Patterns of Freight Flow and Design of a Less-than-Truckload Distribution Network

A Thesis
Presented to
The Academic Faculty

by

Devang Bhalchandra Dave

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Industrial and Systems Engineering
Georgia Institute of Technology
May 2004

Patterns of Freight Flow and Design of a Less-than-Truckload Distribution Network

Approved by:

Dr. John J. Bartholdi, III, Chair

Dr. Eva K. Y. Lee, co-Chair

Dr. C. John Langley Jr.

Dr. Mathieu W. P. Savelsbergh

Dr. Chelsea C. White, III

Ms. Teresa Trussell

Date Approved: 12 April 2004

To my parents,

Ela and Bhalchandra

ACKNOWLEDGEMENTS

A lot of friends and well-wishers have crossed my path. It may not be possible for me to thank each of them individually in here. Besides the few people mentioned here, I would like to extend my deepest gratitude to all of those who helped me be what I am today – directly or indirectly.

Words fall short for me to express my feelings for my parents. My Ph.D. endeavor was as much their dream as was mine and I take the liberty to say that their countless acts of sacrifices and encouragement have been rewarded. Thanks for the words of encouragement when I needed them the most. My sister and her family have always been there for me – providing long escapades from Atlanta, just to relax and with no chores expected in return.

I was extremely fortunate to work under Dr. Bartholdi and Dr. Lee. They have offered so much more beyond their technical expertise – things which I will never find in a book. I would like to thank them for the times they believed in my abilities when I myself did not. And for letting me grow not just as a researcher but also as a person. I will be ever grateful to Dr. Bartholdi for providing me financial support during my entire stay at Georgia Tech and for the several analogies between my dissertation work and Galileo's. I think I did build a "telescope", and a pretty good one at that. Dr. Lee was always there not only to help me with technical expertise but also sharing with me fascinating facts, especially her stories of bird-intelligence. I would like to express my deepest gratitude to both of my advisors for making my stay at Georgia Tech the most challenging, enjoyable and memorable.

I would like to thank Dr. C. John Langley, Dr. C. C. White, Dr. M. W. P. Savelsbergh and Ms. Teresa Trussell, Manager of Operations Research at Yellow Transportation, Inc. for their willingness to serve on my committee and for their valuable feedback.

I would also like to thank the staff at the School of Industrial and Systems Engineering and the Interactive High Performance Computing Laboratory at the College of Computing for providing me access to their vast computational resources. I would also like to express my thanks to Ilog, Inc. for the free CPLEX licenses, much needed for my computational research.

I am very grateful to Caliber Technology, The Logistics Institute at Georgia Tech and its sister organization, The Logistics Institute Asia Pacific in Singapore, the Office of Naval Research (grant #N00014-89-J-1571) who funded this project and without which I would have been unable to pursue

my dream. I would also like to thank Georgia Tech's Trucking Industry Program, which is a member of the Sloan Foundation's Industry Centers Network, for partial support of my research.

I was fortunate to meet Anand Ramesh who beyond being a great friend helped me save the environment (and also some money) by car-pooling. Alisa Kongthon was another great person I stumbled across. Thanks for the several hours of listening, encouraging and sharing experiences. They were monumental for my growth. (And I am still planning on the 5–year reunions beginning 2002.) I would like to thank Ralph Mueller all his help – be it fixing my car, helping me out when I locked myself out of my car and house or sitting in the car the first time I drove on the interstate.

My roommates, Arpan and Aisha, have been very considerate during my entire stay with them. The marathon discussions on religion and other topics will be one of the million fond memories. Both the Nikhils – Buddhiraja and Patil – have been more of mentors than friends. I really appreciate all the help you have offered without which my life would have definitely been noticeably tougher. I was glad to get to know the very many wonderful people at Trikone-Atlanta. Special thanks go to Altaf, Chenoo and Janak for being great friends.

Running helped me maintain my sanity. Had I not been able to maintain my running it would have been a longer search to figure out an alternative venting mechanism. I would like to thank Joseph Hurley whose talks of encouragement and helped my conquer my inertia. And many thanks to my running partners – Mai and Paul Brooks who pushed and slowed me down when needed. Thanks for braving the cold and rain with me.

I have always taken for granted the friendship that Prashant Rao had to offer and last but not the least, I would like to express my thanks to Mr. Ashok Bhave, my neighbor who sparked my interest in Industrial Engineering and encouraged and guided me in my application process for graduate school.

There was a period in my Ph.D. endeavor that was extremely difficult in my personal life. I was extremely fortunate to be surrounded by truly amazing people who helped me emerge a stronger person at the end of those years. On a quiet Sunday afternoon when I will be looking back at my life all of you will be on my mind.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

A less-than-truckload (LTL) carrier typically delivers shipments less than 10,000 pounds (classified as LTL shipment). The size of the shipment in LTL networks provides ample opportunities for consolidation. LTL carriers have focused on hub-and-spoke based consolidation to realize economies of scale. Generally, hub-and-spoke systems work as follows: the shipment is picked up from the shipper and brought to an origin terminal, which is the entry point into the hub-and-spoke system. From the terminal, the freight is sent to the first hub, where it is sorted and consolidated with other shipments, and then sent on to a second hub. It is finally sent from the second hub to the destination terminal, which is the exit point of the hub-and-spoke system.

However, the flow of shipments is often more complicated in practice. In an attempt to reduce sorting costs, load planners sometimes take this hub-and-spoke infrastructure and modify it considerably to maximize their truck utilization while satisfying service constraints. Decisions made by a load planner may have a cascading effect on load building throughout the network. As a result, decentralized load planning may result in expensive global solutions.

Academic as well as industrial researchers have adapted a hierarchical approach to design the hub-and-spoke networks: generate the hub-and-spoke network, route shipments within this hub-and-spoke network (generate a load plan) and finally, balance the empty trailers. We present mathematical models and heuristics for each of the steps involved in the design of the hub-and-spoke network. The heuristics are implemented in a user-friendly graphical tool that can help understand patterns of freight-flow and provide insights into the design of the hub-and-spoke network. We also solved the load planning sub-problem in a parallel computation environment to achieve significant speed-ups. Because of the quick solution times, the tool lays the foundation to address pressing further research questions such as deciding location and number of hubs.

We have used data provided by Roadway Parcel Services, Inc. (RPS), now FedEx Ground, as a case-study for the heuristics. Our solutions rival the existing industry solutions which have been a product of expensive commercial software and knowledge acquired by the network designers in the industry.

CHAPTER 1

LESS-THAN-TRUCKLOAD FREIGHT OPERATIONS

1.1 Introduction

Trucking companies generally specialize in one of the following types of shipments:

1. truckload (TL) shipment
2. less-than-truckload (LTL) shipment

The Interstate Commerce Commission (ICC) defines LTL shipment as one that weighs less than 10,000 pounds, while a TL shipment is one that weighs more than 10,000 pounds. The ICC does not categorize firms as TL or LTL; only shipments are categorized. TL freight is usually an individual shipment from its origin to its destination in a single trailer. This freight does not require any intermediate handling or sorting. A LTL carrier usually handles LTL shipments, though it may also provide TL services. To make economic use of the trailer LTL shipments are usually consolidated (see section 1.3.1).

Certain LTL carriers specialize in parcel delivery. A package carrier is a specialized motor carrier that restricts itself to freight generally less than 50 pounds. The US parcel delivery industry includes regional carriers such as AB Express, Inc. in mid-west US, as well as USPS, FedEx and UPS which serve the entire US.

Case Study

This research was motivated by a project with Caliber Logistics which was subsequently acquired by FDX Corporation and now operates as FedEx Supply Chain Services. The network data provided was for the year 1995 for Roadway Parcel Services, Inc. now FedEx Ground Inc. and throughout the research this data will be used as a case-study. In the remainder of this thesis, this data will be referred to as *FedEx data set*.

However, it should be noted that though the research is based on single data set it is representative of the entire LTL industry and the results presented may be extended to any general LTL carrier with comparable shipment size distributions.

1.2 Operating as a Truckload Carrier

A LTL ground package carrier delivers freight from the point of origin to the point of destination. If there are no service restrictions the optimal policy for dispatching a trailer would be "go when full". Under this policy the trailer utilization would be maximum and each shipment would be routed from the origin to the destination directly. If there is insufficient freight to fill a trailer on a origin–destination route the shipments would be held back at the origin terminal waiting for additional freight. However, it is an impractical strategy because there are no shippers who would be willing to let their freight wait at the terminal with no guarantee on when the freight will be shipped to the consignee. Since service provided by the carrier is extremely important in attracting customers and maintaining the market share, to implement the "'go when full" policy is a very bad strategy.

If there is sufficient freight to fill a trailer either entirely or almost entirely then the full truck would be dispatched directly from the origin to the destination without compromising on service. However, if there is insufficient freight to fill the truck completely then sending a partially filled truck might be an inefficient and expensive way to operate.

