 (4) (1992) 669–694.
[24] B. Gendron, T.G. Crainic, Parallel branch-and-bound algorithms: Survey and synthesis, Operations Research 42 (6) (1994) 1042–1066.
[25] B. Gendron, T.G. Crainic, Parallel Implementations of bounding procedures for multicommodity capacitated network design problems, Publication CRT-94-45, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada, 1994.
[26] B. Gendron, T.G. Crainic, Relaxations for multicommodity network design problems, Publication CRT-965, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada, 1994.
[27] B. Gendron, T.G. Crainic, Bounding procedures for multicommodity capacitated network design problems, Publication CRT-96-06, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada, 1996.
[28] B. Gendron, T.G. Crainic, A. Frangioni, Multicommodity capacitated network design, in: B. Sansó, P. Soriano (Eds.), Telecommunications Network Planning, Kluwer, Norwell, MA, 1998, pp. 1–19.
[29] F. Glover, M. Laguna, Tabu Search, Kluwer, Norwell, MA, 1997.
[30] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA, 1989.

[31] M.F. Gorman, An application of genetic and tabu searches to the freight railroad operating plan problem, Annals of Operations Research 78 (1998) 51–69.
[32] M.F. Gorman, Santa Fe Railway uses an operating-plan model to improve its service design, Interfaces 28 (4) (1998) 1–12.
[33] T. Grünert, H.-J. Sebastian, Planning models for long-haul operations of postal and express shipment companies, European Journal of Operational Research 122 (2000) this issue.
[34] A.E. Haghani, Formulation and solution of combined train routing and makeup, and empty car distribution model, Transportation Research B: Methodological 23 (6) (1989) 431–433.
[35] K. Holmberg, J. Hellstrand, Solving the uncapacitated network design problem by a Lagrangian heuristic and branch-and-bound, Operations Research 46 (2) (1998) 247–259.
[36] K. Holmberg, D. Yuan, A Lagrangian heuristic based branch-and-bound approach for the capacitated network design problem, Report LiTH-MAT-R-1996-23, Department of Mathematics, Linköping Institute of Technology, 1996.
[37] M.H. Keaton, Designing optimal railroad operating plans: Lagrangian relaxation and heuristic approaches, Transportation Research B: Methodological 23 B (6) (1989) 415–431.
[38] M.H. Keaton, Designing optimal railroad operating plans: a dual adjustment method for implementing Lagrangian relaxation, Transportation Science 26 (1992) 263–279.
[39] D. Kim, C. Barnhart, Transportation service network design: Models and algorithms, Report, Center for Transportation Studies, Massachusetts Institute of Technology, Cambridge, MA, 1997.
[40] D. Kim, C. Barnhart, K. Ware, Multimodal Express package delivery: a service network design application, Transportation Science, 1997 (forthcoming).
[41] P. Laarhoven, E.H.L. Aarts, Simulated Annealing: Theory and Applications, Reidel, Dordrecht, 1987.
[42] B.W. Lamar, Y. Sheffi, W.B. Powell, A capacity improvement lower bound for fixed charge network design problems, Operations Research 38 (4) (1990) 704–710.
[43] T.L. Magnanti, P. Mirchandani, R. Vachani, The convex hull of two core capacitated network design problems, Mathematical Programming 60 (1993) 233–250.
[44] T.L. Magnanti, P. Mirchandani, R. Vachani, Modeling and solving the two-facility capacitated network loading problem, Operations Research 43 (1995) 142–157.
[45] T.L. Magnanti, L.A. Wolsey, Optimal trees, in: M. Ball, M.L. Magnanti, C.L. Monma, G.L. Nemhauser (Eds.), Network Models, vol. 7, Handbooks in Operations Research and Management Science, North-Holland, Amsterdam, 1995, pp. 503–615.
[46] T.L. Magnanti, R.T. Wong, Network design and transportation planning: models and algorithms, Transportation Science 18 (1) (1986) 1–55.
[47] M. Minoux, Network synthesis and optimum network design problems: Models solution methods and applications, Networks 19 (1986) 313–360.
[48] G.L. Nemhauser, L.A. Wolsey, Integer and Combinatorial Optimization, Wiley, New York, 1988.
[49] H.N. Newton, C. Barnhart, P.H. Vance, Constructing railroad blocking plans to minimize handling costs, Transportation Science 32 (4) (1998) 330–345.
[50] W.B. Powell, A local improvement heuristic for the design of less-than-truckload motor carrier networks, Transportation Science 20 (4) (1986) 246–357.
[51] W.B. Powell, Y.A. Koskosidis, Shipment routing algorithms with tree constraints, Transportation Science 26 (3) (1992) 230–245.
[52] W.B. Powell, Y. Sheffi, The load-planning problem of motor carriers: problem description and a proposed solution approach, Transportation Research A: Policy and Practice 17 (6) (1983) 471–480.
[53] W.B. Powell, Y. Sheffi, Design and implementation of an interactive optimization system for the network design in the motor carrier industry, Operations Research 37 (1) (1989) 12–29.
[54] J. Roy, T.G. Crainic, Improving intercity freight routing with a tactical planning model, Interfaces 22 (3) (1992) 31–44.
[55] J. Roy, L. Delorme, NETPLAN: A network optimization model for tactical planning in the less-than-truckload motor-carrier industry, INFOR 27 (1) (1989) 22–35.
[56] H.M. Salkin, K. Mathur, Foundations of Integer Programming, North-Holland, Amsterdam, 1989.