Transportation Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Scheduling Direct and Indirect Trains and Containers in an Intermodal Setting

Alexandra M. Newman, Candace Arai Yano,

To cite this article:
Alexandra M. Newman, Candace Arai Yano, (2000) Scheduling Direct and Indirect Trains and Containers in an Intermodal Setting. Transportation Science 34(3):256-270. https://doi.org/10.1287/trsc.34.3.256.12297

Scheduling Direct and Indirect Trains and Containers in an Intermodal Setting

ALEXANDRA M. NEWMAN

Division of Economics and Business, Colorado School of Mines, Golden, Colorado 80401

CANDACE ARAI YANO

Department of Industrial Engineering and Operations Research, University of California, Berkeley, California 94720

The focus of our research is on rail transportation of intermodal containers. We address the problem of determining day-of-week schedules for both direct and indirect (via a hub) trains and allocating containers to these trains for the rail (linehaul) portion of the intermodal trip. The goal is to minimize operating costs, including a fixed charge for each train, variable transportation and handling costs for each container and yard storage costs, while meeting on-time delivery requirements. We formulate the problem as an integer program and develop a novel decomposition procedure to find near-optimal solutions. We also develop a method to provide relatively tight bounds on the objective function values. Finally, we compare our solutions to those obtained with heuristics designed to mimic current operations, and show that a savings of between 5 and 20% can be gained from using our solution procedure.

Intermodal transportation consists of combining modes, usually ship, truck, or rail to transport freight. The focus of our research is on rail transportation of intermodal containers for the long-haul portion of their journey. For distances over 500 miles, train transportation is more efficient than truck transportation, and results in savings in operating costs and labor. Because rail transport diverts some freight traffic from the roads, congestion and wear and tear on highways is partially alleviated (McKenzie, 1989). Despite recent advances in the efficiency of intermodal operations, difficulties remain. Some obstacles the railroads face result from inadequate infrastructure including a shortage of track, and the lack of a fully operational, continuous transcontinental railroad. Other difficulties arise, in part, due to basic management and information limitations, which lead to poor train routes and schedules and inadequate priority rules for sending shipments. Because of the extra delay incurred and the increased potential for mishandled containers at intermodal terminals, it is important that time and cost considerations be taken seriously for intermodal transportation to compete effectively with long-haul trucking.

To improve the scheduling and coordination of trains, we address the problem of how to schedule direct and indirect trains and which containers to send on each train for the rail (linehaul) portion of the intermodal trip to minimize operating costs while meeting on-time delivery requirements. Intermodal rail operations differ from conventional rail operations in several important respects. First, because of the high cost of container handling equipment, intermodal networks have relatively few, widely spaced terminals. For example, the Illinois Central Railroad has only nine intermodal terminals, a few of which are small. Networks with more than about two dozen major intermodal terminals are uncommon. With such a structure, economies of scale can be realized not only in container handling, but also in train movements from terminal to terminal. Transport from the customer to the nearest intermodal terminal is handled by truck or by regional or feeder railroads. Second, because of the distances between intermodal terminals, a typical container makes few stops and is transferred between trains only a few times on its journey. This eliminates the need to consider *blocks*, i.e., groups of railcars that travel as a unit for one or more seg-

Transportation Science, © 2000 INFORMS
Vol. 34, No. 3, August 2000

0041-1655/00/3403-0256 $05.00
pp. 256–270, 1526-5447 electronic ISSN

ments of their journey (to reduce train reassembly time at rail yards), which are essential in conventional rail scheduling and routing decisions. Finally, shorter delivery leadtimes are promised for intermodal freight, and, consequently, there is a greater need to schedule trains to achieve desired levels of customer service. Under conventional operations, some freight may wait while enough railcars accumulate to form a block. The first two factors reduce the number of decisions required for intermodal freight versus conventional freight, but the third factor dramatically increases the importance of careful train scheduling and routing decisions.

Most of the research on train scheduling and container routing uses average demand rates and has the goal of determining steady-state train frequencies and container allocations. KEATON (1989) considered direct and indirect train scheduling, the routing of railcars, and their grouping (or blocking) on a train. CRAINIC and ROUSSEAU (1986) developed a general framework for multimode freight transportation including the design of the network (e.g., which modes to use and what frequency of service to provide) and the traffic routing scheme through this network. MARÍN and SALMERÓN (1996) investigated the problem of determining an optimal train schedule for a rail network, and the optimal assignment of rail cars to these trains such that each train carries cars of a single service class. They used simulated annealing and tabu search to solve the problem. For a recent survey on train scheduling and related problems, see CORDEAU, TOTH, and VIGO (1998).

The literature on intermodal transportation is growing. MORLOK and SPASOVIC (1994) developed a model to reduce drayage costs without affecting the timeliness of pickups and deliveries. DIAL (1994) sought to minimize trailer-on-flatcar costs incurred by United Parcel Service by choosing whether to ship freight with a trailer owned by United Parcel Service, or to lease one from the railroad. BARNHART and RATLIFF (1993) developed a model that sought to minimize transportation costs for a set of trailer movements by truck and/or rail, considering the possibility of pairing trailers from different sources on the same flatcar.

The research most similar to ours is by NOZICK and MORLOK (1997) and GORMAN (1998a,b). Nozick and Morlok took the train schedule as given and addressed freight movement as well as equipment and locomotive repositioning decisions. Their objective was to minimize the cost of on-time delivery subject to constraints on equipment availability. The primary differences between Gorman's model and ours are that he considered yard and rail line capacity in an aggregate way, but focused on the special case of a single origin and single destination with multiple routes between them. He used a tabu-enhanced genetic search to arrive at solutions within 10% of the optimum for this case. He then applied his procedure to a problem with multiple interdependent origins and destinations. The solution provided significant improvements in cost and customer service over current policies, but the solution was not evaluated in an absolute sense.

To the best of our knowledge, this problem of simultaneously determining direct and indirect train-scheduling and container-routing decisions for multiple interdependent origins and destinations using a formal optimization approach has not been addressed in the literature. Four aspects of our problem that make it challenging are: (i) more than one train may be sent on each segment each day, (ii) both direct and indirect trains may be scheduled, (iii) containers arrive dynamically during the decision horizon, and (iv) customer orders have distinct due dates.

The remainder of the paper is organized as follows. In the following section, we present a problem statement and formulation. In Section 2, we introduce a new decomposition approach, which is based on a partitioning of the underlying network. In Section 3, we show how this decomposition approach can be modified to handle larger problem instances more effectively. In Section 4, we develop valid inequalities that allow us to obtain tighter lower bounds on the solutions. In Section 5, we discuss simple heuristics designed to mimic current practice. In Section 6, we present numerical results. A summary and directions for future research appear in Section 7.

1. PROBLEM STATEMENT AND FORMULATION

OUR RESEARCH WAS motivated by the train-scheduling and container-routing problem that we observed at the intermodal division of a major railroad. The railroad has several intermodal terminals on the west coast of the United States, a single major hub in the west-central part of the United States, and a few other intermodal terminals east of the Mississippi River. The flow of traffic eastbound is greater than it is westbound, as it is for most U.S. railroads.

Much of the eastbound rail transport capacity is dedicated to moving sea cargo for major international shipping lines, for which the transoceanic transit time is fairly predictable (approximately one week). Remaining train capacity is utilized to service smaller customers within a few hours' drive of the railroad's intermodal terminals. These customers typically use intermodal retailers to coordinate

the truck and rail movements for their goods. Intermodal retailers often reserve space on trains in advance, and then sell this space to their customers. These reservations also contribute to the predictability of demand for the railroad. Overall, demand exhibits weekly patterns due to freighter schedules, and seasonal patterns due to factors such as traditional cycles in retail demand, and agricultural and manufacturing production.

The railroad offers several speeds (or levels) of service and charges a premium for faster (promised) delivery. Trains may be sent directly from an origin intermodal terminal to the destination terminal without stopping at a hub, providing the fastest available service. Alternately, trains carrying containers bound for several destinations may be sent to a hub, where containers are consolidated by destination onto outbound trains. This consolidation activity may cause a few days' delay for transferring containers or repositioning rail cars between trains. Further delays also occur when inbound and outbound schedules are not coordinated.