A shipment is the entire freight from a particular point of origin to a particular destination. Freight from several shippers still constitute a single shipment. There are various measures for the size of the shipment. General LTL carriers may measure shipment size by either weight or volume. Parcel carriers, such as FedEx Ground, measure shipment size by the number of packages. Typically, most of the shipments are usually of a very small size. For the FedEx data set, we estimated the average size of a FedEx Ground shipment to be less than 1% of truck capacity. Figure 1 shows the distribution of the shipment sizes entering the FedEx Ground distribution network. Over 80% of the shipments utilize less than 1% of a truck capacity.

Based on this shipment information one of the questions that strikes immediately is "If LTL networks implement TL operating strategies how bad can it get?" If consolidation was not allowed then, similar to TL operations, for each of these shipments a truck would be routed directly from the origin terminal to the destination terminal. The only costs to be then considered are the transportation costs for routing trucks on each of these direct routes. In a hub-and-spoke based consolidation model, sorting costs account for a significant portion of the total costs and have to be considered. We used our mathematical model (described in Chapter 2) to generate a hub and spoke based consolidation network for the FedEx Ground data and estimated the operating costs. Based on the FedEx Ground data we estimated the TL based operating costs to be approximately 64 times the hub-and-spoke based operating costs. The operating costs include transportation and

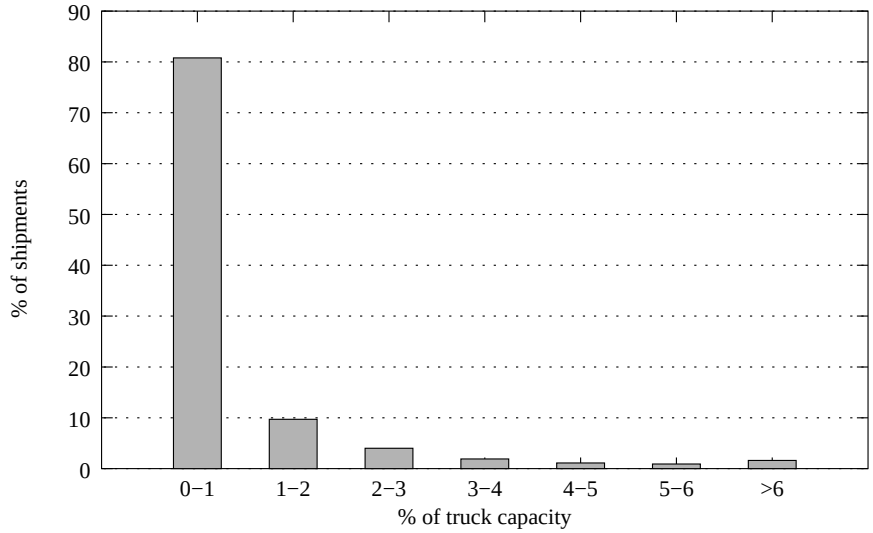


Figure 1: Over 80% of the LTL shipments occupy less than 1% of the truck capacity (= 2 trailers)

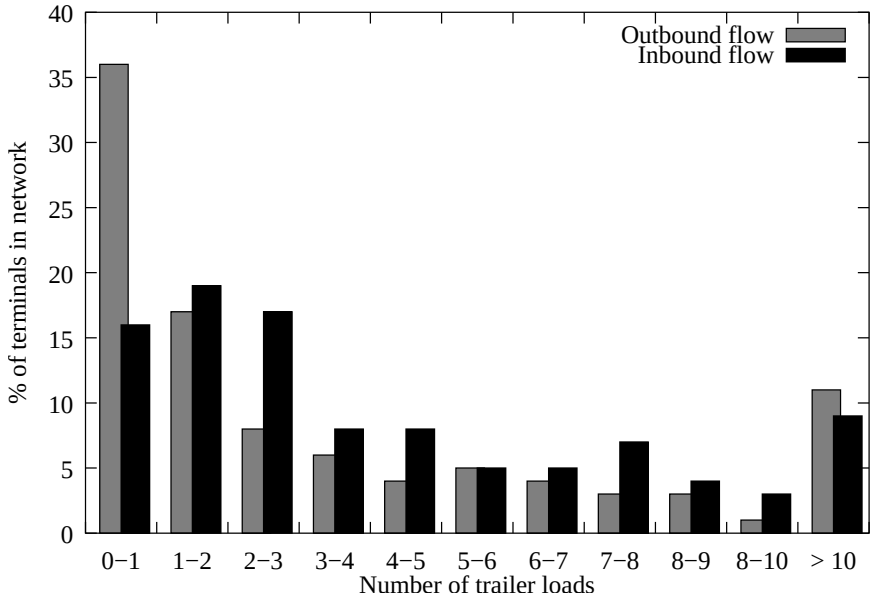


Figure 2: Over 15% of the terminals in the network receive less than a trailerload shipments whereas almost twice as much send out less than a trailerload shipments totally.

sorting costs.[1] This is sufficient incentive to consolidate freight.

It is important to bear in mind that consolidation can only deteriorate service levels. Freight to be consolidated is now routed through one or two hubs increasing the transit time. Also, every time freight is sorted at a hub, handling and sorting times add to the transit time. Typically, handling at one hub adds about a day to the delivery time to the consignee [Braklow, Graham, Hassler, Peck, and Powell, 1992].

To increase trailer utilization, freight is consolidated so that on a majority of the routes the truck is as full as possible at the time it must depart. Current LTL freight network designs have focused on hub-and-spoke based operations.

1.3 Hub-and-Spoke Network

Figure 3 illustrates a typical hub-and-spoke network. A terminal is connected to a hub by a "spoke", also called an assignment. All the hubs are connected and the collection of all the hubs with their spokes form the hub-and-spoke distribution network.

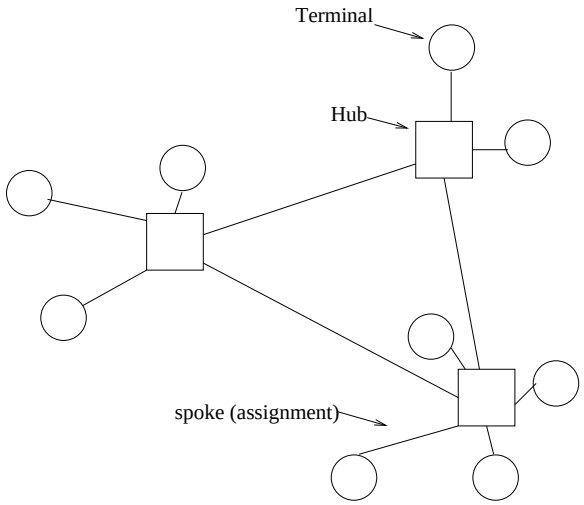


Figure 3: An illustrative hub-and-spoke network

In this section we present the reasons why LTL carriers prefer hub-and-spoke distribution models for their operations. Then we briefly explain the operations that are involved in a hub-and-spoke based network.

[1]We neglected fixed sunk costs (real estate investment in hubs/equipment) and/or other variable costs (rental costs, insurance, salaries of employees at hubs, etc.).

1.3.1 Economics of Hub-and-Spoke Network Operations

The *marginal cost* of a package in a shipment is defined as the incremental increase in cost by adding a package to that shipment. If a trailer on a route has excess capacity this package can be loaded onto that trailer and the only increase in cost is the handling cost at the origin and destination[2]. However, if the trailer is full, or if no trailer is assigned to the route yet, to ship the additional package an entire additional trailer has to be assigned on that route. And in this case the marginal cost of the package is not negligible. Since the marginal cost of a package in a shipment on a trailer with excess capacity is very small, any policy that increases package densities reduces average cost per package per trip. For this reason for truckload carriers direct point-to-point operations can be justified economically. However, the shipment size in a parcel delivery industry provides ample opportunities for consolidation (network economies). With the spokes feeding the packages to the hubs, the hub-and-spoke network configuration increases package densities on the inter-hub routes. The economics of small parcel LTL industry tends to prefer hub-and-spoke network over direct point-to-point operations.

The cost of moving a trailer from one terminal to another is not negligible. This transportation cost can be considered as a "fixed" cost. Consider the situation when an additional terminal is added to the network. In a point-to-point delivery system, this terminal must be connected to all other terminals which would incur fixed costs directly proportional to the number of existing terminals in the network. However, in a hub-and-spoke network only two additional lanes have to be operated — one from the terminal to its hub and second from the hub to the terminal. It is now obvious why that TL network is usually a point-to-point system. Since the marginal cost of additional shipment in TL operation is not small any route other than a direct route from origin to destination increases transportation costs [Starr and Stinchcombe, 1992].

