Logistics Service Network Design for Time-Critical Delivery

Cynthia Barnhart[1] and Su Shen[2]

[1] Massachuesetts Institute of Technology, Cambridge, MA 02139
[2] FedEx, Memphis, TN 38125

Abstract. Service network design is critical to the profitability of express shipment carriers. In this paper, we consider the problem of designing the integrated service network for premium and deferred express shipment delivery. Related existing models adapted to this problem are intractable for realistic problem instances, requiring excessive computer memory and solution times. We extend existing models and introduce a new approach to solve the resulting integrated service network design model. Our approach results in order or magnitude reductions in the numbers of variables to be considered in the integer program, allowing us to solve previously unsolvable problem instances. Applying our approach to the service network design problems of a large express package service provider, we demonstrate the potential for tens of millions of dollars in annual operating cost savings, and reductions in the numbers of aircraft needed to perform the service.

1 Introduction

In 1998, UPS generated revenues of $7.1 billion in domestic, air-express shipment service [12]. In 2002, the revenue of UPS' air-express shipment service grew by more than 15%, to about $8.2 billion [12]. Many Wall street analysts attributed UPS' revenue growth and gain in market share to its emphasis on operating efficiency [10]. Efficient operations give a carrier a decisive competitive advantage, allowing the carrier to price its service more aggressively and gain market share, or use the cash flow generated to make further advantageous investments.

Given the high-revenue and low-operating margins of air express shipment service, even a single-digit percentage reduction in operating costs translates to a significant increase in profitability. Because service network design is at the core of express shipment delivery, it is a critical element in achieving operating efficiency. In this paper, we develop optimization models and algorithms to facilitate the design of cost-minimizing express shipment service.

1.1 Problem Description

Express shipment carriers operate transportation equipment, including both aircraft and ground vehicles, and fixed facilities, such as hubs, to serve customer pick-up and delivery requests within tight time windows. Fig. 1 depicts a partial

express shipment delivery service network. Typically, packages are transported by ground vehicles to *ground centers,* or more specifically, *origin ground centers.* A ground center can serve as both an origin and a destination ground center, depending on whether the operation is a pickup or delivery. A ground center is usually associated with a city, although there might be several ground centers for a large city. After packages arrive at the origin ground center, an *origin sort* is conducted to determine the routing for each package based on its destination and a pre-specified package service plan. Although there are exceptions, the shipment is transported to a *gateway* (that is, an airport) either by a ground vehicle or a small aircraft. Packages at gateways are then loaded onto jet aircraft and transported over a *pickup route* to a *hub.* Upon arrival at a hub, packages are sorted, consolidated by destination, and loaded for delivery along *delivery routes* to final destination gateways. At the destination gateway, packages are offloaded onto ground vehicles or small aircraft and transported to their respective destination ground centers. At the destination ground center, a *destination sort* is conducted, and packages are loaded onto ground vehicles for final delivery.
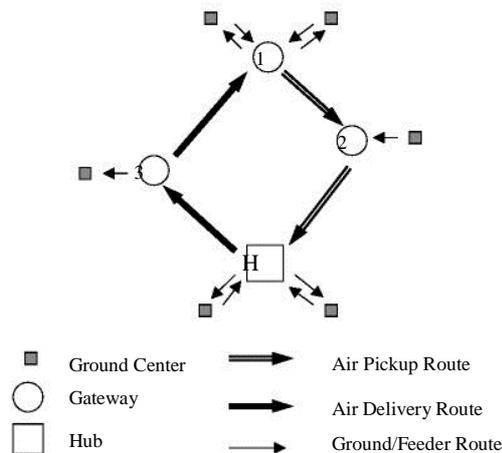
Fig. 1. Express Package Service Operations

Carriers usually offer different levels of service and charge higher premiums for higher levels of service. The level of service is characterized by the time from pick-up to delivery. For example, UPS offers both next-day and second-day services. For shipments picked up on a given day, next-day service has guaranteed delivery by the early morning of the next day, typically before 10

AM, and second-day service has guaranteed delivery by the end of the second day. For both services, the full premium is refunded to customers if delivery is not made on time (UPS [13]).

Operations for different services are similar, with the same equipment and facilities used, although at different times. More specifically, the same aircraft is used to deliver next-day shipments during the night and second-day shipments during the day. Although next-day and second-day operations are performed sequentially, the two services are linked. Fleet position as a result of the pickup and delivery operation of the next-day operation affects fleet position and costs associated with the second-day operation, and vice versa. In the carrier's current practice, because of problem size and complexity, tactical planning for next-day and second-day services is done sequentially, solving two independent problems, one for next-day and another for second-day service, with fleet position fixed in the second problem based on the results of the first. Planning next-day and second-day services simultaneously, that is, considering the *Integrated next-day and second-day problem*, is the focus of the research we describe in this paper.

1.2 Contributions and Paper Outline

The contributions of our research include:

- Designing a solution methodology to solve the integrated next day and second-day express shipment service problem. Because existing approaches are intractable for large-scale problems, we introduce a new approach that allows us to solve previously unsolvable problem instances. In addition to its relevance to express package delivery, our approach can also be applied to other problem types, including multi-commodity flow problems and crew-scheduling problems, to reduce model size and improve solution speed; and Demonstrating the efficacy of our approach on problem instances provided by a large U.S. carrier. Our results indicate that tens of millions of dollars in annual operating costs can be saved, with even greater potential savings in aircraft ownership costs and hub set-up and maintenance costs.

In Section 2 of this paper, we present our modeling approach for the integrated next-day and second-day express shipment service design problem. Then, in Section 3, we detail our solution approach involving decomposition and column generation. In Section 4, by applying our approach to data representing the integrated next-day and second-day express shipment operation of UPS, we demonstrate the scalability and practical significance of our work.