Each train has a limited capacity, where the capacity is expressed in terms of number of containers in our model. We assume that containers are homogeneous in terms of their use of train capacity, which depends upon the power of the locomotives and the terrain over which the train must travel. Typically, decisions regarding locomotive capacity for each transportation segment are determined in advance on the basis of demand forecasts. We assume that the capacities of all trains on a given segment are the same, which reflects the situation in our motivating application.

Yard storage space for containers waiting to be shipped, awaiting a transfer at the hub, or waiting to be picked up, is limited at all terminals. As the number of containers in storage increases, containers are stacked higher and more densely. This increases the time required to retrieve a container and places a further burden on material-handling equipment, which may already be a bottleneck. Our model does not constrain the number of containers that can be stored at a yard, but we do assess a cost for container storage (discussed below) to deter unnecessary container inventory.

From our observations of intermodal terminal operations, the train schedules and container routing decisions do not appear to be affected strongly, if at all, by what speed of delivery has been promised, or what rate has been charged to the customer. This motivated us to investigate how to schedule trains and route containers to achieve on-time delivery at minimum cost.

We address a short-term, finite-horizon, discrete-time scheduling problem for the linehaul portion of the intermodal trip. Given container demands differentiated by origin, destination, arrival date at origin, and due date, the objective is to determine a train schedule (for both direct and indirect trains) and container-shipment plan to minimize the total cost while meeting on-time delivery requirements and adhering to train capacity restrictions. We discuss the cost elements in more detail below.

The costs incurred by the rail company for transportation on a segment consist of both a fixed-charge component, or *fixed cost*, for each train and a variable (per container) component. The fixed cost consists primarily of operators' wages and the opportunity cost of locomotive use. We assume that each train on a specific segment incurs the same fixed cost.

The variable cost per container consists of three main components: (i) transportation costs, such as fuel, oil, and track maintenance; (ii) handling costs incurred for moving containers on and off the rail cars, or for repositioning the cars at an intermediate terminal; and (iii) yard storage costs associated with holding containers in inventory at the origin or at an intermediate terminal. We assume that the variable transportation costs are constant over time and that they depend only on the route. The assumption of constant transportation costs over time is quite reasonable over a short horizon (e.g., a week or two), and the assumption that costs depend only on the rail segment is consistent with our assumption regarding homogeneity of containers in terms of their use of transport capacity.

Handling costs for moving containers or repositioning rail cars depend more heavily on the equipment used for such operations than they do upon the container itself, or its origin or destination. Thus, the assumption of constant handling costs for a given terminal is quite mild. Inventory costs consist primarily of yard storage costs and the opportunity cost of having a container unavailable for use, because the opportunity cost of capital for in-transit goods is borne either by the shipper or by the consignee. Yard storage costs in our model are assumed to be equal for all containers for all locations and time periods. We assume that customers will accept delivery upon arrival at the destination, so no inventory is held at the destinations. Generalizations to consider other linear cost structures and delivery-acceptance rules are straightforward.

We assume that hub delay and transit times are deterministic, constant across time, and that both transit times and delays at the hub are expressed as an integral number of time periods, where a time period is typically one day. In practice, transit times

are rarely predictable, but because time is expressed in days, not hours or minutes, there is implicit slack in the schedule. As with airlines and other transportation providers, additional slack in both scheduled transit times and scheduled hub delays may be built into the schedule to help ensure on-time delivery. For instance, one major railroad adjusts schedules with the goal of ensuring a 12-hour delivery window for a container with a 4-day transit time.

Finally, we assume there is no limit on the number of trains that can be sent each day, although in reality, locomotive and crew availability may be limited with respect to location and time. Our model can be generalized to handle limits on the number of trains, provided train availability is adequate to ensure on-time delivery, which we enforce as a hard constraint. In practice, terminal operators might choose to delay the shipment of some containers to avoid sending a train with a small number of containers. Our model can be generalized to allow tardy deliveries and associated penalties.

To summarize, our problem is to choose train schedules and container routes for each day over a short horizon to achieve on-time delivery at minimum cost, where the total cost consists of a fixed charge per train, a variable transportation cost per container, both of which are dependent on the rail segment, handling costs per container dependent upon the location, and inventory holding costs for containers held at any terminal prior to their arrival at the destination.

Note that our model is intended to aid in establishing schedules in the "bottleneck" direction and does not address locomotive repositioning. Empty container repositioning can be handled as part of the demand. The notation for our model is described below.

Subscripts:

i = index of origins
j = index of hubs
k = index of destinations
t = index of periods in the time horizon, $t = 1, 2, \ldots, T$
l = index of level of service, i.e., the due date of the container at the destination

Parameters:

α_{ik} = direct transportation time between origin i and destination k
β_{ij} = transportation time between origin i and hub j
γ_{jk} = transportation time between hub j and destination k
δ_j = delay time incurred from passing through hub j
C = capacity of a train (number of containers)
h = holding cost of a container ($/container/day)
c^a_{ik} = variable unit cost of transporting a container directly from i to k
c^e_{ijk} = variable unit cost of transporting a container from i via j to k
S^{ao}_{ik} = fixed cost of running a train directly between origin i and destination k
S^{eo}_{ij} = fixed cost of running a train between origin i and hub j
S^h_{jk} = fixed cost of running a train between hub j and destination k
g^o_i = cost of placing a container on the train at origin i
g^h_j = cost of rearranging a container at hub j
g^d_k = cost of removing a container from the train at destination k
b_{iktl} = the number of containers that arrive at i on day t bound for k, due at time l

Decision variables:

I^o_{iktl} = container inventory held at i at time t, which is due at k by time l
I^h_{ijktl} = container inventory from i, held at j at time t, due at k by time l
I^d_{iktl} = container inventory from i due by time l which is held at k at time t
x^{ao}_{iktl} = number of containers shipped directly from i to k at time t, due by time l
x^{eo}_{ijktl} = number of containers shipped from i to j at time t, due at k by time l
x^h_{ijktl} = number of containers from i, shipped at time t from j to k, due by time l
z^{ao}_{ikt} = the number of trains sent directly from origin i to destination k at time t
z^{eo}_{ijt} = number of trains sent from origin i to hub j at time t
z^h_{jkt} = number of trains sent from hub j to destination k at time t

The formulation follows.

(P)

$$\min Z = \sum_{iktl} hI^o_{iktl} + \sum_{ijktl} hI^h_{ijktl} + \sum_{ijkl} \sum_{w=t+\beta_{ij}}^{t+\beta_{ij}+\delta_j} hx^{eo}_{ijkwl}$$
$$+ \sum_{iktl} c^a_{ik}\, x^{ao}_{iktl} + \sum_{ijktl} c^e_{ijk}\, x^{eo}_{ijktl}$$
$$+ \sum_{iktl} g^o_i\, x^{ao}_{iktl} + \sum_{ijktl} g^o_i\, x^{eo}_{ijktl} + \sum_{ijktl} g^h_j\, x^h_{ijktl}$$
$$+ \sum_{iktl} g^d_k\, x^{ao}_{iktl} + \sum_{ijktl} g^d_k\, x^h_{ijktl} + \sum_{ikt} S^{ao}_{ik}\, z^{ao}_{ikt} + \sum_{ijt} S^{eo}_{ij}\, z^{eo}_{ijt}$$
$$+ \sum_{jkt} S^h_{jk}\, z^h_{jkt}$$

subject to

$$b_{iktl} + I^o_{ik(t-1)l} = I^o_{iktl} + x^{ao}_{iktl} + \sum_j x^{eo}_{ijktl} \quad \forall i,\ k,\ t,\ l \tag{1}$$

$$I^h_{ijk(t-1)l} + x^{eo}_{ijk(t-\beta_{ij}-\delta_j)l} = I^h_{ijktl} + x^h_{ijktl} \quad \forall i,\ j,\ k,\ l,\ t \ni t \geqslant 1 + \beta_{ij} + \delta_j \tag{2}$$

$$I^d_{ik(t-1)l} + x^{ao}_{ik(t-\alpha_{ik})l} + \sum_j x^h_{ijk(t-\gamma_{jk})l} = I^d_{iktl} + b_{iktl} \quad \forall i,\ k,\ l,\ t \ni t \geqslant 1 + \alpha_{ik} \tag{3}$$

$$\sum_l x^{ao}_{iktl} \leqslant C z^{ao}_{ikt} \quad \forall i,\ k,\ t \tag{4}$$

$$\sum_{kl} x^{eo}_{ijktl} \leqslant C z^{eo}_{ijt} \quad \forall i,\ j,\ t \tag{5}$$

$$\sum_{il} x^h_{ijktl} \leqslant C z^h_{jkt} \quad \forall j,\ k,\ t \tag{6}$$

All variables restricted to be nonnegative and integer.