1.3.2 Typical Hub-and-Spoke Based Consolidation

Typically, freight is picked up from the point of origin and transferred to a local consolidation facility known as *terminal* or *satellite*. The network has established terminals that are the entry/exit points of the freight in the distribution network. The terminal collects freight from all the points of origin within its area of control. Collection of the freight from their points of origin is known as *local pickup*. All the freight from a particular origin terminal to a specific destination terminal is a *shipment*, also

[2]Assuming that it takes about 1 minute to unload and then load a packages, based on hourly wages of a package handler to be $10, we estimate the handling charges to be about $0.20/package.

known as a *flow*[3]. To minimize costs, the packages in a shipment may have different routes from the origin to the destination (see section 3.3). The shipments are brought to the terminal and divided into *inbound freight* (that which is to be delivered locally) and *outbound freight* (that which is to be delivered outside the region served by the satellite). The outbound shipments are consolidated at the terminal and typically delivered to central terminals (also known as *hubs*). The movement of trucks on the spokes of a hub is known as *shuttle operation*. Shuttle movements are usually a few hundred miles, usually under 350 miles. At the hub the shipments are sorted and loaded onto trucks destined to the hub serving its destination region. From the destination hub, the shipments are sent to the terminals serving the respective destination cities. From the terminals the shipments are delivered to individual consignees. Figure 4 (page 7) illustrates a typical route of a shipment [Wyckoff, 1974].

1.3.2.1 Direct loads and direct runs

Freight flow is not as simple as suggested above, mostly due to ad hoc modifications to further reduce handling in the hubs and to improve service. Both of these goals can be achieved by "intelligently" consolidating freight. For example, suppose that many shipments destined to terminal j arrive at hub i. If a truck can be sufficiently filled up by these shipments then it can be dispatched directly to terminal j bypassing the second hub. This is called a hub-to-terminal direct load. Similarly, several shipments, destined to the same geographical region, originating at terminal i can be loaded onto a single truck and sent to the second hub bypassing the first hub. This is called a terminal-to-hub direct load.

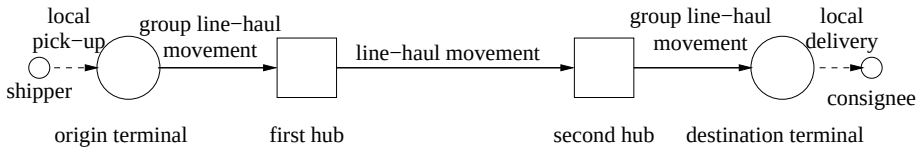
A pair of trailers may be pulled directly to their destinations if nearly full because the rig and the driver would then be fully utilized and there is no need for additional handling. This is known as a *direct run*. In the FedEx Ground network this happens very infrequently.

1.3.2.2 Consolidation at freight terminals

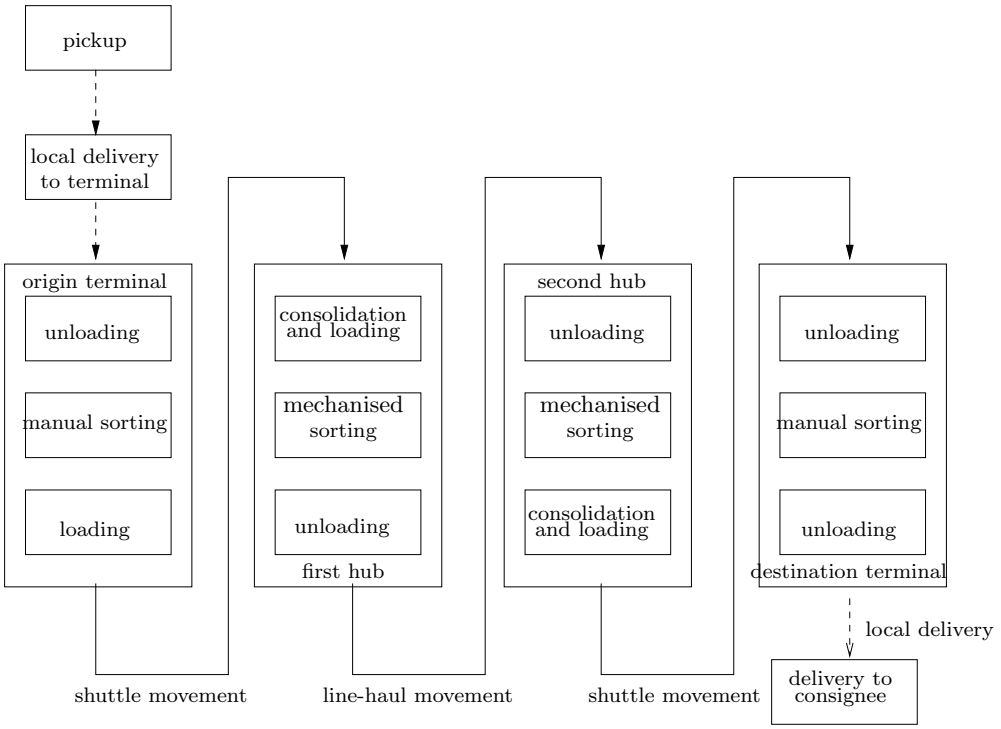
Other patterns of freight flow are possible too. For example, an Overnight-Transfer-Point (OTP) is a hub that handles special freight that has been guaranteed for overnight delivery. This freight is sent from a satellite terminal to an OTP and thence to its destination satellite. Design of the overnight delivery network is not encompassed in this research.

Another pattern of flow is to route freight through a Relay Point, which is a terminal with no sorting capability. Essentially it is a place to park and/or swap trailers and where drivers can sleep.

[3]In the industry, shipment refers to freight between a particular origin–destination pair from a particular shipper and flow refers to a collection of shipments. Since we do not classify shipments based on customer information in this research, we will use the shipments to denote all the freight between a particular origin–destination pair.



(a) Shipment route through the hub-and-spoke network



(b) Detailed flow of a shipment from the sender to consignee

Figure 4: Typically a shipment is routed through two hubs. In the case that the origin and destination terminals are assigned to the same hub it is routed only through that hub.

In the US, these are occasionally used out west, where inter-hub distances can be quite long. Since it does not have a substantial impact on the network and freight operations this is not considered in this research.

Finally, Spider Leg routes are those in which a truck/trailer from a terminal picks up additional freight from another terminal along the way to its assigned hub. This freight routing pattern is very unusual in the FedEx Ground network. Only very special cases of terminals have spider legs. Spider legs are ignored in this research.

1.3.2.3 Head loads

Head loads are concerned with the way shipments are loaded on to a trailer. If a trailer is known to be traveling all the way to satellite terminal i then any freight bound for terminal i will be stored in the front of the trailer so that it need not be handled again en route. Head loads are generated in an attempt to further reduce handling at hubs but these are of less significance. In this research we do not consider head loads.

1.4 Terminals and Hubs

A terminal is essentially the entry or exit point of the hub-and-spoke system. A particular geographic area is serviced by each terminal. The terminal manages local pick-up from shippers and dispatch to the consignees within its service area. Since the local pick-up and dispatch is not considered a part of the hub-and-spoke network it is beyond the scope of this research and the terminals, rather than shippers and consignees, are considered to be points of origin/destination for the freight. Much work has been done to improve the local pick-up and delivery operations. The interested reader may refer to Ball, Magnanti, Monma, and Nemhauser [1995].

Once freight arrives at the terminal, it is checked and after the administrative procedures are completed they undergo a local sort. The inbound shipments are taken to the appropriate loading zones for local delivery (usually the next morning) and the outbound shipments are taken onto the appropriate trucks (or trailers) to be sent to a hub. Most terminals do not have automated sorting capabilities. Freight is sorted manually while transferring it (either using two-wheel trucks, forklifts, drag lines or conveyor belts) from where it was unloaded to the appropriate loading dock.

A terminal has one or more *load doors*. Load doors are essentially loading docks that are available so that trucks (or trailers) can back up to the load door for loading or unloading the freight. The number of load doors is essentially the maximum number of trucks (or trailers) that can be simultaneously loaded (or unloaded) at the terminal. More load doors may keep the terminal less