2 Modeling the Integrated Next-Day Air and Second-Day Air Problem

for next-day and second-day operations simultaneously. Costs are incurred for aircraft operation (including *ferry flights),* ground vehicle operation and package handling. Ferry flights represent the repositioning of empty aircraft, usually to increase aircraft productivity. Aircraft operating cost includes two components: 1) *block time cost* including crew and fuel costs resulting from operating a flight leg; and 2) *fixed cycle cost* incurred on each flight leg, typically including the landing fees and other one-time charges. Ground vehicle operating costs, largely based on the distance traveled, are much smaller than aircraft operating costs, and hence, we consider them to be zero. Package handling cost also includes two components, a cost based on block time and a fixed handling cost. Block time cost is a proxy for the marginal fuel cost, and handling cost largely includes the package handling cost at ground centers and hubs. Package handling costs are insignificant compared to aircraft operating costs, and hence, we consider them also to be zero.

The shipments for each origin-destination pair must follow a pre-defined service plan specifying the origin and destination gateways, and the hub at which the packages will be sorted. Given this, we aggregate shipments by origin gateway-destination gateway pairs, and refer to them as *origin-destination (O-D) commodities or origin-destination volumes* hereafter. We consolidate OD commodities originating from the same gateway and assigned to the same hub into a single gateway-hub demand, defined as the *pickup demand* for the gateway-hub pair. Similarly, we consolidate O-D commodities destined to the same gateway and assigned to the same hub into a single gateway-hub demand, defined as the *delivery demand* for the gateway-hub pair. We assume all demands are deterministic.

In addition to serving all demands within specified time windows, express shipment service network design is subject to a number of restrictions, including:

1. Conservation of aircraft at gateways and at hubs - the number of arriving aircraft of a specified type must equal the number departing, for each location;
2. Airport capacity - the number of aircraft arrivals at a hub cannot exceed the number of aircraft parking spots at the hub;
3. Aircraft count - the number of aircraft of each fleet type used must not exceed the available number;
4. Aircraft capacity - the packages assigned to each aircraft cannot exceed the aircraft capacity; and
5. Hub sort capacity - the packages routed through a hub must not exceed its sort capacity.

Various forms of the express shipment service network design problem have been studied. Griinert and Sebastian (2000) identify planning tasks faced by postal and express shipment companies and define corresponding optimization models. Leung and Cheung (2000) propose models for the ground distribution network design problem. Kuby and Gray (1993) consider the limited capacity, single-hub problem and apply the formulation to a case study involving Federal

Express' west-coast hub. Barnhart and Schneur (1996) present a formulation for the uncapacitated single-hub problem and Kim et al. (1999), Krishnan et al. (2002) and Armacost, Barnhart and Ware (2002) consider a capacity-restricted, multi-hub problem with flexible hub assignment, and conclude that service network design models, containing both integer aircraft route variables, referred to as *design variables,* and continuous package flow variables, have associated tractability issues. Their corresponding linear programming (LP) relaxations have solutions that are often fractional and difficult to transform into good-quality feasible solutions. Armacost, Barnhart and Ware (2002) report success in overcoming these tractability issues by applying extended formulation techniques that embed package flow decisions within the design variables. Given this, we address the integrated next-day and second-day express shipment service design problem by adapting the modeling approach of Armacost, Barnhart and Ware and developing a new decomposition algorithm.

2.1 A Daily Model for the Integrated Problem

Fig. 2 depicts the service timeline for the next-day operation, assuming packages are collected on Day 1. Carriers schedule pickup of packages from customers as late as possible to allow customers sufficient time to prepare their packages. Hence, packages arrive at origin ground centers in the late afternoon or early evening. After the origin sort in the evening, packages moving by air service are transported to origin gateways at night and loaded onto aircraft. From origin gateways, aircraft are transported along next-day air (NDA) pickup routes and arrive at hubs in the late night or early morning of the next day. The hub sort for NDA packages starts around midnight and lasts for 2-3 hours. After the hub sort, packages are delivered to their destination gateways, and then their destination ground centers, arriving in the early morning of the next day. At that point, the destination sort occurs at the ground center and packages are loaded onto ground vehicles and delivered to customers to meet delivery requirements. The same next-day air operation, starting with air pickup and ending with air delivery, is repeated each day except Sunday.

The second-day operation is similar to the next-day operation except for an expanded service time. Fig. 2 depicts the service timeline for the second-day operation, assuming packages are collected on Day 1. At the origin ground center, the origin sort for second-day packages begins at night after the origin sort for next-day packages is completed. Then, second-day packages to be transported via air service stay at the origin ground center overnight, while others are transported to destination ground centers or hubs via ground service. On the morning of the next day, second-day packages at origin ground centers are transported to gateways and loaded onto aircraft that have just completed their NDA delivery routes. Aircraft then follow second-day air (SDA) pickup routes, arriving at hubs before noon. After the hub sort, packages are delivered either to destination ground centers via ground service or to destination gateways via air service. In the case of air delivery, aircraft carrying SDA packages arrive at destination gateways in the evening of Day 2. After SDA packages are unloaded, aircraft

Next-Day Customer Pickup	Next-Day Origin Sort	Next-Day Air Pickup	Next-Day Hub Sort	Next-Day Air Delivery	Next-Day Destination Sort	Next-Day Customer Delivery
Late Afternoon	Evening	Night	Mid-Night	Early Morning		Mid-Morning
Day 1				**Day 2**		**Time**

Fig. 2. Next-Day Air Operations

are available to begin their NDA pickup routes. The unloaded SDA packages are transported to destination ground centers, where they wait overnight for other second-day packages transported via ground. On the morning of Day 3, the destination sort for second-day packages begins after the completion of the destination sort for NDA packages collected on Day 2. SDA packages are then delivered to customers in the afternoon. Note that compared with the next-day service, the extended service time allows more extensive use of ground transport.