where I^o_{iktl} is set equal to 0 unless $t \geqslant 1$ and $l > t + \alpha_{ik}$; I^h_{ijktl} is set equal to 0 unless $t \geqslant 1 + \beta_{ij} + \delta_j$ and $l > t + \gamma_{jk}$; and I^d_{iktl} is set equal to 0 unless $t \geqslant 1 + \alpha_{ik}$ and $l > t$. Note also that $x^{eo}_{ijktl} = 0$ if $l < t + \beta_{ij} + \gamma_{jk} + \delta_j$ and $x^h_{ijktl} = 0$ if $l < t + \gamma_{jk}$. These constraints ensure that appropriate inventory variables are initialized to zero; others are constrained to be non-negative. Similarly, indirect container shipments are constrained to be zero if due dates necessitate a direct routing.

In our analysis, we assume only one hub, although the formulation is written for the more general case in which there may be multiple hubs and each container may pass through at most one hub. We also assume that direct travel time between an origin and a destination is strictly less than the total transit and delay time for a container shipped indirectly, i.e., $\alpha_{ik} < \beta_{ij} + \gamma_{jk} + \delta_j$.

The objective function consists of inventory-holding costs at the origin and at the hub; transportation costs for directly and indirectly shipped goods; handling costs at the origin for both direct and indirect shipments, handling costs at the hub for indirect shipments, and handling costs at the destination for direct and indirect shipments; and finally, fixed costs at the origin for direct trains and trains bound for a hub, and fixed costs at the hub for indirect trains.

Constraints 1, 2, and 3 represent conservation of flow of containers at the origins, hubs, and destinations, respectively. Although we assume that no inventory is held at the destination, the variables I^d_{iktl} are used for bookkeeping purposes to ensure that demand is satisfied. To properly reflect that containers are not actually held, we set the corresponding inventory-holding cost coefficients to zero. Constraints 4 require that, for all origins, destinations, and time periods, the number of containers sent on direct trains must not exceed the total capacity of the trains operating on that segment. Likewise, constraints 5 and 6 ensure that train capacity is not exceeded on trains bound for the hub and trains leaving the hub, respectively. Finally, nonnegativity and integrality constraints are imposed on all decision variables.

A typical problem has thousands of general integer variables, and the nature of the tradeoffs contributes further to the difficulty of the problem. Direct trains are both faster and less expensive than indirect trains, but they service only one destination. Thus, one faces difficult choices such as whether to send containers on a relatively full indirect train today, or, alternately, to send them within the next few days on a direct train that may be underutilized. Indeed, even when the train schedule is fixed, the best priority scheme for assigning containers to trains is not evident. In particular, it may not be optimal to ship containers bound for the same destination in earliest due-date order.

Our problem has the structure of a multiple-fixed-cost, multicommodity network flow problem. Figure 1 depicts such a network for two origins, two destinations, and two time periods, assuming transit times and hub delays are zero for simplicity. In addition to a source and a sink, the network contains three sets of nodes, with each node representing a location–time period pair, where the location may be an origin, hub, or destination. An arc links two nodes representing different locations if a container may travel between the locations beginning and ending exactly at the time periods associated with the two nodes. Additional arcs from a location in one time period to the same location in the subsequent time period permit inventory flows from period to period. The network has a multicommodity structure because flows on the same arc may differ with respect to their origin, destination, and/or due date, depending upon the specific arc in question. Different variable costs may be associated with each commodity on each arc. A commodity is assigned an infinite cost if its due date prohibits it from traveling on a particular arc. Upper bounds on the total flow on each arc depend on the decisions regarding the number of trains between the relevant locations in a given time period, and a fixed charge is assessed for each train. Multicommodity (uncapacitated) net-

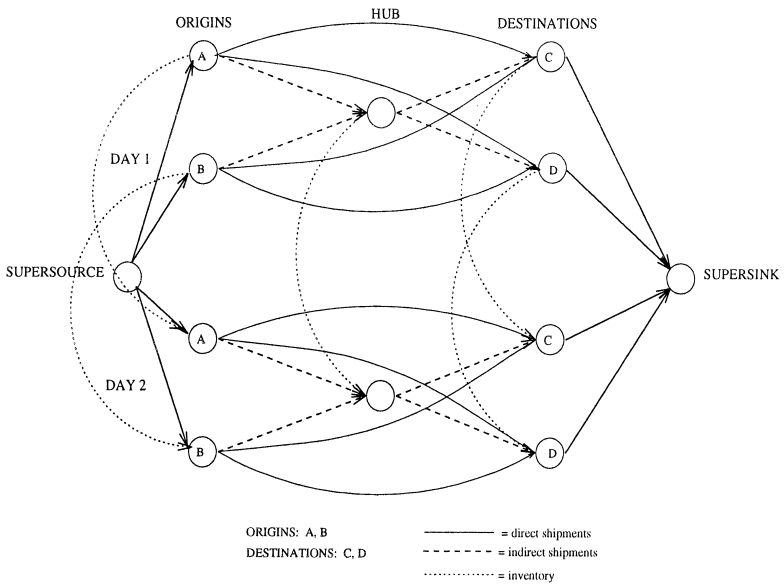

Fig. 1. Multi-commodity network depiction of our problem.

work flow problems with a fixed-charge structure are known to be NP-complete (GAREY and JOHNSON, 1979), and our problem is complicated further by the multiple fixed costs. Although algorithms have been developed for fixed-charge networks, none can provide near-optimal solutions efficiently for our problem. The problem may also be formulated as a single commodity network flow with bundle constraints whose capacities depend upon the train decisions, but the problem is no less difficult to solve in this form (see AHUJA, MAGNANTI, and ORLIN, 1993).

We attempted to develop solution procedures using Lagrangian relaxation and Benders' decomposition. Relaxing the train capacity constraints using Lagrange multipliers led to poor results because the multipliers could not accurately capture the step-function nature of the costs associated with the train variables. In our application of Benders' decomposition, the train decisions are included in the master problem, and the container flow variables appear in the subproblem. Although this decomposition appeared to be the most natural one, the dual price information from the subproblem is insufficient to aid in selecting better train decisions because of the large fixed charge associated with each train. For further details on Lagrangian relaxation and Benders' decomposition, see NEMHAUSER and WOLSEY (1988), and for further discussion of the application of these techniques to our problem, see NEWMAN (1998). Because of the difficulty of adapting traditional techniques to our problem, we developed a new decomposition that takes advantage of the physical structure of the system and the underlying network flow problems.

2. NEW DECOMPOSITION TECHNIQUE

OUR DECOMPOSITION APPROACH is motivated, in part, by the observation that, if the optimal pattern of container arrivals at the hub (by origin, destination, arrival date at origin, and due date) were known, we could infer which containers were to be sent on direct trains. Moreover, for any given pattern of container arrivals at the hub, the problem decomposes into three sets of subproblems:

P_0: scheduling direct trains and containers for each origin–destination pair;
P_1: scheduling trains and containers into the hub from each origin; and
P_2: scheduling trains and containers from the hub to each destination.

Each of these subproblems has a single origin and a single terminus (the hub or a destination), and, although they remain network flow problems with multiple fixed charges on each arc, they can be solved optimally in polynomial time under the assumptions of our model. See YANO and NEWMAN (1998) for details. For convenience, let us refer to the objective of P_0 for origin i and destination k as $Z_{P_0}(i, k)$, the objective of P_1 for origin i as $Z_{P_1}(i)$ and the objective of P_2 for destination k as $Z_{P_2}(k)$.