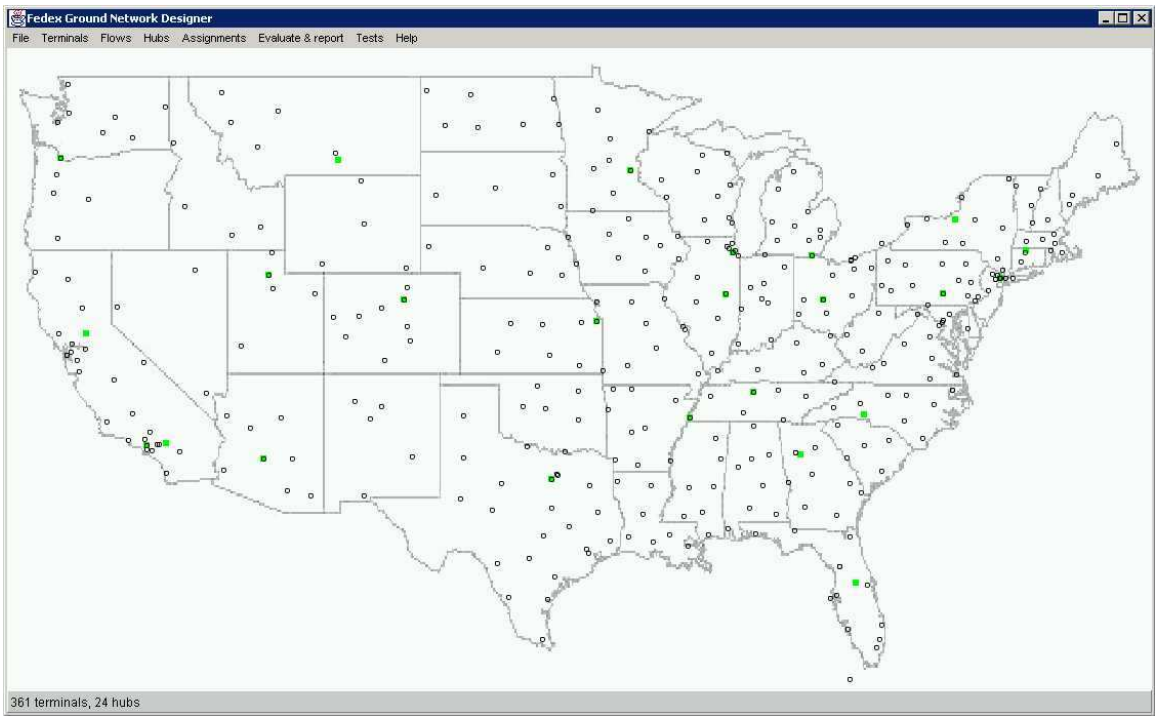


Figure 5: Terminals and hubs for FedEx Ground in mainland USA. The empty circles represent the terminals and the squares represent the hubs.

congested but may result in inefficient utilization of the load doors. On the contrary, fewer load doors result in higher load door utilizations but may keep the freight waiting at the terminal. This may result in congestion and inefficient freight movement at the terminal and increase the chances of health hazard and damage to the freight.

A hub, also known as *breakbulk terminal*, is a consolidation center in the hub-and-spoke network. Since many terminals are assigned to a hub, the freight volumes handled at a hub are significantly larger than that handled at a terminal. Hubs are provided with automated sorting equipment when economically justified.

Figure 5 shows the terminals and hubs for FedEx Ground in mainland USA.

1.5 Freight

The nature of freight handled is perhaps one of the most important considerations in the design of motor carrier operations. The nature of freight plays a very important role in

- selection of transportation equipment: Depending upon the type of freight moved either single trailer or twin trailer trucks may be preferred. To transport bulky freight twin trailer may tend to be under-utilized. Freight in a regular shape can fill up a trailer better than irregular

shaped freight.

- sorting methods: The nature of freight dictates whether sorting will be manual or automated. With the current sorting technology available regular shaped packages can be efficiently sorted automatically. However, irregularly shaped freight cannot be sorted automatically and has to be sorted manually.

This research focuses on small parcel shipments. Most of the shipments accepted by FedEx Ground are severely constrained by the weight, shape and size. FedEx Ground usually uses twin-trailer trucks and all of the hubs are automated for sorting.

For purposes of tactical planning, we will assume that all the packages are homogeneous. This seems to be a reasonable assumption because of the strict restrictions on the dimensions of the packages accepted.

1.6 Transportation Equipment

The LTL carrier industry uses trailer(s) pulled by truck tractors. This combination is called a *truck trailer*. The advantages of using truck trailers as opposed to single unit trucks are

- it is not necessary to unload the vehicle to release the power unit for other work
- increased maneuverability of a truck trailer as compared to a single wheelbase truck of equal capacity.

Because of changes in regulation over the past decade, a more noticeable trend in the LTL motor carrier industry is the use of *twin trailer trucks*. A single tractor pulls two 28-ft trailers (*pups*) in tandem instead of the one 45- or 48-ft trailer. This combination is also known as *double-bottom*. Figure 6 shows a twin-trailer truck used by FedEx Ground.

Figure 6: Twin-trailer truck used by FedEx Ground

Some of the advantages of twin trailers over single trailer trucks are increased cubic capacity, better response to freight and reduced operational costs as listed below:

1. **Local pickup and dispatch:** A single trailer can be towed by the tractor on local pick-up and dispatch route while the other trailer is being loaded/unloaded at the dock.

2. **Loading direct trailers:** Avoiding sorting at a hub reduces the delivery time to consignee by about a day. Consider the following example: Miami sends half a truck worth of freight to Boston. In case of a single 48ft. truck, when this truck reaches a hub, say Orlando, all the freight will be unloaded and sorted. The crucial information that half of the truck is filled with freight destined to Boston now becomes irrelevant. Instead, if two 28ft. pups are used, this information can be retained by filling all the Boston bound freight in a single trailer and the other freight in the second trailer. The Boston bound trailer is closed and locked. When this trailer reaches Orlando, it is not be opened and sorted. This reduces sorting costs and the associated times.

3. **Smoother terminal/hub operations**. At a load door, a trailer destined for a hub, say Miami, can be first loaded. While the second trailer waits for additional Miami bound freight, a trailer destined for another hub, say, Boston can begin loading. Thus freight moves faster in and out of the docks and results in smoother terminal/hub operations. If a 48ft. truck is used, freight destined for Boston will sit at the terminal/hub until additional freight arrives to fill the Miami-bound truck.

Triple-trailer equipment is used in some states where its legal. Since single-trailer and triple-trailer equipment is not commonly used we will assume for the purposes of this research that only twin trailer equipment is used and is homogeneous, that is, all trailers are identical.

FedEx Ground estimated 1000 packages fit into a pup, based on historical averages and in the rest of this research, for numerical computations and results, we will be considering the trailer capacity to be 1000 packages so that the capacity of a truck (two trailers) is 2000 packages.

1.7 LTL Network Design Problem

Given that a hub-and-spoke model is to be used for consolidation, the network has to be designed in a way that a bad design does not limit the profitability of the operations. In the remainder of this chapter we first list the different steps involved in designing/building the network and then briefly explain why these steps are relevant in the design process.

The design of the hub-and-spoke based LTL network involves the following:

1. Network design

(a) Physical network design (location and design of hubs and terminals)

(b) Service network design (determining linkages between the hubs, truck schedules, etc.)

(c) Hub-and-spoke network design

2. Routing shipments through the network (load planning).

3. Routing empty trailers.

Though all of these are highly inter-related the network design procedure can be viewed as a three step hierarchical procedure.

Several other operational issues such as truck/trailer scheduling, driver scheduling, delivery and pickup time window constraints are also important but are currently beyond the scope of this research.

Each of these steps is explained in detail in the following sections.

1.8 Physical Network Design

The physical design of a network involves decisions regarding locations of terminals and hubs and allocation of capacity at the hubs.

1.8.1 Terminal Locations

In our experience, though some kind of economic decision analysis is usually involved in determining the location of the terminals, in practicality, location of terminals are strongly influenced by pre-existing terminal locations of the competitors. This is because of the intense competition among the domestic carriers, who provide competition based on geographical area coverage and costs, as well as the regional carriers who provide competition based on faster service. Customer (shipper) demand forms one of the most important inputs to the analysis. In a good economy, in order to attract customers (shipments) and in a customer-driven industry, most terminals are located near the influential customers. This, more than often, results in excess number of terminals than that are needed. As the market gets over-capacitated with service providers and competition increases, resulting in very low profit margins, the problem of terminal locations changes from adding new terminals (to the existing network) to deleting terminals (from the existing network) [pers. com. Trussel, 2002].

However, for our research, motivated by FedEx Ground, the locations of the terminals are provided and is part of the data.

1.8.2 Hub Locations

The design of the hub network primarily addresses how to distribute sorting capacity within the network. Sorting capacity plays a very important role in the network design as it affects the service provided by the carrier.

1.8.2.1 How the LTL network grows incrementally

Typically, in the parcel delivery industry the trend for a company has been to grow from a small market regional carrier into a large domestic carrier. For a small regional carrier, when the network is small, with maybe one hub, locating that hub approximately based on single facility location analysis seems to be a good strategy. However since most customers prefer their shipments to be handled by one carrier, instead of the regional carrier inter-lining freight with other carriers, and to exploit economics of scale, the regional carrier considers expanding service into adjacent geographical areas [Braklow et al., 1992, pg. 149]. As the market expands into an adjacent region, approximate location of another hub may again seem an obvious solution – single facility layout problem for the extended region.

As the network expands operations into a larger region, say domestic United States, its operations and profitability will be severely constrained by these parochial decisions. The motor carrier could be operating at the best it can for the given network configuration. But its profits could be increased by modifying the network. Since the cost of shutting down (closing) a hub are high sunk costs, usually closing one hub and reopening another is not common.

To allocate sorting capacity within the network we need to decide the following

1. **Number of hubs**
2. **Location of hubs**
3. **Hub capacities:** Excessive sorting capacity is undesirable since it implies investment in an under-utilized expensive equipment. A hub operating near its full capacity is also undesirable since it leads to increased costs due to congestion. A hub operating at its capacity may not be able to absorb variations in freight demand. Also a highly utilized hub may be inflexible in accommodating seasonalities in shipment demand.