Second-Day Customer Pickup	Second-Day Origin Sort	Second-Day Air Pickup	Second-Day Hub Sort	Second Day Air Delivery	Second-Day Destination Sort	Second-Day Customer Delivery
Late Afternoon	Night	Morning	Noon	Afternoon	Morning	Afternoon
Day 1			**Day 2**		**Day 3**	**Time**

Fig. 3. Second-Day Air Operation

Although the complete second-day operation spans three days, as depicted in Fig. 3, we can model the SDA operation as a *daily problem,* that is, the same operation is repeated daily, because a new second-day operation initiates each day. We illustrate this concept as follows. In Fig. 4, we depict second-day operations over three days. The number in the parenthesis at the upper left corner of each box indicates the starting day of the corresponding SDA operation. We refer to a second-day operation starting on Day n as *second-day Operation n.* On any given day n, there are three sets of second-day activities underway, one set for packages entering the system on Day $(n-2)$, one set

for those entering on Day $(n-1)$, and finally, one for those on Day n. In the morning of Day n, the destination sort for second-day Operation $(n-2)$ is conducted and packages of second-day Operation $(n-1)$ are transported by air to hubs. Around noon, packages of second-day Operation $(n\ 1)$ are sorted at hubs, and then, in the afternoon, packages of second-day Operation $(n\ 2)$ are delivered to customers. Next, in the late afternoon, packages of second-day Operation $(n\ 1)$ are delivered by air to destination gateways, and packages of second-day Operation n are collected from customers. Finally in the night, the origin sort for the second-day operation n is conducted. As is evident in Fig. 4, the same air operation is repeated daily in second-day operations.
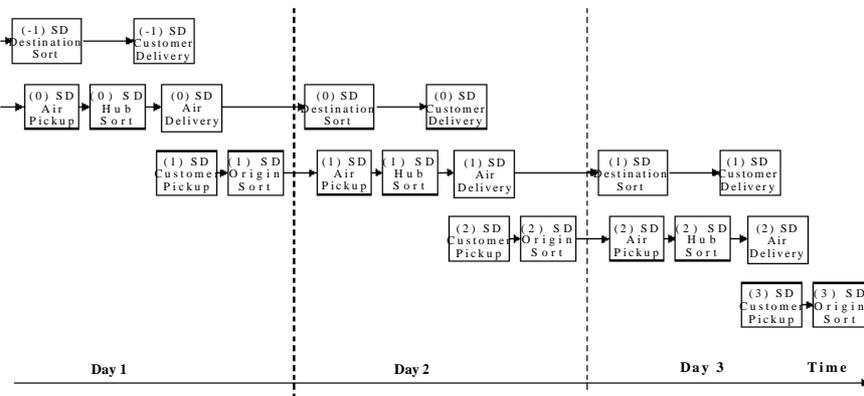
Fig. 4. Daily Second-Day Air Operation

By recognizing that the second-day express shipment service operation can be captured by a single, representative day, we are able to model the integrated next-day and second-day operation as a daily problem. This allows us to minimize the number of variables and constraints in our models, and helps to reduce the challenges associated with solving these very large-scale formulations.

2.2 Integrated Problem Formulation

Armacost, Barnhart and Ware (2002) present a new model for express shipment service network design using *composite variables* to reduce fractionality of the LP relaxation and enhance tractability. They define a *demand composite* to be a set of aircraft routes providing sufficient capacity to transport *all* the demand between the nodes contained in the selected aircraft routes. We illustrate the concept through a simple example in which we have 3 units of demand to be

transported from gateway i to hub h. There is a single fleet type with capacity of 2 units. The operating cost of each aircraft on the route i to H is 10 units. One possible composite variable, denoted c, is two aircraft from i to h with cost 20, providing 4 units of capacity to transport all 3 units of demand. Note that one aircraft from i to h is not a valid composite variable because 2 units of capacity is insufficient to serve all the demand from i to h. In conventional network design models, to ensure that the 3 units of demand are served, we specify a constraint

$$2\ y > 3,$$

with variable y representing the number of aircraft selected. The optimal solution to the LP relaxation is then 1.5 aircraft, with 15 units of operating cost. In contrast, with composite variables, the condition that all demand must be served can be specified as

$$c > 1.$$

In the optimal solution to the LP relaxation using composite variables, c equals one, implying that two aircraft are selected to serve the demand, with a total operating cost of 20 units. This small example illustrates the improved LP bound achievable with composite variables.

We apply the demand composite modeling concept to the integrated problem and introduce the following notation. Let T indicate the type of service— next-day (denoted N) or second-day (denoted S); and 0 indicate the operation—pickup (denoted P) or delivery (denoted D). We define the following additional sets and variables.

Sets

F set of fleet types.
H set of hubs.
Ar set of gateways.
C^T set of demand composites for NDA $(T = N)$ or SDA $(T = S)$ network.
$C7)$composites for   set of   pickup $(0 = P)$ or delivery $(0 = D)$ demand NDA $(T = N)$ or SDA $(T = S)$ network.

Data

a^T {number of aircraft parking spots at hub h for NDA $(T = N)$ or SDA $(T = S)$ network. J pickup $(0 = P)$ or delivery $(0 = D)$ demand between gateway i
$b\ i^h,\ 0$ and hub h for NDA $(T = N)$ or SDA $(T = S)$ network.
    number of aircraft routes r in demand composite c.
$d,$ cost of demand composite c, $d_e = \mathrm{E}_{re}\ e$ -yr$_e$d$_r$.
df_j ferrying cost for an aircraft of type f ferried from gateway i to j.
$,\ r$ {number of aircraft of type f available for NDA $(T = N)$ or SDA $(T = S)$ network.

number of aircraft of type *f* in demand composite *c*.

number of aircraft of fleet type *f* originating at gateway *i* (or hub *h)*
[7Ri)] in demand composite *c*.

number of aircraft of fleet type *f* destined to gateway i (or hub *h)*
[7.(D] in demand composite *c*.