Let us define $B(i, k, t', t, l)$ as the number of containers that arrive at origin i in period t', arrive

at the hub at period t, and are due at destination k in time period l. Then, assuming instantaneous travel time for simplicity, the matrix, **D**, of arrivals at the origin that must be shipped on direct trains is defined by the values

$$D_{ikt'l} = b_{ikt'l} - \sum_{t=t'}^{l} B(i, k, t', t, l) \quad \forall i, k, t', l,$$

and the aggregate number of containers arriving at the hub at time t, bound for destination k and due at time l can be represented as

$$\sum_{i,t'} B(i, k, t', t, l) \quad \forall k, t, l.$$

With these definitions, problem (P) can be restated as

$$\min_{\mathbf{B}} \left\{ \min \sum_{i,k} (Z_{P_0}(i, k)|\mathbf{D}) + \min \sum_{i} (Z_{P_1}(i)|\mathbf{B}) + \min \sum_{k} (Z_{P_2}(k)|\mathbf{B}) \right\}.$$

The matrix **B** is constrained by the pattern of arrivals at the origin and the on-time delivery constraints, and directly influences the matrix **D**. The problem of finding the optimal matrix **B** is difficult, even without considering the on-time delivery constraints. Many different **B** matrices may lead to similar solutions for the individual subproblems because each non-urgent container may take one of several routes with similar, or even identical, costs. Moreover, because the total cost is the sum of the costs of many subproblems, there may be many different **B** matrices that lead to similar overall costs. For example, one solution in which a given origin sends many containers indirectly and another sends many containers directly may have a similar cost to one in which the allocation of direct and indirect shipments is reversed. Our strategy is based on the conjecture that finding "good" **B** matrices should provide the foundation for identifying a near-optimal solution.

Rather than searching for good **B** matrices directly, we solve a problem of the following form, which we term the "origin scheduling problem" to determine the direct and indirect train schedules and related container flows outbound from each origin i:

$$\min \sum_{k} Z_{P_0}(i, k) + Z_{P_1}(i), \tag{7}$$

where, in $Z_{P_1}(i)$, we assume that for each train inbound to the hub from origin i, there is a train outbound from the hub whose fixed cost is the demand-weighted average of the costs of trains from the hub to the various destinations. Also, the handling cost per container is the sum of the handling cost at the origin and at the hub. Thus, rather than using fixed and handling costs that reflect only the first transportation segment, we use adjusted costs that reflect estimates for the entire route:

$$\widetilde{S^{eo}_{ij}} = S^{eo}_{ij} + S^{h}_{j} \qquad \widetilde{g^{o}_{ij}} = g^{o}_{i} + g^{h}_{j},$$

where S^h_j is the fixed cost at hub j, obtained by using an average or a weighted average of S^h_{jk} across destinations.

Our motivation for making these cost adjustments is to incorporate the first-order effects of sending trains from the origin to the hub on the costs that are incurred after the train reaches the hub. In other words, the cost adjustment is an estimate of the "cost to go" outbound from the hub. Although the number of trains into and out of the hub may not be exactly equal within a short time horizon, in practical applications, these values are fairly well balanced. If the train capacities are well utilized inbound to the hub, on-time delivery requirements make it difficult to hold containers at the hub for long enough to achieve significant additional consolidation outbound from the hub. Here, we are implicitly assuming that the capacities of trains inbound to and outbound from the hub are the same. If train capacities vary by segment, then appropriate adjustments can be made in the cost-to-go estimates.

More formally, the origin scheduling problem for each origin i is

$$\min \sum_{ktl} hI^{o}_{iktl} + \sum_{ktl} c^{a}_{ik}\, x^{ao}_{iktl} + \sum_{jktl} c^{e}_{ijk}\, x^{eo}_{ijktl} + \sum_{ktl} g^{o}_{i}\, x^{ao}_{iktl} + \sum_{jktl} \widetilde{g^{o}_{ij}}\, x^{eo}_{ijktl} + \sum_{kt} S^{ao}_{ik}\, z^{ao}_{ikt} + \sum_{jt} \widetilde{S^{eo}_{ij}}\, z^{eo}_{ijt}$$

subject to constraints 1, 4, 5, and nonnegativity and integrality constraints on all variables.

Having solved the above problem for each origin i, we solve P_2, the "hub scheduling problem," for each destination k given container flows into the hub, i.e., x^{eo}_{ijktl} from the origin scheduling problem. Letting c^e_{jk} denote the transportation cost between hub j and destination k, the problem for destination k is

$$\min \sum_{ijtl} hI^{h}_{ijktl} + \sum_{ijtl} (g^{h}_{j} + g^{d}_{k}) x^{h}_{ijktl} + \sum_{ijtl} c^{e}_{jk} x^{h}_{ijktl} + \sum_{jt} S^{h}_{jk}\, z^{h}_{jkt},$$

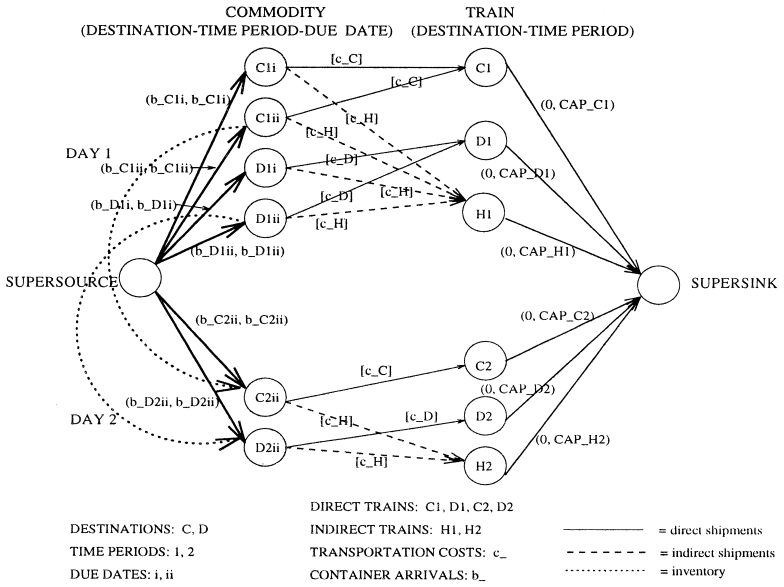

Fig. 2. Network depiction of the origin scheduling problem for a given train schedule.

subject to constraints 2, 6, and nonnegativity and integrality constraints on all variables. More complex methods based on the same general strategy appear in NEWMAN and YANO (2000).

Observe that this solution strategy allows us to solve each origin scheduling problem independently, and to solve an independent hub scheduling problem for each destination. By simultaneously considering direct and indirect shipments from each origin in constraint 7 and approximating the cost to go for the indirect shipments, we are able to find solutions that reflect tradeoffs related to the type (direct or indirect), number, and timing of trains to service goods arriving at the same origin. We expect there will be some loss of optimality from our approximation of the cost to go, but we trade this off against the loss of optimality from suboptimal solutions to the original, monolithic problem. We explore these issues further in Section 6.

Figure 2 depicts the network representation of the origin scheduling problem for a single origin, a hub, and two destinations for a given train schedule, and

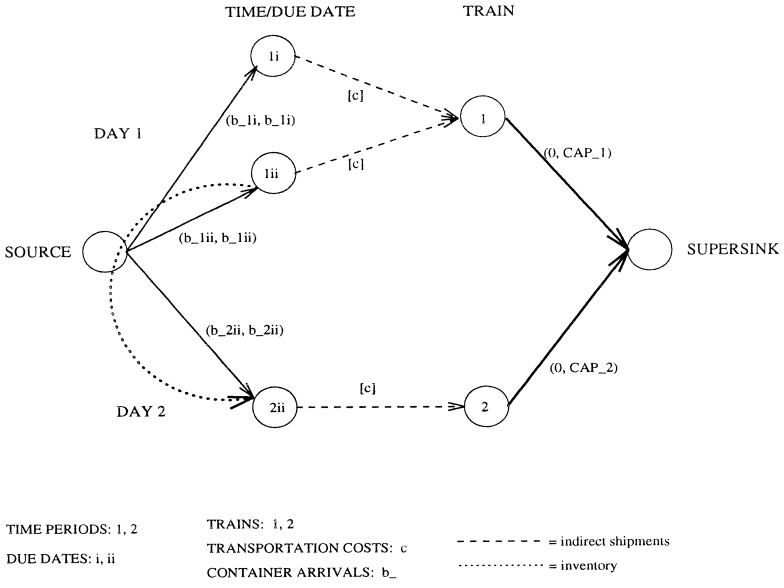

Fig. 3. Network depiction of the hub scheduling problem for a given train schedule.