It may be the case that though the cumulative sorting capacity may be sufficient for the network it may be inefficiently allocated amongst the hubs. This may result in one hub being congested while the other hub has extra capacity. This imbalance may also lead to inefficient freight routing if hub utilization is an important performance measurement criteria for the network operation.

This step also has to address the scenario when the market expands within the region, that is, the average size of shipments has increased. One common problem is to decide whether to expand the capacity of a currently existing hub or to build new hub *close by* to service that region.

Hub location is beyond the scope of our current research. However, one of the goals of this research is to lay the foundation to address the problem of hub location and to help the LTL carrier improve the profits by redesigning the network. Hub location is briefly discussed in Chapter 7.

1.9 Service Network Design

Once a carrier decides to offer service between two locations it commits to the customers some kind regularity in dispatching schedules between those locations.

In our approach, we assumed daily services between each terminal and its assigned hub. Also we assumed daily services between all the hubs. Our basic assumption was that a terminal will send out at least one truck to its hub at the end of the day. However, in practical scenarios, a terminal may store freight overnight and send it with the additional freight the next day. Though this may deteriorate service it may substantially reduce transportation costs. The service network design problem looks at aggregated freight flows over a longer time period, such as a week, and involves decisions regarding the service frequency of trailers over that time period. Powell [1986], Lamar and Sheffi [1987], Powell and Sheffi [1989], Roy and Delorme [1989], Powell and Koskosidis [1992] have considered designing the service network shipment routing with minimum frequency constraints between hubs and for direct loads.

In this research we will not address this issue. Since our research is aimed to help (re)design the network at a tactical level, a solution to the service design problem can be yielded by providing maximum service to the customers and then, based on that output, reducing service frequencies on low volume lanes. FedEx Ground did not think it was critical to constrain their network as per the current operations [pers. com. McMurtry, 2000].

1.10 Hub-and-Spoke Network Design

After the physical network and the service network have been established, we have to design the hub-and-spoke network. We need to decide which hub will be a *primary* or *parent* hub of a terminal. The idea of *assigning* (connecting via spokes) a terminal to a parent hub is that, usually, all the less-than-truckload shipments will be first sent to the parent hub for consolidation. The assignment problem is not trivial since the assignment of one terminal may affect the assignments of other terminals.

Based on the operating policies the following classification of assignments can be made

1. *Single assignments*: A terminal is assigned to a single hub. Since all the shipments in and out of the terminal will be sent to its parent hub, single allocation provides maximum utilization of the trucks on the shuttle and longhaul segments but at the cost of higher transportation cost and costs of sorting the shipments at the hub(s). Since all the shipments at a terminal will be sent to its parent hub, single assignment policy also simplifies planning and control [O'Kelly and Miller, 1994].

2. *Multiple assignments*: A terminal may be assigned to two or more hubs by exploiting the freight patterns. Load planners, generally, seek opportunities provided by multiple assignment policy to minimize operational costs.

In the single assignment policy load planners may not take advantage of certain freight patterns to reduce sorting and/or transportation costs whereas multiple allocation can save transportation costs by tailoring the selection of hubs to the eventual destinations of the flows being shipped from an origin node thus reducing the distance traveled. There is a trade-off between reduced truck utilizations and reduced distance traveled by the truck. Most package carriers adopt an hybrid policy.

The design of the hub-and-spoke network is discussed in Chapter 2.

1.11 Shipment Routing

As explained in section 1.3.2 hub-and-spoke systems work as follows: the shipment is picked up from the shipper and brought to an origin terminal, which is the entry point into the hub-and-spoke system. From the terminal, the freight is sent to the first hub, where it is sorted and consolidated with other shipments, and then sent on to a second hub. It is finally sent from the second hub to the destination terminal, which is the exit point of the hub-and-spoke system.

However, the flow of shipments is often more complicated in practice. In an attempt to reduce sorting costs, load planners sometimes take this skeletal hub-and-spoke infrastructure and modify it considerably to maximize their truck utilization, while satisfying service constraints. Unfortunately, a load planner has a local perspective, and conflicting operating policies. For example, a load planner at the origin terminal may want to hold on to shipments so that he can collect sufficient freight to fill a truck and send it directly to the second hub, bypassing the first hub (terminal-to-hub direct load). However, the load planner at the first hub, for his part, may be planning to use that

freight to fill a truck to send directly to the destination terminal, bypassing the second hub (hub-terminal direct load). Thus a decision taken by a load planner may have a cascading effect on load building throughout the network. Therefore, decentralized load planning may result in unnecessarily expensive global solutions.

Our goal is to centralize shipment routing operations with a focus on reducing operating costs. Shipment routing is discussed in detail in Chapters 3–5.

1.12 Recirculation of Empty Trailers

Ideally a LTL network would like to balance the freight flows. However, since it may not be entirely possible to balance the freight flows LTL networks are designed to reduce the imbalance as much as possible. In order for operations to continue efficiently in spite of the imbalance, the empty trailers have to be routed from accumulation points to points of deficit. Routing empty trailers, also known as backhaul, is expensive and most LTL carriers aim to minimize the miles driven by empty trailers. In chapter 6 we propose models to route the empty trailers to minimize transportation costs.

1.13 Network Design Methodology

As mentioned earlier, the design of a LTL network can be viewed as a hierarchical problem. To integrate all of these sub-problems into a single model is challenging and our experience suggests that solving any of the sub-problems to near optimality is very difficult for any company of the size of FedEx Ground. Hence, it is a reasonable approach to not attempt to integrate all of the hierarchies but to use the procedure described below to design the network.

Step 1: Select location for the hubs, if not provided

Step 2: For these hub locations generate the hub-and-spoke network

Step 3: Once the hub-and-spoke infrastructure is generated, extract direct loads and route freight

Step 4: Route empties to balance movements of tractors and trailers.

Step 5: Estimate total operational costs

Step 6: If stopping criterion is satisfied then stop.

Step 7: If hub locations are to be decided then perturb the hub locations and go to step 2, else stop.

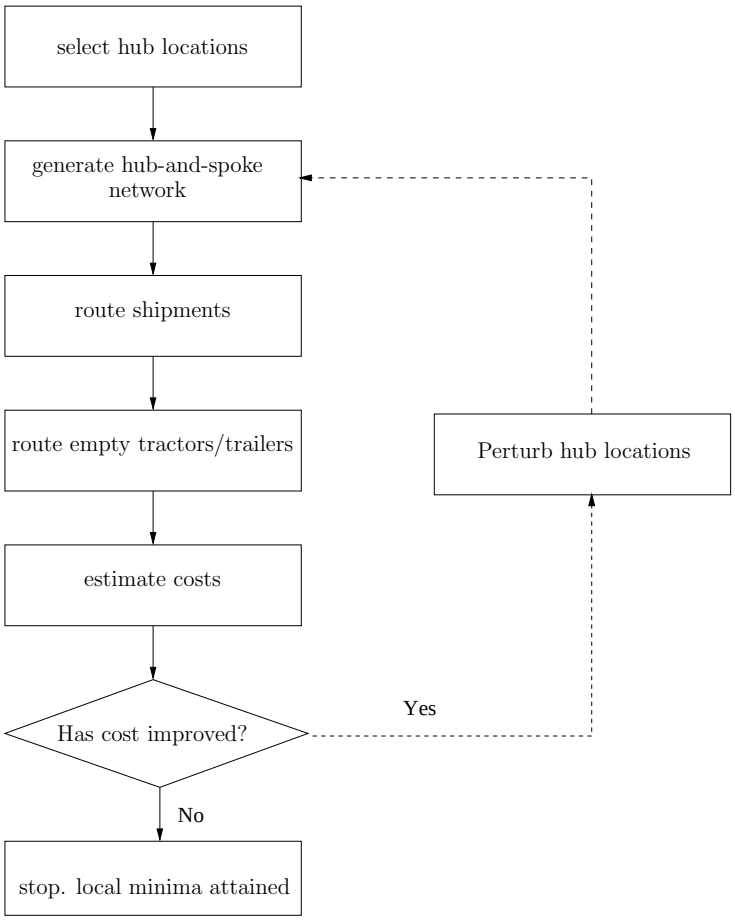


Figure 7: Solution approach for the designing the LTL network

CHAPTER 2

HUB-AND-SPOKE NETWORK DESIGN

2.1 Problem Description

To generate a hub-and-spoke network we assign every terminal to a hub. Assignment of a terminal to a hub determines the routing of most of the shipments originating from and destined to the terminal. Typically, every shipment is routed from the origin terminal to its assigned hub and then via the assigned hub of the destination terminal to the destination terminal. In a hub-and-spoke system, when direct loading (explained in section 1.3.2.1) is allowed, some of the shipments may bypass one or both of the intermediate hubs.

Assigning a terminal to a hub can have an cascading effect throughout the network as it utilizes an expensive and limited resource — sorting capacity — at a hub.