[ih] 1 if demand composite *c* covers NDA *(T = N)* or SDA *(T = S)*
pickup *(0 = P)* or delivery *(0 = D)* demand between gateway *i*
and hub *h,* and
0 otherwise.

Decision Variables

equals 1 if demand composite *c* is selected, and 0 otherwise.

number of aircraft of type *f* on the ground at gateway (hub) *i*
[,-7-''°] during NDA *(T = N)* or SDA *(T = S)* pickup *(0 = P)* or
f,z delivery *(0 = D)* operation. $'_f^T$, $',, = '_{.1,',,}$, [T D if•] z cE *H.*
J

number of aircraft of type *f* ferried from gateway (hub) *i* to *j* after
the NDA *(T = N)* or SDA *(T = S)* operation.

We present the following formulation (INS) for the Integrated NDA-SDA
problem:

$$\min \sum_{T=IN,,SleECT} \sum d_o v_e \sum \sum \sum_{jEN\cdot} {}^{T,f}_{ij} \qquad (1)$$

subject to

$$i\ (z)\ e\sum_{eEai,}{}_q;\qquad ,COv_e- \qquad {}^{S,D}$$

$$,ie \qquad feF \qquad (2)$$

$$\sum_{je.V,joi} \text{Ŀ†C)} \quad {}_{7e\ e} \quad {}_{ryc\ g)v\ e}{}^{N,D} \quad {}^{S,P}+$$

$$+\sum_{j\in\mathcal{N},j\neq i} \qquad eeq; \qquad = 0,\ i\ Ar,\ f\ EF \qquad (3)$$

$$+\varpi^{T,P}_{f,h}-$$

$$\sum_{7):(\_1)ve \atop eecT,} \qquad (T)\ v_e - \qquad = 0,\ h\ e\ H,\ f\ e\ F,\ T = \{N,\ .5\}$$

$$(4)$$

$$\substack{ry \\ eEe\text{l};}$$

$f \, EF, T = \{N, S\}$ (5)

$$\sum_{f \in F} \sum_{c \in E_i^f} r_{yi:(11)} v_c \quad , \quad h \in H, \ T = \{N,S\} \tag{6}$$

$$\sum_{c \in C} f_{T,O,c}^{ish} \qquad > (i,h) \qquad > 0, \ T = \{N, S\}, \ O = \{PD\}, \ i \in N, \ h \in H \tag{7}$$

$$v_c \in \{0, 1\} \text{ for all } c \in C^N \cup C^S \,,$$

$$T_{O}^{f} \in Z_{\pm} \text{ for } T = \{N,S\}, \ O = \{P,D\}, \ i \in N \ c$$
$$Z_{+} \text{ for } i, j \in A^{lh}, i \ j, \ f \in F, \ T = \{N, S\}$$

The objective is to minimize the sum of the total NDA and SDA operating costs and the ferry costs between the operations. Constraints (2) and (3), the *boundary balance constraints,* ensure that aircraft at gateways are balanced between NDA and SDA operations. Constraints (2) require that the number of aircraft of type *f* at a gateway (hub) *i* at the start of the NDA pickup operation equals the number of aircraft of type *f* at gateway (hub) *i* at the end of the SDA delivery operation, adjusted by the number of aircraft of type *f* ferried into and out of gateway (hub) *i* at that time. Constraints (3) similarly require that the number of aircraft of type *f* at a gateway (hub) *i* at the end of the NDA delivery operation equals the number of aircraft of type *f* at gateway (hub) *i* at the beginning of the SDA pickup operation, adjusted by the number of aircraft of type *f* ferried into and out of gateway (hub) *i* at the end of the NDA operations. Constraints (4) are *hub balance* constraints that ensure conservation of flow of aircraft by type at each hub, for both the NDA and SDA operation. The *count constraints* (5) limit the number of aircraft of each fleet type selected in the NDA and in the SDA operation to be no more than the number available. We need only to specify these constraints for pickup routes because conservation of flow constraints ensure that aircraft count will also be satisfied for delivery. The *landing constraints* (6) ensure that the number of aircraft arriving at a hub during NDA and during SDA operations does not exceed the parking spots available. We similarly only specify the landing constraints for pickup routes because aircraft conservation of flow ensures satisfaction for delivery. The *cover constraints* (7) ensure that at least one composite is selected to cover each nonzero gateway-hub demand. Because each demand composite is guaranteed to serve the associated gateway-hub demands fully, the cover constraints also ensure satisfaction of the aircraft capacity constraints. Finally, the last set of constraints ensure that the solution is comprised of a non-negative, integer number of composite variables, representing a set of aircraft routes, some of which will be flown by more than one aircraft.

3 Solving the Integrated NDA-SDA Formulation

Populating the **INS** formulation with all possible variables results in an intractable model: computer memory requirements and solution times are excessive. To address this issue, we use column generation to reduce the number of columns considered in solving the IP.

In column generation, we maintain a restricted version of the original model, called the Restricted Master Problem (*RMP*), which includes only a limited set of columns. At each so called master iteration, we solve the *RMP* to obtain a set of dual prices. Using this set of dual prices, we can either compute the reduced cost of each column explicitly, or solve a pricing sub-problem, as in Dantzig-Wolfe decomposition (Dantzig and Wolfe [4]), to identify columns that potentially can improve the objective value of the *RMP*. If a problem has a diagonal block structure, pricing sub-problems can be specified for each block, resulting in simpler sub-problems. We repeat the process until no column is generated in one master iteration.