Figure 3 depicts the hub scheduling problem for a single hub–destination pair for a given train schedule. The examples have two time periods, two due dates, and both travel times and hub delays are assumed to be zero. Upper and lower bounds on container flows are given in parentheses, and arc costs are given in brackets.

It is important to note that, with the train schedules fixed, both the origin and hub scheduling problems can be represented as single-commodity network flow problems. In contrast, the original problem with a fixed train schedule remains a multicommodity network flow problem. Thus, our solution strategy takes advantage of not only the direct benefits of the decomposition in creating smaller subproblems, but also the indirect benefits due to the structure of the resulting subproblems. In particular, the fact that the embedded networks are single-commodity network flow problems allows us to relax the integrality constraints on the container flows when solving the origin and hub scheduling problems. In the next section, we show how our methodology can be modified to provide good solutions for larger problems.

3. PREPROCESSING METHOD TO DETERMINE DIRECT TRAINS FOR MANY DESTINATIONS

OUR METHOD TO OBTAIN solutions for problems with a larger number of destinations relies on a heuristic preprocessing step to set the values of direct train variables in the origin scheduling subproblems. Our rationale for heuristically setting the direct train variables (to reduce the size of the remaining problem) is that these decisions depend primarily upon the demand between a single origin and a single destination, and are only indirectly affected by when and how containers are sent to other destinations. Moreover, the primary indirect effect can be captured largely in the flows of containers sent via the hub from the designated origin to all other destinations. Our preprocessing method is motivated by these observations.

For each origin, the preprocessing step proceeds as follows. We construct K different subproblems, where K is the number of destinations. In the kth subproblem, $k = 1, \ldots, K$, we partition the set of destinations into two groups: a single destination, k, and the remaining $K - 1$ destinations, which we aggregate into a "super-destination." Demands are aggregated across destinations within the super-destinations, taking into account differences in travel times. (In effect, demands with the same latest departure dates are grouped together.) Weighted average fixed and variable costs are assessed for the direct and indirect routes between the origin and the aggregated destination.

This problem is now treated as an origin scheduling problem with two destinations. Direct and indirect train schedules for both the single (kth) and the aggregated destination are derived, along with the corresponding container routing schemes, but only the direct train schedule for the kth destination is retained. Therefore, at the end of this preprocessing step, for each origin, we have established direct train schedules for all K destinations. Having set the direct train variables for all origin–destination pairs in the preprocessing step, we solve the origin scheduling problems to determine indirect train schedules and all container flows.

This procedure generally will not provide an optimal solution to the original origin scheduling problem because the direct train schedules are determined without full consideration of the details of the indirect train schedules. However, recall that the origin scheduling problem is an approximation in itself. From the viewpoint of solving the original problem, it would appear that there is greater loss of optimality from decoupling the origins to create the origin scheduling problems (and from the inability of commercial software to find an optimal solution to the original origin scheduling problem) than there is from the use of this aggregation procedure in solving the individual origin scheduling problems. Moreover, because this preprocessing step entails collapsing a many-destination problem into a two-destination problem, it can be used even when the number of destinations is large.

In the next section, we develop valid inequalities to obtain tight lower bounds for the original model.

4. LOWER BOUNDS

TO EVALUATE THE performance of our procedure, we use lower bounds as one type of benchmark. The lower bounds provided by commercial software are based on linear programming relaxations, which are poor because they ignore the fixed-charge nature of the train costs. These bounds can be improved considerably by adding valid inequalities (cuts) to the monolithic problem. These cuts are used to tighten the lower bound on the original problem (but do not necessarily yield an improved integer solution).

The cuts pertain to the minimum number of trains required to service certain subsets of the demand. Because of the substitutability of trains across time to service non-urgent containers, it is difficult to obtain tight lower bounds on the number of trains on any segment in any time period. We can, however, obtain fairly tight lower bounds on the following:

i. the total number of trains outbound from each origin to each destination during the horizon (single origin–single destination constraint type 1, or sosd1);
ii. the total number of trains outbound from each origin to all destinations (collectively) during the horizon (single origin–all destinations, or soad);
iii. the total number of trains inbound to each destination from each origin during the horizon (single origin–single destination constraint type 2, or sosd2); and
iv. the total number of trains from all origins inbound to each destination during the horizon (all origins–single destination, or aosd).

Let M_{ik}, M_i, and M_k be the lower bounds on the number of trains sent during the horizon associated with the origin–destination pair (i, k), with an origin i, and with a destination k, respectively. The four types of cuts are stated algebraically as:

$$\sum_{t=1}^{T} z_{ikt}^{ao} + \sum_{t=1}^{T} \sum_{j} z_{ijt}^{eo} \geq M_{ik} \quad \forall i, k, \qquad \text{(sosd1)}$$

$$\sum_{t=1}^{T} \sum_{k} z_{ikt}^{ao} + \sum_{t=1}^{T} \sum_{j} z_{ijt}^{eo} \geq M_{i} \quad \forall i, \qquad \text{(soad)}$$

$$\sum_{t=1}^{T} z_{ikt}^{ao} + \sum_{t=1}^{T} \sum_{j} z_{jkt}^{h} \geq M_{ik} \quad \forall i, k, \qquad \text{(sosd2)}$$

$$\sum_{t=1}^{T} \sum_{i} z_{ikt}^{ao} + \sum_{t=1}^{T} \sum_{j} z_{jkt}^{h} \geq M_{k} \quad \forall k. \qquad \text{(aosd)}$$

To obtain values of M_{ik}, M_i, and M_k for all i and k, we introduce the concept of a "supertrain," a fictitious train type with the advantages of both direct and indirect trains. A supertrain emanating from the origin can deliver shipments to any destination, giving it the geographical consolidation advantage of indirect trains. However, like direct trains, it incurs neither the cost nor time delays associated with passing through a hub. Travel time to any destination is the direct travel time for the origin–destination pair. To obtain lower bounds on the number of trains into each destination (sosd2 and aosd), we reverse the network and reindex time appropriately.

Before solving each supertrain problem, we first determine whether there are any demands that necessitate direct trains. We set the corresponding direct train variables to 1 (or to a value large enough to accommodate these demands, if more than one train is needed for an origin–destination pair). We then solve the problem of finding a supertrain schedule and container flows on both the supertrains and pre-set direct trains that minimize the total number of supertrains outbound from the origin while satisfying on-time delivery requirements. To establish the value for M_i, we solve the supertrain problem for a single origin and all destinations. To establish the value for M_k, we solve the supertrain problem on a reversed version of the network and consider all origins and a single destination. To establish the value for M_{ik}, we solve the supertrain problem for just a single origin–destination pair on either the original or the reversed network.

We now demonstrate that the optimal objective value for the supertrain problem (which has the objective of minimizing the number of trains) is a lower bound on the optimal number of trains for the cost-minimization (i.e., the original) problem. Note that the assumptions in the supertrain problem lead to differences in the constraints, not just in the objective function. Therefore, for completeness, we demonstrate this result formally.

We consider the train schedule component of the optimal solution to the cost minimization problem, say $\{z_{ikt}^{ao*}, z_{ijt}^{eo*}, z_{jkt}^{h*}\}$. From this schedule, we show how to construct a feasible solution to the supertrain problem with the same number of trains. Because the optimal solution to the supertrain problem has an equal or fewer number of trains than any feasible solution to that problem, the optimal solution to the supertrain problem provides a valid lower bound on the number of trains in the cost minimization problem.