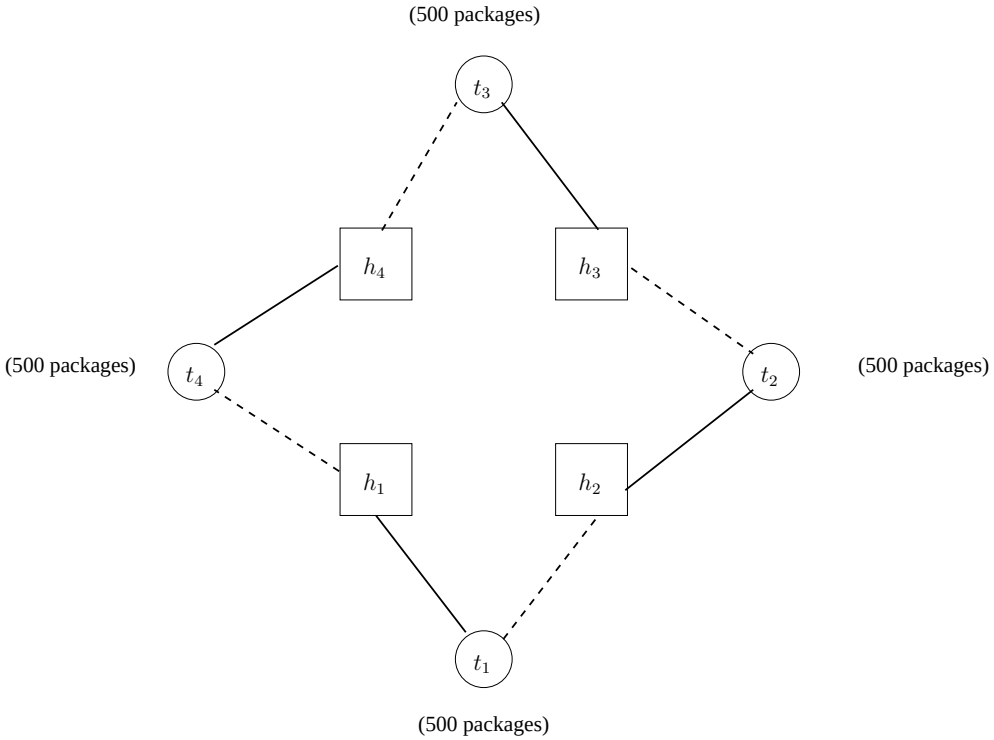


Figure 8: Because of the combinatorial nature, the assignment of one terminal affects the assignment of other terminals in the network.

Example 2.1 *Consider four terminals, $t_1, \ldots, t_4$, and four hubs, $h_1, \ldots, h_4$, in the network. Terminal t_i sends P packages to terminal t_{i+1}, $1 \leq i \leq 3$ and terminal t_4 sends P packages to terminal t_1. Each of the four hubs has sorting capacity of $2P$ available. This restricts only one terminal to be assigned to a hub, since each terminal sends and receives a total of $2P$ packages that need to be sorted.*

One possible assignment policy is to assign terminal t_i to hub h_i, shown by solid lines in figure 8. However, if terminal t_1 is assigned to hub h_2 then this affects the assignments of the remaining 3 terminals, shown by dashed line.

The assignment of a terminal to a hub also affects the transportation cost amongst all the line-haul movements since a terminal sends shipments to and receives shipments from almost every other terminal in the network. Based on the FedEx data set, a terminal sends freight to and receives freight from approximately 70% of all other terminals in the network.

In the remainder of the thesis, we will say that a terminal *communicates* with another terminal if it either receives or sends freight from that terminal. Also, the *intensity* of communication is directly proportional to the size of the shipments sent between those two terminals. That is, a terminal t has greater intensity of communication with terminal t_1 than with terminal t_2 if t sends to and receives from t_1 more packages than from t_2. The *resistance* of a hub to a terminal is measured in terms of the total intensity of communication between the terminal and all other terminals assigned to that hub.

Models to assign terminals to hubs have typically been notoriously hard combinatorial Quadratic Integer Programing (QIP) formulations, such as introduced by O'Kelly [1987]. The quadratic term determines the total flow between two hubs, which depends on the terminals assigned to each hub. Our approach is to develop a fast heuristic to generate the hub-and-spoke network. This seemed like a reasonable approach for two reasons: firstly, this was intended for tactical planning of the entire network. Rather than focusing on the minute operational details and incremental savings the objective was to have a broader, maybe approximate, understanding of the network and its operations. Secondly, the goal of our research was to provide reasonably fast solutions for what-if analysis of various scenarios. Solving the QIP for networks much smaller than the FedEx Ground network is computationally exhausting and because of limitations on solution times a heuristic approach is justified [O'Kelly, 1987].

To generate the hub-and-spoke network we implemented a greedy heuristic. We did not consider other meta-heuristics such as local search, tabu-search, simulated annealing and genetic algorithms.

The terminal assignment problem has a very large neighborhood to search. For a given feasible set of assignment we have,

$$\text{size of the neighborhood} = (\text{number of hubs})^{(\text{number of terminals})}$$

The greedy heuristic was designed to narrow the search region and also quickly construct a local optimum.

2.2 Input Data

Understanding the data and the parameters will be helpful in understanding the heuristic and any assumptions made.

Terminal data

For each terminal we require the following:

1. **Location:** The location of a terminal is specified in terms of latitude and longitude.

   The latitude and longitude is used to determine the great-circle mileage distance of the terminal from another terminal, if this distance is not available in the distance database, in which case an appropriate factor is applied for better approximation to the actual over-the-road distances.

   Latitudes and longitudes are also used for plotting the terminals and locations on a map.

2. **Timezone:** Each terminal's timezone is provided.

   The timezone is important for service constraints. Freight traveling west across a timezone may have an extra hour to make a sort at the destination hub. This means a hub may serve more distant terminals to the east than to the west.

3. **Number of load doors:** The number of load doors determine the number of direct loads that can be built at this terminal.

   A load door is a loading dock that is dedicated to a trailertraveling to a certain destination. Typically, a terminal sends most of its freight to its assigned hub unless it has sufficient freight to fill up a direct truck to any other hub/terminal. The truck waits at a load door to be loaded as and when freight arrives at the terminal. If all the load doors are occupied, even if there is sufficient freight to build a direct load, the freight will be loaded onto a truck destined to the assigned hub of the terminal, in order to avoid congestion and maintain smoother operations.

   For the heuristic, we use this piece of information only for accounting and reporting the final flows and costs and not as an active constraint. Rather, any terminals violating this constraint

can be viewed as potential candidates for re-design. We can justify this assumption since we are looking at designing the network rather than designing the operations constrained by the resources of the network.

Hub data

Besides the data provided for each terminal, each hub is provided with the following data:

1. **Sorting capacity (packages per day):** This is sorting capacity of the automated sorting equipment installed at the hub.

   When designing the network we would like to generate the hub-and-spoke network unrestricted by current design. However, this is a reasonable approach when modifications to current designs are not expensive. Sorting equipment is very expensive and adding capacity is usually a decision made at the strategic level. Hence, we cannot enforce capacity constraints passively. To see the effect of capacity constraints on the network, the heuristic has the option to either enforce the constraints or ignore them. The role that sorting capacity plays in network design is discussed in section 2.3.11.

2. **Sorting cost per package ($ per package):** Typically, the sorting cost is estimated based on the amortized cost of the sorting equipment.

   The current industry practice is to consider the sorting cost per package to be independent of the utilization of the sorting equipment at the hub. Hence, in our data we have the same sorting cost (25 cents) per package for all the hubs. However, some researchers consider the sorting cost to be a function of the hub utilization because as the utilization of the sorting equipment increases, it becomes the bottleneck in the hub operations and leads to congestion of freight traffic at the hub, which in turn disrupts the entire hub operations and increases operating costs. This cost increase can be viewed as an indirect cost rather than direct cost.

Shipment data

Each shipment is characterized by the following:

1. Origin terminal
2. Destination terminal
3. Number of packages

For the parcel delivery industry, the shape, weight and dimensions of the packages are severely constrained. The estimate of the trailer capacity is based on the historical average size of a package. So at the tactical planning level it is justified to assume that all the packages are homogeneous.

Network parameters

These global parameters capture the business rules implemented by the LTL carrier.

1. **Cost per truck per mile:** We assume a uniform cost for driving a truck per mile. Though this may be a function of how much the truck is loaded, the variation is insignificant from tactical planning perspective and assuming a uniform transportation cost is reasonable.

2. **Truck speed:** We also assume the truck speed to be uniform miles per hour over all the routes. Again, there may be slight variations regionally which are ignored for purposes of tactical planning.

3. **Maximum time to sort at hub:** This represents the service constraint imposed by the carrier. A package is tracked every time it is sorted at one of the hubs. One of the main reasons to try to sort a package within this time limit is to keep track of a package early on so that a warning can be issued for missing (or lost) packages without much delay.