In this section, we explore different solution approaches for the **INS** formulation. We refer to the first approach as *naive column generation*; a standard column-generation approach in which demand composite variables with negative-reduced cost are generated as needed, with restrictions on the number of variables generated per iteration. In the second approach, referred to as *aggregate information-enhanced column generation*, smaller hub pickup or delivery sub-problems are solved to generate the necessary variables, and a master column represents the network design for the pickup or delivery operation of a hub. In the third approach, referred to as *disaggregate information-enhanced column generation*, we similarly solve hub pickup or delivery sub-problems, but each master column represents a demand composite variable, and we partition the hub sub-problem solution, that is, the solution to the pricing problem, into individual demand composites when adding columns to the *RMP*. In each of these solution approaches, we limit column generation to the root node LP relaxation, and consider only columns generated in solving the root node LP in branch-and-bound.

To evaluate these solution approaches, we first apply them to UPS' NDA problem only, not the integrated NDA-SDA problem, to gain insights into their respective effectiveness. The UPS NDA network includes 101 gateways, 7 hubs, 9 fleet types, 198 pickup and 195 delivery gateway-hub demands. Formulation statistics are reported in Table 1.

Columns	195,009
Rows	3,302
Nonzeros	2,062,466

All computations were performed on an HP 03000 workstation with 400MHz CPU and 2GB RAM, running HPUX 10.20. The models and column generation processes were compiled using HP's aCC compiler with calls to the ILOG CPLEX 6.5 Callable Library [3]. CPLEX MIP Solver settings are reported in Table 2. For parameters not indicated, the CPLEX default values were used.

Parameter	Setting
Backtrack	0.85
Branching Direction	Up direction selected first
Node Selection	Best estimate search
Variable Selection	Based on strong branching
Relative Best IP-Best Bound Gap Tolerance	0.0001

Table 2. Settings for CPLEX 6.5 MIP Solver

3.1 Naive Column Generation

In *naive column generation,* we evaluate the cost of demand composite variables explicitly using the dual prices obtained from solving the *RMP.* Denote the objective coefficient vector for demand composite variables as d, and the constraint matrix for demand composite variables in constraints (2), (3), (4), (5), (6) and (7) as B_1, B2, H, N, A and C, respectively, and let the dual vector of the corresponding constraints be denoted $7r^{131}$, $7r^{132}$, $7C^H$, $7r^N$ $7r^A$ and ir^e. The reduced cost vector of demand composite variables is given by

$$d' — (7r^{131})13_1 — (7r^{B2})132 — \qquad — (7r^{1\backslash 1}\backslash T — (7r^{-A})4. — (rr^e)C.$$

Demand composite variables with negative reduced cost are generated when solving the LP relaxation. In order to limit the size of the integer programming model, we evaluate the effect of limiting the number of columns generated in one iteration to at most 100, 500, 1000, 2000, and 4000, respectively, and determine that generating at most 1000 columns in an iteration results in the fewest number of columns generated.

Our results for the naive column generation approach, limiting the number of columns generated in one iteration to at most 1000, are reported in Table 3. For comparison, we also solve the problem with *all* demand composite variables present, referred to as the *all-column approach.* "AC" represents the all-column approach, and "NCG" represents the naive column generation approach. In both approaches, the optimal LP value is the same. The objective value of the best IP solution using the naive column generation approach is 0.01% higher than that obtained with the all-column approach. This difference is explained by the fact that we generate columns only at the root node of the branch-and-bound tree,

and hence, we do not consider certain demand composite variables whose reduced cost becomes negative as we branch in the branch-and-bound solution algorithm. This small degradation of the objective value is compensated for by the reduction in algorithmic complexity resulting from limiting column generation to the root node. In comparing running times, the naive column generation approach takes less than one fifth of the time required by the all-column approach.

Solution Approach		AC	NCG
Columns. Generated		-	16259
IP Objective Value		-	+0.01%
Run Time (sec.)	Root Node LP	28	23
	IP	8692	1550

Table 3. All-Column and Naive Column Generation Results for the UPS NDA Problem

3.2 Aggregate Information-Enhanced Column Generation

In our information-enhanced column generation approach, instead of generating *individual* demand composite variables with negative reduced cost, we generate a *set* of demand composite variables that is both feasible and has, summing over the demand composites in the set, a negative reduced cost.

We define a *set* of pickup (or delivery) demand composites to be a *hub pickup* (or *delivery) composite* if it: (1) includes integer numbers of aircraft routes; and (2) satisfies the count constraints, and the landing and cover constraints specified for the pickup (or delivery) gateway-hub demands, at a set of hubs. We introduce the following additional notation.

Sets and Data

N^T set of hub composites for the NDA *(T = N)* or SDA *(T = S)* network.

T $\{$ set of pickup *(O = P)* or delivery *(O = D)* hub composites for the NDA *(T = N)* or SDA *(T = S)* network.

d_e cost of hub composite O, $d_e = \sum_{e}\sum_{e}$, d_e.

ry^f_o number of aircraft of type f in hub composite O.

7^I:9 (7) number of aircraft of type f originating at gateway (hub) i in hub composite O.

$7j$;)(i) number of aircraft of type f destined to gateway (hub) i in hub composite O.

ih
TAO,- 1 if hub composite O covers NDA *(T = N)* or SDA *(T = S)* pickup *(O = P)* or delivery *(O = D)* demand between gateway i and hub h, and

0 otherwise.

Decision Variables

ye   equals 1 if hub composite 0 is selected, and 0 otherwise.