We now show how to construct a feasible solution to the supertrain problem from $\{z_{ikt}^{ao*}, z_{ijt}^{eo*}, z_{jkt}^{h*}\}$. Let $\tilde{z}_{it}^{o}$ be the number of supertrains departing from origin i at time t and $\tilde{z}_{kt}^{d}$ be the number of supertrains arriving at destination k at time t. These supertrains will substitute for both direct and indirect trains in the original minimum-cost solution. First, substitute a supertrain for each direct train outbound from origin i in the minimum cost solution and retain the container assignments. The supertrain has the same transit time as a direct train, and thus satisfies on-time requirements for the containers assigned to that train. Next, substitute a supertrain for each indirect train outbound from origin i in the minimum cost solution, again retaining the container assignments. The supertrain also has the same transit time as a direct train for any destination, and thus also satisfies on-time requirements for the containers on that train.

We can now set $\tilde{z}_{it}^{o} = \Sigma_k z_{ikt}^{ao*} + \Sigma_j z_{ijt}^{eo*}$. The analysis for trains inbound to destination k parallels the analysis above, and we can set $\tilde{z}_{kt}^{d} = \Sigma_i z_{ikt}^{ao*} + \Sigma_j z_{jkt}^{h*}$. Thus, we have constructed a feasible solution to the supertrain problem from $\{z_{ikt}^{ao*}, z_{ijt}^{eo*}, z_{jkt}^{h*}\}$ that has the same

number of trains as in the minimum cost solution. Consequently, the minimum number of trains in the supertrain problem provides a lower bound on the number of trains in the minimum cost problem.

5. SIMPLE HEURISTICS

WE DEVISE TWO heuristics that are designed to mimic current operating policies. These simple heuristics provide additional benchmarks and allow us to estimate the potential savings from using our procedure. At the intermodal operation that motivated our research, most, if not all, containers are sent via a hub, which leads to some late deliveries. Because we require on-time delivery in our model, these two simple heuristics include the provision for direct shipments when required to satisfy on-time delivery requirements.

In both heuristics, all containers are sent out as soon as possible after they arrive at the origin or at the hub. Containers requiring expedited service are sent on direct trains. Non-urgent containers are assigned to direct trains (which have lower transportation and handling costs) to the extent space is available. Then, all remaining non-urgent containers are assigned to indirect trains. Thus, the heuristics need only specify the number of trains of each type.

The main difference between the two heuristics lies in the conditions for sending direct trains, beyond the minimum required to service expedited containers. In Heuristic 1, direct trains also are sent when there are enough containers to fill a train, even if direct service is not required. We also allow up to one additional direct train with non-urgent containers for each origin–destination pair, provided it is at least θ full, $0 \leqslant \theta \leqslant 1$. In practice, θ is a parameter determined by management, taking into account the tradeoff between the opportunity cost of operating a less-than-full direct train and the additional costs for containers sent through the hub. In Heuristic 2, only necessary direct trains are sent; no trains that contain exclusively non-urgent containers are sent directly. The first heuristic attempts to minimize costs by sending as many containers as possible, or as practicable, on direct trains (which have lower costs). The second heuristic foregoes this opportunity, emphasizing instead the opportunity for consolidation on trains traveling into and out of the hub.

We now describe Heuristic 1 in more detail. For each origin–destination pair, compute the number of direct trains:

$$\text{Set} \quad z^{ao}_{ikt} = \left\lfloor \frac{\sum_{l=t+\alpha_{ik}}^{t+\beta_{ij}+\gamma_{jk}+\delta_j-1} b_{iktl}}{C} \right\rfloor + y \quad \forall i,\ k,\ t,$$

where y represents the smallest number of additional direct trains to be sent other than those carrying containers requiring expedited service, and is the smallest nonnegative integer such that

$$\left\{ \sum_{l \geqslant t+\beta_{ij}+\gamma_{jk}+\delta_j} b_{iktl} - \left(C * z^{ao}_{ikt} - \sum_{l=t+\alpha_{ik}}^{t+\beta_{ij}+\gamma_{jk}+\delta_j-1} b_{iktl} \right) \right\}^+ > \theta * C + (y-1) * C.$$

The corresponding number of containers that are sent directly is given as

$$x^{ao}_{iktl} = b_{iktl} \quad \forall i,\ k,\ t,\ t+\alpha_{ik} \leqslant l < t+\beta_{ij}+\gamma_{jk}+\delta_j,$$

$$x^{ao}_{iktl} = b'_{iktl} \quad \forall i,\ k,\ t,\ l \geqslant t+\beta_{ij}+\gamma_{jk}+\delta_j,$$

where $b'_{iktl} \leqslant b_{iktl}\ \forall i,\ k,\ t,\ l \geqslant t+\beta_{ij}+\gamma_{jk}+\delta_j$ containers not requiring expedited service are chosen for direct shipment because either they fill a direct train (with or without expedited containers), or they constitute at most one train not carrying any expedited containers, but which is at least θ full.

Then, for each origin and time period, compute the number of indirect trains needed to service the remaining containers not sent on a direct train:

$$\text{Set} \quad z^{eo}_{ijt} = \left\lfloor \frac{\left\{ \sum_k \left\{ \sum_{l \geqslant t+\beta_{ij}+\gamma_{jk}+\delta_j} b_{iktl} - \left(C * z^{ao}_{ikt} - \sum_{l=t+\alpha_{ik}}^{t+\beta_{ij}+\gamma_{jk}+\delta_j-1} b_{iktl} \right) \right\} \right\}^+}{C} \right\rfloor \quad \forall i,\ j,\ t.$$

For the case of a single hub, the corresponding number of containers sent indirectly to the hub is given as

$$x^{eo}_{ijktl} = b_{iktl} - b'_{iktl} \quad \forall i,\ k,\ t,\ l \geqslant t+\beta_{ij}+\gamma_{jk}+\delta_j, \quad \text{and} \quad j = 1.$$

Compute the number of containers from each origin i ready to depart the hub for destination k at time $t + \beta_{ij} + \delta_j$ as

$$\sum_l x^h_{ijk(t+\beta_{ij}+\delta_j)l} = \left\{ \sum_{l \geqslant t+\beta_{ij}+\gamma_{jk}+\delta_j} b_{iktl} - \left(C * z^{ao}_{ikt} - \sum_{l=t+\alpha_{ik}}^{t+\beta_{ij}+\gamma_{jk}+\delta_j-1} b_{iktl} \right) \right\}^+ \quad \forall i,\ k,\ t, \quad \text{and} \quad j = 1.$$

TABLE I

Test Problem Characteristics

Problems	Number of Origins–Hubs Destinations	Relative Proportion Expedited Service Demanded
1–5	3–1–3	~20%
6–10	3–1–3	~10%
11–13	3–1–4	~20%
14–15	3–1–4	~10%
16–18	4–1–3	~20%
19–20	4–1–3	~10%
21–25	6–1–6	~20%
26–30	6–1–6	~10%

TABLE II

Parameters for Test Problem Instances

Parameter	Range Used in Test Problems
Container arrival rate per day	0–65
Fixed cost at origin (direct train) ($/train)	11000–15000
Fixed cost at origin (indirect train) ($/train)	5900–8500
Fixed cost at hub ($/train)	6300–8500
Transportation cost ($/container)	40–100
Handling cost (all locations) ($/container)	1–2
Inventory holding cost ($/container/day)	1.5–2

Finally, for each destination and time period, compute the number of trains needed to accommodate the number of containers ready to depart from the hub:

$$\text{Set} \quad z^h_{jk(t+\beta_{ij}+\delta_j)} = \left\lceil \frac{\sum_i \sum_{l \geq t+\beta_{ij}+\gamma_{jk}+\delta_j} x^h_{ijk(t+\beta_{ij}+\delta_j)l}}{C} \right\rceil$$

$$\forall k, t, \quad \text{and} \quad j = 1.$$

Heuristic 2 can be described as follows: Send a direct train between an origin and a destination when necessitated by the due dates of containers. After the containers requiring urgent service have been given priority, fill the remaining space with non-urgent containers in any order (which we justify below). Send as many indirect trains from the origin as necessary each day to ship all remaining containers (i.e., those not sent directly). Send indirect trains from the hub to each destination each day to ship the containers arriving at the hub.