4. **Trailer capacity:** Each trailer (pup) has a capacity. Since our research was motivated by a parcel carrier, we measured trailer capacity by the number of packages it carried. For the parcel carrier industry, the trailer capacity determined by the number of packages is tighter as compared to other dimensions such as weight and volume. However, for LTL carriers that handle more general freight, other measures such as weight and volume are widely used.

5. **Minimum direct load factor (β_{min}^{direct}):** This factor captures the operating policy for sending direct tractors/trucks. A truck will not be sent directly from a terminal to any other hubs than its assigned hub unless its is $100\beta_{min}^{direct}\%$ utilized. Since in our heuristic we do not search for direct loads besides the obvious terminal to terminal direct loads, we use this piece of data passively to issue a warning when reporting the terminal-to-hub direct loads after the terminals are assigned.

6. **Minimum load factor ($\beta_{min}^{routing}$):** This factor captures the operating policy for sending tractors/trucks from a hub to its assigned terminal and on longhaul routes. If a truck is less than $100\beta_{min}^{routing}\%$ utilized then the packages will be delayed. They will be held back at the hub and

sent the next day after it has sufficient packages (freight) to fill the truck over the minimum utilization. We use $\beta_{min}^{routing}$ passively, to issue a warning when trucks are under-utilized.

In the following sections we describe the heuristic and analyze it.

2.3 Description of the Heuristic

We use a greedy least-cost heuristic to assign a terminal to a hub. The heuristic was designed to generate the assignments (spokes) in the hub-and-spoke system very quickly for analysis of what-if scenarios. In order to keep the completion time of the heuristic small, we analyzed the network from a broader perspective, which was in accordance with our objective of tactical planning of the network. By broader perspective, we mean that we do not consider the detailed flow of each and every individual shipment. Rather, we assume that shipments between terminals assigned to different hubs are routed through two hubs and for terminals assigned to the same hub the shipments are routed through just one hub. One of the advantages of this approach is that now each pair of origin-destination terminals has exactly one route for all its shipments, viz.,

origin terminal → hub of origin terminal → hub of destination terminal → destination terminal

or

origin terminal → common parent hub of origin and destination terminal → destination terminal

If direct loads were considered, the heuristic would have to specify the routing for each package individually and this would make the heuristic computationally exhausting. For this reason, we do not consider direct loads in the heuristic while assigning terminals to hubs. However, some obvious direct loads are reported after all the terminals are assigned.

Instead of treating the problem as a combinatorial problem and assigning all the terminals at once we sequentially assign terminals, one after the other. In the heuristic, the two cost components to be considered when assigning a terminal to a hub are:

1. Sorting cost
2. Transportation cost

2.3.1 Sorting Cost

The volume of packages that are sorted make sorting costs for the entire network a significant portion of the operating costs. Reducing sorting costs for each terminal can lead to significant savings at the network level. Sorting costs can be avoided either by bypassing a hub for distant terminals or by

shipment being routed through just one hub for nearby terminals (figure 9). However, since we do not search for the direct loads when assigning a terminal to a hub, we try to reduce sorting costs by assigning terminals that have greater movement of shipments amongst themselves to a single hub.

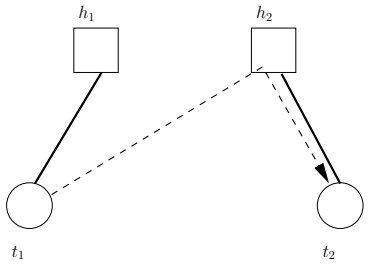
Let s_h be the sorting cost per package at hub h. If f_{ij} is the shipment size from i to j then the entire shipment need not be sorted. Only the less-than-trailerload amount of shipment needs to be sorted. If a portion of a shipment can fill up an entire trailer then that trailer need not be opened at the hub. Let f_{ij}^{ltl} denote the less than trailer-load portion of shipment from i to j.

The total cost of sorting freight from/to t_i if we assign t_i to h_1:

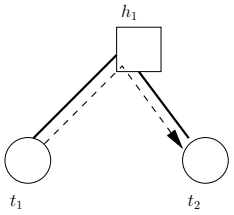
$$\sum_{j:hub(j)\neq h_1} (s_{h_1} + s_{hub(j)}) \cdot f_{t_i j}^{ltl} + \sum_{j:hub(j)=h_1} s_{h_1} \cdot f_{t_i j}^{ltl} \tag{1}$$

Sorting cost for the entire network is obtained by adding the sorting costs for all the shipments.

$$\sum_{t_i} \sum_{j:hub(j)\neq h_1} (s_{h_1} + s_{hub(j)}) \cdot (f_{t_i j}^{ltl} + f_{j t_i}^{ltl}) + \sum_{t_1} \sum_{j:hub(j)=h_1} s_{h_1} \cdot (f_{t_i j}^{ltl} + f_{j t_i}^{ltl}) \tag{2}$$



1. Loading direct



2. Movement of freight between two terminals assigned to the same hub

Assignment

Shipment route

Figure 9: Reducing Sorting Cost

2.3.2 Transportation Cost

For each terminal the transportation cost is comprised of

1. **Shuttle transportation costs**

   (a) *Outbound truck cost* which is the cost to send all the shipments from the terminal to its assigned hub in trucks.

   (b) *Inbound truck cost* which includes the cost to receive the shipments at the terminal from its assigned hub.

2. **Longhaul cost:** This is the cost to send (receive) the shipments from (to) the terminal's assigned hub to (from) the assigned hubs of the destination terminals.

The estimates for these costs depend on the type of costing model used.

2.3.3 Transportation Costing Models: Continuous and Marginal

Typically, the average utilization of a truck is higher on longhaul movements than on shuttle segments because of increased shipment concentrations on longhaul segments. Hence, it is a reasonable assumption to implement the *continuous costing model*, which charges for the fractional number of trucks on longhaul movements rather than the rounded up value of the trucks used. This means that our heuristic is not based on the idea of marginal cost, where a shipment can ride at no additional transportation cost in a truck with available capacity in longhaul segments.

Using the concept of marginal cost leads to a some sense of discontinuity in assignments because of available capacity on most longhaul trucks. Consider the following example where the first terminal, say t_1, is assigned to hub h_1. Since the terminal t_1 sends freight to most other terminals, the hub h_1 sends trucks with available capacity to almost all other hubs. Now with this information, let us consider the assignment for a second terminal, say t_2. If the continuous costing model is used, hub h_1 has no advantage over hub h_2 which has no terminal assigned to it yet. Hub h_1 has information about the terminals already assigned to it. If the marginal cost model is used, because of the available capacity on all longhaul segments, there will be no increase in longhaul cost if t_2 is assigned to h_1. Savings in longhaul costs will dominate any increase in shuttle transportation costs. In fact, hub h_1 will keep on attracting regional terminals as long as its sorting capacity is available. At some point, to assign a terminal t_j to hub h_1 will require sending a new truck from h_1 to another hub h_3. However, trucks on other longhaul movements still have available capacity and to increase in transportation cost on just one longhaul segment rather than all longhaul segments, terminal t_j is assigned to hub h_1. Hence, to dampen the effect of initial assignments we use the continuous cost model while assigning terminals to hubs. However, when we are reporting final costs we account

for transportation costs based on rounded up values of truck to be transported instead of fractional values of trucks.

The disadvantage of using the continuous costing model arises when the average utilization of trucks on longhaul segments is low. Consider two hubs which are closely located and both have relatively low truck utilizations on most of their longhaul lanes. The idea of the hub-and-spoke system is to consolidate freight and increase shipment concentrations along longhaul segments. However, the continuous cost model does not attract freight for consolidation by providing incentives to consolidate freight on longhaul segments. Rather, in this case it actually reduces the truck utilizations for the longhaul segments involving these two hubs by distributing the terminals among the two hubs. However, practically, since the scenario where two hubs are very closely located is rare in practice, the heuristic should rarely encounter this problem on industry data. For example, in the FedEx Ground network only two hubs, Sacramento and Rialto are located close to each other (under 100 miles).[1] Moreover, a simple way to deal with this disadvantage is to delete one of the two hubs in turn and analyze the effect it has on the hub-and-spoke network and its costs.

2.3.4 Trade-offs

We shall now consider the trade-offs in selecting a hub amongst a set of candidate hubs for assignment of a terminal.

Observation 1 *A terminal may be assigned to a distant hub which is en route to greater number of destinations than to a closer hub which is out of the way.*

Let us first consider the case without sorting costs. If a hub is along the path from origin to destination hub (or terminal) for most of the shipments then these shipments are not substantially diverted and hence do not contribute to a significant increase in transportation costs.

Consider a terminal t and two candidate hubs, h_1 and h_2, for assignment. Suppose h_1 is located nearer to t than h_2. If t is assigned to h_1 then the transportation costs on the shuttle segments are lower as compared to when t is assigned to h_2. The longhaul costs depending on the location of h_1 and h_2 with respect to the other hubs in the LTL network. It might be cheaper to assign t to a farther hub if the savings in longhaul transportation costs compensate the increases in shuttle transportation cost. Hence, neglecting sorting costs momentarily, the trade-off lies in assigning a terminal to a hub which maybe near and thus reducing the shuttle transportation cost and assigning

[1] It should be noted that a hub was built at Rialto which is close to Sacramento because Sacramento was over utilized and an in-house analysis found that it was cheaper to build a hub at Rialto than add capacity to the existing hub at Sacramento.

to a farther hub which possibly reduces the longhaul cost.