We re-write the INS formulation with hub composite variables (INS-H) as follow.

$$\min \sum_{T=\{N,S\}} \sum_{H \in H} d_e v_e + \sum_{T=IN,S1} \sum_{i \in Ar} \sum_{J \in N} \quad {}_{i\,j}^{T,f} \tag{8}$$

subject to

$$\sum_{ee7-c;;} 'Y(^{f9} (Dye \qquad \sum_{\Theta \in \mathcal{H}_i^N} \quad {}^{Ye}\!\!\!/_{Z}) \backslash ye \quad N,P \quad S,D$$

$$je.V,JAi \qquad \sum_{j \in \mathcal{N}, j \neq i} \qquad , i \qquad f EF \tag{9}$$

$$\sum_{\Theta \subset \mathcal{H}_i;} 7(f9(i)ve \qquad \sum_{\Theta \subset \mathcal{H}_i^p} {}^{r\backslash}\!\!-yek,z)ye \qquad ,D \qquad tv \quad {}_{f,i}^{S,P}$$

$$+ \sum_{\substack{jEN,J\#i}} E \quad = 0, i c V, f c F \tag{10}$$

$$\sum_{\substack{E7\cdot(1;}} f_{\substack{(9(h)ve +\\ fT,P}} \sum_{\substack{jeN,Joi\\ f\\ ·yeAve\\ HE7\cdot(1,}} {}^{,D} \qquad = 0, h c H, f c F, T = \{N, S\}$$

$$\sum_{HEHT,} 7iLv_e \qquad f F, T = \{N, S) \tag{12}$$

$$\tag{13}$$

$$\sum_{\substack{fEF,cEZ\cdot g;\\ >1,}} \sum \quad : b^i P_{,o} > 0, T = \{N, S\}, 0 = \{P, D\}, i \qquad h H$$

$$\tag{14}_7$$

$$(A)ve \quad a_h, h E H, T= \{N, S\}$$

$$E \quad 4,0$$
$$_{HE7\cdot 0)} \qquad ,eve$$

$$ye E \{0, 1\} \text{ for all } (7\cdot) \; 7\text{-}1^N U \; 7\text{-}i\cdot,$$
$$E Z_\pm \text{ for } T = \{N, S\}, 0 = \{P, D\}, i Ar$$
$$E z_+ \text{ for } i,jEN, i \; j, f F, T = \{N, S\}$$

The formulation is the same as the INS formulation except that demand composite variables are replaced with hub composite variables. It is straightforward to show (Shen 2004)that the INS-H formulation is at least as strong as the INS formulation. In the following example we describe a case in which the INS-H formulation is strictly stronger than the INS formulation.

Consider the example in Fig. 5. There is a single fleet type with 2 units of capacity. We want to cover all gateway-hub demands in the example. We

only consider the pickup operation for simplicity, but we can easily expand the examples to include delivery operations and aircraft balance without affecting formulation strength.

1-H: 1 unit
2-H: 1 unit
3-H: 1 unit

(12, 2)

(6, 2)

(8, 2)

(5, 2)

(12, 2)

(d_r, u_r)

Demand composite variables:
dc1: one route 2-H covering demand 2-H.
dc2: one route 3-H covering demand 3-H.
dc3: one route 1-2-H covering demands 1-H and 2-H.
dc4: one route 1-3-H covering demands 1-H and 3-H.
dc5: one route 2-3-H covering demands 2-H and 3-H.

Hub composite variables:
hc1: one route 1-2-H and one route 3-H
      covering demands 1-H, 2-H and 3-H.
hc2: one route 1-3-H and one route 2-H
      covering demands 1-H, 2-H and 3-H.

Fig. 5. Example of Hub Composite Variable

We consider only the demand composite variables and hub composite variables in the figure. (Other demand and hub composite variables do not affect the optimal integer or LP relaxation solution to the **INS** and **INS-H** formulation.) Excluding the balance, count, and landing constraints, the **INS** formulation is

$$dc_3 + dc_4 = 1$$
$$dc_1 + dc_3 + dc_5 = 1$$
$$dc_2 + dc_4 + dc_5 = 1$$
$$dc_i \in \{0, 1\}, \quad i = 1, 2, ..., 5.$$

The resulting optimal solution to the LP relaxation is $\{dc_1 = 0, dc_2 = 0, dc_3 = 0.5, dc_4 = 0.5, dc_5 = 0.5\}$, with objective value 16.

Excluding the balance, count, and landing constraints, the **INS-H** formulation is

$$hc_1 + hc_2 = 1.$$

The optimal solution to the LP relaxation is $\{hc_1 = 1, hc_2 = 0\}$, with objective value 17.

We denote the dual vector for constraints (9), (10), (11), (12), (13) and (14) as π^{B_1}, π^{B_2}, π^H, π^N, π^A and π^C, respectively. Let $C^{T,O,h}$ be the subset of NDA ($T = N$) or SDA ($T = S$) demand composite variables covering subsets of gateway-hub demands to hub h in the case of pickup ($O = P$) or subsets of gateway-hub demands from hub h in the case of delivery ($O = D$), and $v^{T,O,h}$ the vector indicating the selection of those demand composite variables. Following the matrix notation introduced in describing naive column generation, we denote the constraint matrix for demand composite variables in $C^{T,O,h}$ in constraints (2), (3), (4), (5), (6) and (7) as $\mathbf{B}_1^{T,O,h}$, $\mathbf{B}_2^{T,O,h}$, $\mathbf{H}^{T,O,h}$, $\mathbf{N}^{T,O,h}$, $\mathbf{A}^{T,O,h}$, $\mathbf{C}^{T,O,h}$,

respectively. Denote the right-hand-side vector of constraints (5) and (6) for T = *{N, S}* as nT and aT. Denote the right-hand-side vector of constraints (7) for gateway-hub demands for T = *{N, S}*, 0 = *{P, D}* and $h \ e \ H$ as I'''h. We define the following sub-problem for T = $\quad 0$ = *{P, D}*, and $h \ e \ H$:

$$\min \ [d' \qquad O,h \ (_{7r}s_{2\ )13TO,h} \ (_{7r}11)_HT,O,\ _h$$

$$(_rA)_AT,o,h, \ (_{ir}c)c,o,hi_vT,O,h \qquad (15)$$

subject to

$$AT,O,h \ vT,O,h \ aT \ vT,0,h < \ _{nT} \qquad (16)$$

$$cT,O,h \ _vT,O,h > IT,O,h \qquad (17)$$

$$y_e \ C \qquad c \ C \ C^{TPA} \qquad (18)$$

$$(19)$$

Constraints (16) ensure that the selected demand composite variables at hub h satisfy the landing constraints at *all* hubs. We consider all hubs because the demand composite variables in $C''^{,h}$ might include routes entering or departing hubs other than h. Constraints (17) are the count constraints, specified for each fleet type, and constraints (18) are the cover constraints specified for each gateway-hub demand to or from hub h.

The solution to a sub-problem is a hub composite, and the objective value is its reduced cost. If the objective value of a solution is negative, we add the corresponding hub composite variable to the *RMP*. The process terminates if after solving all sub-problems, one for the pickup operation and one for the delivery operation at each hub, and for NDA and SDA, not one sub-problem solution has a negative objective value. Because we ensure the set of columns generated are feasible and the sum of their reduced cost is negative, we call this approach *information-enhanced column generation.* We refer to the information-enhanced column generation approach in which a sub-problem solution is introduced into the *RMP* in its *aggregate* form, that is, as a hub composite variable, *aggregate information-enhanced column generation.*

We apply aggregate information-enhanced column generation to the same UPS NDA problem instance that we solved with the naive column generation and the all-column approaches. Our results are reported in Table 4. Compared with the INS formulation, the optimal LP objective value increases by 0.001%. The MIP solver, however, runs out of memory and fails to find a feasible integer solution after 20 hours with the set of columns generated. The best bound achieved at that point is 2.4% higher than the true IP optimal objective value.

3.3 Disaggregate Information-Enhanced Column Generation

The columns generated by the aggregate information-enhanced column generation at the root node of the branch-and-bound tree fail to provide a feasible

Columns. Generated		7101
Master Iterations		270
Objective Value	Root Node LP	+0.001
	IP	N/A
Run Time (sec.)	Root Node LP	2842
	IP	N/A

Table 4. Aggregate Information-Enhanced Column Generation Results for the UPS NDA Problem

solution. The issue is that too many decisions are embedded in a column of the *RMP*. To overcome this issue, we introduce *disaggregate information-enhanced column generation.*

We replace the INS-H formulation with the INS formulation as the *RMP*. At each master iteration, we similarly solve the pricing problem (15)-(19) for the pickup and delivery operation of each hub and for NDA and SDA. If the objective value of a sub-problem is negative, instead of adding to the *RMP* a single column representing all the demand composite variables in the sub-problem solution, we partition the solution into individual demand composite variables and add to the *RMP* those that are not currently included. (Some demand composite variables might have been included in the *RMP* in earlier iterations.)

We apply disaggregate information-enhanced column generation to the same UPS NDA problem instance and report our results in Table 5.

Columns Generated		1535
Master Iterations		34
IP Objective Value		+0.11%
Run Time (sec.)	Root Node LP	307
	IP	185

Table 5. Disaggregate Information-Enhanced Column Generation Results for the UPS NDA Problem

Using disaggregate information-enhanced column generation, we generate less than 1% of all possible columns, and less than 10% of the number of columns generated using naive column generation. This indicates that the hub-based sub-problems are more effective than naive column generation in identifying columns that can be used in an optimal solution. The root node LP converges to the true objective value, but the IP objective value is somewhat worse than that obtained with naive column generation, because columns are again generated only at the root node of the branch-and-bound tree. Compared with the naive column generation and the all-column approaches, the root node LP relaxation takes longer to solve, but the IP solution time is significantly reduced using disaggregate information-enhanced column generation. Overall, disaggregate information-enhanced column generation achieves a 70% reduction in total solution time

compared with naive column generation, and a 95% overall reduction compared with the all-column approach.

Compared with aggregate information-enhanced column generation, disaggregate information-enhanced column generation not only produces fewer columns, but also converges in fewer master iterations. Most importantly, solutions with objective values close to the optimal value can be identified with the set of columns generated.

4 Case Study

We apply the disaggregate information-enhanced column generation approach to the integrated UPS NDA-SDA problem, with the objective to minimize daily operating costs. Problem statistics are reported in Table 6.

To compare the integrated solution to sequential solutions, we use disaggregate information-enhanced column generation to solve in sequence the NDA and SDA problems. To solve the SDA problem, which is relatively small compared to the NDA problem, we generate only 3113 columns, or 1.4% of all variables, using the disaggregate information-enhanced column generation approach.

Composite Variables	NDA	SDA
	168372	59969
Ferry and Ground Variables	76215	
Rows	4623	
Master Iterations	33	
Generated Demand Composite Variables	3113	

Table 6. UPS Integrated NDA-SDA Problem Statistics

In Table 7, we compare the results of the sequential and integrated approaches with the solution generated by planners at UPS. Costs are reported as the percentage difference from those of the UPS solution. In the UPS solution, the SDA network is designed manually, while the design of the NDA network is accomplished using the composite variable approach of Armacost, Barnhart and Ware (2002).