Any allocation of remaining non-urgent containers to indirect trains is feasible and has the same cost. Feasibility follows because the containers can be sent on either train type. Total variable costs are equal because of our cost structure. Total fixed cost is equal because the number of daily direct trains depends only on pre-determined demand, and the number of daily indirect trains is independent of the slack in the non-urgent containers' schedules.

6. NUMERICAL RESULTS

WE GENERATED 30 problems with one hub, three to six origins and destinations, and with different container-demand patterns and cost structures. We summarize problem characteristics in Tables I and II. All problems have eight time periods in which containers become available at the various origins. All trains have a capacity of 200 containers. Container demand was generated for each origin–destination–arrival time–due date combination with a probability of 0.55 of being randomly generated from a discrete uniform distribution between 10 and 65, and a probability of 0.45 of being 0. Scenarios in which less expedited service is demanded are generated as described above, except that, for each origin–destination–arrival time–due date combination such that the shipment necessitated transport via direct train (i.e., $t + \alpha_{ik} \leq l < \beta_{ij} + \gamma_{jk} + \delta_j$), demands that were originally positive are independently set to zero with probability 0.4–0.5.

Table II gives ranges of values for container arrival rates and the cost parameters. Industry data suggest that fixed and variable transportation costs for shipping a full train are approximately equal. We set the fixed charge associated with each train to be proportional to the distance, based on our observation that train-operator labor constitutes the majority of this cost. Handling costs per container are based on the hourly wage of yard operators and the approximate time needed to load, unload, or reposition a container. The yard storage cost per container per day is assigned a small value that provides incentive to ship earlier rather than later, all else being equal. For the first heuristic, we set $\theta = 0.65$ for all problems.

The problems were solved on a Sun SparcStation 20 with 128 megabytes of RAM using CPLEX 6.0. We obtained solutions and lower bounds for the monolithic problem by using CPLEX (with a time limit of 9000 seconds), with and without the valid inequalities discussed in Section 4. We also obtained solutions using our decomposition procedure and the two simple heuristics. For all executions of the CPLEX software, we use depth-first search, and strong branching, i.e., the branching variable is selected whose resolution is most likely to yield the greatest improvement in the objective function value. Additionally, we had CPLEX implement its built-in heuristic to find integer solutions. This combination of rules provided the best results for our set of problems. We also implemented a priority branching scheme, but it did not lead to significant performance improvement.

TABLE III

Objective Values and Lower Bounds as a Ratio of the Objective from the Decomposition Procedure

Problem	Heuristic 1	Heuristic 2	CPLEX Best Integer Solution	Our Lower Bound	CPLEX Lower Bound
1	1.12	1.12	1.02	0.98	0.80
2	1.12	1.12	1.00	0.98	0.80
3	1.14	1.14	1.01	0.96	0.85
4	1.17	1.17	1.02	0.99	0.81
5	1.15	1.15	1.00	0.97	0.86
6	1.14	1.15	1.01	0.93	0.85
7	1.14	1.15	1.00	0.90	0.81
8	1.11	1.12	1.03	0.92	0.86
9	1.19	1.19	1.03	0.92	0.85
10	1.19	1.18	1.03	0.94	0.84
11	1.09	1.09	1.01	0.94	0.82
12	1.13	1.13	1.01	0.97	0.80
13	1.15	1.16	1.01	0.96	0.84
14	1.16	1.16	1.04	0.91	0.82
15	1.15	1.16	1.01	0.90	0.83
16	1.10	1.10	1.01	0.96	0.79
17	1.18	1.18	1.01	0.97	0.81
18	1.11	1.12	1.00	0.96	0.84
19	1.19	1.19	1.03	0.93	0.82
20	1.22	1.24	1.02	0.93	0.84
21	1.05	1.05	0.99	0.91	0.79
22	1.06	1.06	0.98	0.92	0.79
23	1.08	1.08	0.99	0.93	0.81
24	1.08	1.08	1.00	0.94	0.80
25	1.06	1.06	1.00	0.93	0.79
26	1.07	1.05	0.99	0.90	0.83
27	1.08	1.06	1.00	0.90	0.82
28	1.07	1.06	1.00	0.89	0.81
29	1.08	1.07	0.97	0.87	0.83
30	1.08	1.08	0.99	0.90	0.83

Although the valid inequalities were quite effective for improving the lower bounds (as we discuss in more detail later), we found them to be ineffective for improving the best integer solution found prior to the time limit or for reducing overall computational effort. One reason is that the cuts specify bounds on the total number of trains outbound from an origin or inbound to a destination, and thus also on the sum of fixed costs associated with the trains. With such cuts, we are able to get much larger lower bounds (on costs) at all levels of the search tree, and thus, also, a tighter lower bound upon termination. In contrast, such aggregate constraints provide little guidance in the search process. Meanwhile, modest computational effort must be expended to compute the right-hand-side values for the valid inequalities, and the introduction of the valid inequalities slows the overall execution of CPLEX in solving the problem, often resulting in inferior solutions. For this reason, in what follows, we report solutions and CPU times for the monolithic problem without valid inequalities.

Results appear in Table III. All results are reported as the ratio of the relevant objective value to that of the objective value from the decomposition procedure. As shown in the second and third columns, the simple heuristics yield solutions that are about 12% more costly than those found using our decomposition technique. The results demonstrate that significant cost savings can be realized from our proposed procedures over systematic but simpler heuristics. Results in the fourth column show that our decomposition procedure yields an average improvement of approximately 1% versus the best integer solution obtainable from a straightforward implementation of CPLEX with a 9000-second time limitation. In 6 of the 30 problems, the objective value from our decomposition approach is 1–3% above that of the straightforward CPLEX solution, but, as we will show later, this small loss of solution quality comes with a significant reduction in CPU time. Moreover, some or all of this loss can be regained by using refinements of our basic approach that require very little computing effort (see Newman and Yano, 2000).

The fifth and sixth columns report the ratio of the decentralized objective value to lower bounds, where the lower bounds are obtained by solving the monolithic problem with and without the valid inequalities (discussed in Section 4), respectively. It is evident that the addition of the valid inequalities improves the bounds substantially. Furthermore, the tighter bounds demonstrate that our decomposition procedure performs well in an absolute sense. On average, the solutions from the decomposition procedure are within 6.6% of the corresponding lower bounds, which is remarkable considering the strong role of the fixed costs in our problem and the fact that the bounds are (still) based on linear programming relaxations. The bounds are tighter in instances for which more expedited service is required because many direct train variables must be set to 1 (or more). This, in turn, makes the valid inequalities effectively tighter. By contrast, the straightforward application of CPLEX yields bounds that average 18% less than the corresponding objective function values from the decomposition procedure.

Table IV contains CPU times that we report in two different ways. The third column reports the sum of the CPU times for all origin scheduling subproblems and the hub scheduling subproblems. The fourth column reports the elapsed time that would be required if the origin scheduling problems could be solved in parallel, and, subsequently, the hub scheduling problems could be solved in parallel. The solution times from the monolithic problem are reported as the time at which the best integer solution

TABLE IV
CPU Time Performance

Problem	Number of Origins–Hubs Destinations	Serial Run Time (Decentralized Approach) (sec.)	Parallel Run Time (Decentralized Approach) (sec.)	Centralized Approach (sec.)*
1	3–1–3	†	†	1825‡
2	3–1–3	†	†	9000‡
3	3–1–3	†	†	6802‡
4	3–1–3	†	†	6051‡
5	3–1–3	†	†	4605‡
6	3–1–3	†	†	7892‡
7	3–1–3	142	142	8622‡
8	3–1–3	103	103	5662‡
9	3–1–3	†	†	192‡
10	3–1–3	†	†	5008‡
11	3–1–4	250	250	3824‡
12	3–1–4	97	97	1266‡
13	3–1–4	†	†	534‡
14	3–1–4	517	290	2258‡
15	3–1–4	2112	1800	4700‡
16	4–1–3	†	†	3008‡
17	4–1–3	†	†	683‡
18	4–1–3	†	†	497‡
19	4–1–3	†	†	2910‡
20	4–1–3	†	†	2800‡
21	6–1–6	972	221	1295‡
22	6–1–6	158	69	6469‡
23	6–1–6	114	69	1223‡
24	6–1–6	736	139	5870‡
25	6–1–6	810	113	1185‡
26	6–1–6	2492	355	2642‡
27	6–1–6	1019	226	6383‡
28	6–1–6	912	896	1000‡
29	6–1–6	536	207	5498‡
30	6–1–6	670	111	8708‡

*Indicates time at which best integer solution is first identified.
†Indicates CPU time is less than five seconds.
‡Time limit of 9000 seconds is reached.

is identified within the preset time limit of 9000 seconds. In all cases, the 9000-second time limit is reached without a verified optimal solution. For virtually all of the problems, the CPU time for the decomposition procedure is a small fraction of that required to first identify the best integer solution found within the 9000-second time limit, and the latter times could be achieved in practice only if one were clairvoyant about the best time to terminate the search.