In the first scenario, the *unconstrained NDA and SDA problem*, boundary balance conditions are not enforced between the NDA and SDA operations, and the two problems are solved independently, without aircraft balance constraints. Their combined solution value provides an upper bound on the potential savings achievable through integration of the NDA and SDA problems. In the second scenario, the NDA problem is first solved without aircraft balance constraints. Then the SDA problem is solved with balance constraints ensuring that the NDA operations can be executed as planned. The resulting total cost is slightly better than that of the UPS solution. Notably in this case, ferry costs increase significantly because many ferry flights are required to re-position aircraft before or

after the NDA operation to perform the SDA operations. These ferry costs more than offset the savings achieved in the NDA solution. In the third scenario, a reverse sequence is followed, the SDA problem, without aircraft balance conditions, is first solved, and the NDA problem, with balance constraints ensuring the execution of the SDA operations, is then solved. The resulting operating costs of the NDA solution are greater than those of the UPS solution, but the daily total cost is much lower. This sequential approach produces less expensive solutions than the previous one for the following reasons:

Because the SDA operation uses only about one third of the fleet used in the NDA operation, there is sufficient flexibility to position the unused aircraft in the SDA operation to match the needs of the NDA operation; and

Most aircraft re-positioning for the SDA operation can be accomplished with revenue flight movements in the NDA operation, given the large number of NDA gateway-hub demands to be served.

Scenario	Daily Revenue Flight Cost	Daily Ferry Flight Cost	Total Daily Cost	Fleet Usag
Unconstrained	-23.4%	-100%	-15.9%	
Unconstrained NDA	-7.3%			
Unconstrained NDA	-7.3%	+903.6%	-0.3%	-4
Constrained SDA	-17.5%			
Unconstrained	-23.4%	+218.5%	-5.9%	-3
Constrained NDA	+1.9%			
Integrated SDA	-19.5%	+140.7	-8.1%	-5
Integrated NDA	-1.2%			

Table 7. Sequential and Integrated Approach Results for the UPS NDA-SDA Problem

In the last scenario, we solve the integrated NDA-SDA problem with disaggregate information-enhanced column generation. Although ferrying costs are more than double those in the UPS solution, the NDA and SDA operating costs are both reduced, reflecting the better coordinated aircraft movements. The daily operating cost savings of the integrated approach translates into tens of millions of dollars annually. Compared with the best sequential approach, the savings from the integrated approach come from: (1) reduced ferry costs; and (2) better coordinated NDA and SDA fleet movements. Beyond the tens of millions of dollars in operating cost savings, two fewer aircraft are needed in the integrated solution than in the sequential solution. This is significant because annual ownership costs for aircraft measure in the millions of dollars.

In all scenarios, savings attributable to the NDA operation are small or nonexistent, whereas savings attributable to the SDA operation are large, reflecting the carrier's use of the Armacost, Barnhart and Ware (2002) optimization approach to design their NDA network, but not the SDA network.

We acknowledge that some operating requirements are not considered explicitly in our models. The staging of package arrivals at hubs is one example. Hence, the savings reported here might not be fully realized.

5 Summary

In this paper, we adapt the Armacost, Barnhart and Ware (2002) model to solve the integrated next-day and second-day express shipment service design problem, and present a new solution approach designed for large-scale problems. Our disaggregate information-enhanced column generation approach is shown to generate many fewer columns and help reduce IP solution time significantly. By solving the *integrated* NDA-SDA problem, we demonstrate potential savings of tens of millions of dollars.

We make the following observations about column generation approaches. First, high quality columns, that is, columns that are likely to be present in the optimal solution, and fewer generated columns can be achieved if interactions among columns are considered. This point is seen by comparing the performance of the naive and disaggregate information-enhanced column generation approaches. Second, better convergence and fewer generated columns can be achieved if a column in the restricted master problem includes fewer decisions. Specifically, in disaggregate information-enhanced column generation, each column in the *RMP* represents a single demand composite variable, indicating the selection of a small number of aircraft and routes. In contrast, each column in the *RMP* in aggregate information-enhanced column generation represents decisions for all aircraft routes at a hub.

In the service network design problem, using the disaggregate information-enhanced column generation approach, we generate a *set of columns* representing a solution to a *sub-network* of the overall network design problem. This approach greatly reduces the total number of columns generated, and is efficient in identifying columns that are likely to be in an optimal solution. We can extend this idea to other classes of problems. For example, in the multi-commodity network flow problem, we can establish at each iteration, a feasible flow in part of the network, instead of a single commodity flow.

References

2. Barnhart, C., R. R. Schneur. 1996. Air network design for express shipment service. *Oper. Res.* 44(6) 852-863.

3. ILOG CPLEX 6.5 User's Manual, 1999.

4. Dantzig, G. B., P. Wolfe. 1960. Decomposition principle for linear programs. *Oper. Res.* 8 101-111.

5. Grtinert, T., H. J. Sebastian. 2000. Planning models for long-haul operations of postal and express shipment companies. *Eur. J. Oper. Res.* 122(2) 289-309.

6. Kim, D., C. Barnhart, K. Ware, G. Reinhardt. 1999. Multimodal express package delivery: a service network design application. *Transportation Sci. 33(4)* 391-407.

7. Krishnan, N., C. Barnhart, D. Kim, K. Ware. 2002. Network design for express shipment delivery. *Comput. Optim. Appl.* 21(3) 239-262.

8. Kuby M., R. G. Gray. 1993. The hub network design problem with stopovers and feeders: the case of Federal Express. *Transportation Res.* A 27A(1) 1-12.

9. Leung L., Cheung W. 2000. An integrated decision methodology for desgining and operating an air express courier's distribution network. *Decision Science* 31(1) 105-126.

10. Salomon Smith Barney. 2000. *Equity Research: United Parcel Service Inc.*

11. Shen S. 2004. *Logistics Service Network Design: Models, Algorithms and Applications.* PhD dissertation, Massachusetts Institute of Technology.

12. UPS 2002 Annual Report.

13. UPS 2003. *Terms and Conditions of Service.*