The majority of the CPU time is associated with solving the preprocessing step required for the larger origin scheduling subproblems (i.e., those with six destinations in our data sets). This preprocessing step is executed (optimally) in about one minute or less, on average, for our problems, but, as the problem size grows, a greater number of these problems must be solved. Although the CPU times for serial processing increase with the number of origins and destinations, parallel run times remain very modest. The computing effort is dominated by the origin scheduling problems; the solution times for the hub scheduling problems are negligible in almost all instances. In general, problems with more expedited service are solved more quickly, as certain direct-train variables can be fixed (or are set in advance in the CPLEX pre-solve operation).

7. CONCLUSIONS

WE HAVE ADDRESSED the problem of simultaneously determining train scheduling and container routing decisions in a rail intermodal setting. We have developed a decomposition procedure that takes advantage of the embedded network structure, and yields near-optimal solutions in less than one-third the time of commercial optimization software. We have also developed methods for obtaining tight lower bounds using valid inequalities. The entire decomposition procedure yields solutions with objective function values about 12% lower, on average, than those obtained with the simple heuristics described in Section 5. Managerially, our procedure has the advantage of allowing decisions to be made locally at each terminal while achieving, on average, better performance than the best solutions obtained from commercial software for the monolithic problem.

For problems with three or four origins and destinations, we are able to obtain optimal solutions for the origin scheduling subproblems in our decomposition procedure. It is difficult to obtain (verified) optimal solutions to these subproblems (which must consider all destinations simultaneously) when the number of destinations grows larger. To deal with such situations, we have developed a variation of our decomposition method that relies on a preprocessing step to set certain direct-train variables heuristically before the remainder of the problem is solved. For a network with a single hub, this preprocessing step is effective in reducing the size of the remaining origin scheduling subproblems, allowing us to solve them to optimality (for problems of the size considered in our experiments). The method for constructing valid inequalities described in Section 4 provides much tighter lower bounds than those provided by a straightforward implementation of CPLEX and demonstrates that our decomposition procedure yields solutions within 6.6% of the optimum, on average.

Future research is needed to capture other practical aspects of the problem. We mentioned earlier that, in practice, some containers may be shipped late to avoid sending a train with only a few containers. Our model can be generalized to allow tardy

deliveries (with tardiness penalties). Other generalizations include incorporation of hub capacity for container handling, and networks with multiple hubs. There may be parallel hubs, with each shipment passing through (at most) one hub, or serial hubs, with shipments passing through one or more hubs en route to their destinations.

In the case of parallel hubs, the origin scheduling problem is similar to that for the single hub case; the decisions outbound from the origin include direct train schedules and corresponding container shipments to each destination, and indirect train schedules and corresponding container movements to each of several hubs. Estimates of the cost to go can be made in much the same way as for a single-hub problem. Once the origin scheduling problems are solved, a hub scheduling problem is solved for each hub–destination pair, because these problems can be decoupled once the arrivals to the hubs are known.

For scenarios involving two or more hubs in series, intermediate hub scheduling subproblems must be solved sequentially starting from the origins and moving toward the destinations. Because there may be several rail segments on a path between an origin and destination, accurately estimating the cost to go may be more difficult. Preliminary computational studies yield good results even without highly accurate cost-to-go estimates. Methods for obtaining valid inequalities for the single hub case can be extended in a straightforward manner for multiple (parallel or serial) hub scenarios.

We believe that the decomposition concepts proposed here provide an opportunity for considerable cost reduction while maintaining the advantages of local decision-making. The benefits derive from looking ahead, not only with respect to downstream freight flows, but also with respect to forecasted future demands and opportunities to consolidate shipments across time.

ACKNOWLEDGMENTS

THIS WORK HAS BEEN supported in part by National Science Foundation Grant GER/HRD 93-96288 to the University of California, Berkeley, and by funding from the United States Department of Transportation and the California Department of Transportation, awarded by the University of California Transportation Center. We also appreciate comments of the referees and Editor on an earlier version of this paper.

REFERENCES

R. AHUJA, T. MAGNANTI, AND J. ORLIN, *Network Flows: Theory, Algorithms and Applications,* Chapter 17. Prentice Hall, Englewood Cliffs, NJ, 1993.

C. BARNHART AND H. RATLIFF, "Modeling Intermodal Routing," *J. Business Logist.* **14,** 205–223 (1993).

J. CORDEAU, P. TOTH, AND D. VIGO, "A Survey of Optimization Models for Train Routing and Scheduling," *Transp. Sci.* **32,** 380–404 (1998).

T. CRAINIC AND J. ROUSSEAU, "Multicommodity, Multimode Freight Transportation: A General Modeling and Algorithmic Framework for the Service Network Design Problem," *Transp. Res.* **20B,** 225–242 (1986).

R. DIAL, "Minimizing Trailer-on-Flat-Car Costs: A Network Optimization Model," *Transp. Sci.* **28,** 24–36 (1994).

M. GAREY AND D. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP Completeness,* 216–217. W. H. Freeman and Company, New York, 1979.

M. GORMAN, "An Application of Genetic and Tabu Searches to the Freight Railroad Operating Plan Problem," *Ann. Opns. Res.* **78,** 51–69 (1998a).

M. GORMAN, "Operating Plan Model Improves Service Design at Santa Fe Railroad," *Interfaces* **28,** 1–12 (1998b).

M. KEATON, "Designing Optimal Railroad Operating Plans: Lagrangian Relaxation and Heuristic Approaches," *Transp. Res.* **23B,** 415–431 (1989).

A. MARÍN AND J. SALMERÓN, "Tactical Design of Freight Networks. Part I: Exact and Heuristic Methods," *Eur. J. Oper. Res.* **90,** 26–44 (1996).

D. MCKENZIE, *Intermodal Transportation: The Whole Story,* Simmons-Boardman, Omaha, NE, 1989.

E. MORLOK AND L. SPASOVIC, "Redesigning Rail–Truck Intermodal Drayage Operations for Enhanced Service and Cost Performance," *J. Transp. Res. Forum* **34,** 16–31 (1994).

G. NEMHAUSER AND L. WOLSEY, *Integer and Combinatorial Optimization,* 323–343. Wiley Interscience Series in Discrete Mathematics and Optimization, New York, 1988.

A. NEWMAN, Optimizing Intermodal Rail Operations, Ph.D. dissertation, Department of Industrial Engineering and Operations Research, University of California, Berkeley, 1998.

A. NEWMAN AND C. YANO, "Centralized and Decentralized Train Scheduling for Intermodal Operations," *IIE Transactions,* **32,** 743–754 (2000).

L. NOZICK AND E. MORLOK, "A Model for Medium-Term Operations Planning in an Intermodal Rail–Truck Service," *Transp. Res.* **31A,** 91–107 (1997).

C. YANO AND A. NEWMAN, "Scheduling Trains and Containers with Dynamic Arrivals and Due Dates," Working paper, Department of Industrial Engineering and Operations Research, University of California, Berkeley, 1998, *Transp. Sci.*, forthcoming.

(Received: October 1998; revisions received: June 1999, October 1999; accepted: October 1999)