As an extreme example, consider h_1 is located on the east, nearer to t, and h_2 is on the west. If t sends all of its shipments westward then t will be assigned to hub h_2 than to h_1. Though h_1 is closer if t is assigned to h_1 all the shipments travel east before going westward.

The first trade-off lies in a terminal getting assigned to a nearer hub and having lower transportation costs versus it getting assigned to a farther hub and possibly lower longhaul costs.

Observation 2 *Any terminal t_i is drawn toward any hub that already sorts much of its frieght.*

If a terminal is assigned to a hub which already sort much of its freight then all those shipments will be routed through only one hub and will thus be sorted only once. Savings in sorting costs may offset increases in shuttle and longhaul transportation costs.

Example 2.2 *When sorting costs are considered, the terminal t may get assigned to hub h_2, if t sends many of shipments to terminals assigned to h_2. All of these shipments will then be sorted only once, at hub h_2. Instead, if terminal t is assigned to h_1 the shuttle transportation costs may be lower but all the shipments going to terminals assigned to hub h_2 will be sorted twice, first at h_1 and then at h_2.*

The second trade-off lies in a terminal getting assigned to a nearer hub and having lower shuttle transportation costs versus it getting assigned to a farther hub and not having to sort some of the shipments for the second time.

Based on the marginal cost model, a terminal may be willing to send its shipments to farther hubs, incurring greater shuttle costs but increasing the utilization of the longhaul trucks and thereby decreasing longhaul costs. Another consequence of the marginal cost model is that the network may have hubs that are not assigned to any terminals since assigning terminals to those hubs may entail sending trucks on longhaul segments that are not economically justified. This implies that all the shipments in the network are now routed through fewer hubs. Some shipments that were routed through two hubs are now routed through a single hub. This also reduces the overall sorting costs in the network.

2.3.5 The Greedy Assignment Heuristic

We shall first present the heuristic formally and then discuss it,

In the heuristic, we select a terminal, say t, that is to be assigned. Let us first consider the uncapacitated version where we neglect hub capacities. We greedily assign t to the nearest hub, say h_1. For the shuttle transportation cost we determine the total shipments in and out of the terminal.

Algorithm 1 Assign terminals

```
1: Sort terminals by decreasing LTL flow
2: repeat
3:    for each terminal t in 𝒯 do
4:       Assign t to a feasible hub with minimum cost
5:       if no feasible hub exists then
6:          Assign t to a closest hub
7:       end if
8:    end for
9: until Assignments remain unchanged
```

Cost Estimate 2 minCostHub(Terminal t): Determine the "least cost" assignment hub for terminal t

```
   if there are no candidate hubs for terminal t then
2:    assign t to its closest hub in the network
   else
4:    Sort the candidate hubs by increasing distance from t
      Assign t to the closest hub
6:    repeat
         Assign t to the closest candidate hub not yet considered, say h_next
8:       if estimated cost of assigning t to h_next is lower than that of assigned hub then
            replace the current assigned hub with h_next
10:      end if
      until the farthest candidate hub is considered
12: end if
   return the hub assigned to t
```

This gives us an estimate for the total number of trailers and trucks required for shuttle operations. To estimate the shuttle transportation cost we consider the maximum of the inbound and outbound trucks. This is reasonable because if terminal t is sending n_1 trucks to its hub h daily whereas the hub h sends back n_2 trucks to t then at the end of the day the difference ($|n_2 - n_1|$) has to be balanced for next days transportation needs.

Let f_{ij} be the shipment size from terminal i to terminal j, C be the truck capacity and c be the cost per truck per mile. Let $d_{i,j}$ be the distance between location i and j.

Total inbound trucks on shuttle segment are

$$\left\lceil \frac{\sum_{j \in \mathcal{T}} f_{t_1 j}}{C} \right\rceil$$

and the total outbound trucks on the shuttle segment are

$$\left\lceil \frac{\sum_{j \in \mathcal{T}} f_{j t_1}}{C} \right\rceil$$

To estimate the longhaul cost we consider a shipment that either originates or terminates at the terminal t. Since for the longhaul movement we use the continuous cost model, adding up the longhaul costs for each of the inbound and outbound shipments yields the total longhaul transportation cost.

The fractional truck for a shipment on longhaul segment is

$$\frac{f_{t_1 j}}{C}, \qquad \text{if } t_1 \text{ is the origin, and}$$

$$\frac{f_{j t_1}}{C}, \qquad \text{if } t_1 \text{ is the destination.}$$

We also estimate the shuttle transportation cost at other end of the route. Consider a shipment whose origin is t_1 and destination is t_2. Let t_2 be assigned to hub h_2. When considering the assignment of terminal t_1 we also need to account for the shuttle transportation costs between t_2 and its hub h_2. But for the same reason of accounting simplicity as on longhaul segments we use continuous costing model for the shuttle transportation costs between t_2 and h_2. This is an underestimate of the total shuttle cost on the segment. The underestimate in the approximation is less significant if on this shuttle segment the trucks are almost entirely utilized. The inaccuracy is justified by the simplicity in cost calculations and the reductions in solution time.

Fractional truck for shuttle transportation of a shipment at the other end is

$$\frac{f_{t_1 j}}{C}, \qquad \text{if } t_1 \text{ is the origin, and}$$

$$\frac{f_{j t_1}}{C}, \qquad \text{if } t_1 \text{ is the destination.}$$

Then, the total transportation costs for assigning terminal t_1 to hub h_1 are,

$$2 \cdot c \cdot d_{t_1,h_1} \cdot max \left\{ \left\lceil \frac{\sum_{j \in \mathcal{T}} f_{t_1 j}}{C} \right\rceil, \left\lceil \frac{\sum_{j \in \mathcal{T}} f_{j t_1}}{C} \right\rceil \right\} + \sum_j c \cdot \frac{(f_{t_1 j} + f_{j t_1})}{C} \cdot (d_{h_1,h_2} + d_{h_2,j}) \quad (3)$$

The total transportation costs for the network are

$$\sum_{t_1} 2 \cdot c \cdot d_{t_1,h_1} \cdot max \left\{ \left\lceil \frac{\sum_{j \in \mathcal{T}} f_{t_1 j}}{C} \right\rceil, \left\lceil \frac{\sum_{j \in \mathcal{T}} f_{j t_1}}{C} \right\rceil \right\} + \sum_{t_1} \sum_j c \cdot \frac{(f_{t_1 j} + f_{j t_1})}{C} \cdot (d_{h_1,h_2} + d_{h_2,j}) \quad (4)$$

Adding the sorting (equation 2) and total transportation costs (equation 4) gives us an estimate for assigning a terminal to a hub.

2.3.6 Sequence of Terminal Selection

Because the heuristic assigns the terminals sequentially, decisions made early in the heuristic will affect subsequent decisions and the quality of the solution.

Selection of the terminal is motivated by the approximation algorithm for bin packing problem [Vazirani, 2001]. Since transportation costs account for a significant proportion of the total operating costs, it seems to be a reasonable approach to select a terminal with the largest LTL shipment size to be sorted first. The idea is to first "pack" the bins, trucks in this case, with terminals having larger less-than-truckload shipments that are to be sorted.

A terminal is considered more *influential* than another terminal if it sends and receives more less-than-truckload shipment amount than the other terminal. The larger is the less-than-truckload amount that originates and terminates at the terminal, the more *influential* the terminal is. With this approach, the most influential terminals get assigned first and then the lesser influential terminals follow. This is a very intuitive approach because otherwise the least influential terminals dictate the shipment routes, and hence the assignments of the terminals with greater flow.

Example 2.3 *Consider terminals t_1 and t_2 as shown in figure 10. Also assume that terminal t_1 is more influential than terminal t_2. Moreover, t_2 only receives packages from t_1. Terminal t_2 is considered for assignment before terminal t_1. Since t_1 is not yet assigned, the cost estimates to assigned t_2 will be based on that simplifying assumption that the shipment will be sent directly from terminal t_1 to the parent hub of t_2. Hence, in order to save on shuttle transportation costs t_2 will be assigned to h_2, which is closer. So the shipment route from t_1 to t_2 is*

$$t_1 \rightarrow h_2 \rightarrow t_2$$

Now when t_1 is considered for assignment, it gets assigned to hub h_1 because it mostly communicates with terminals towards the east. So then the shipment route from t_1 to t_2 is

$$t_1 \rightarrow h_1 \rightarrow h_2 \rightarrow t_1$$

If t_2 had been assigned after t_1 it may be assigned to hub h_3 and the shipment route would then be

$$t_1 \rightarrow h_1 \rightarrow h_3 \rightarrow t_1$$

in which case the reduction in longhaul distance traveled could offset the increase in shuttle transportation cost.

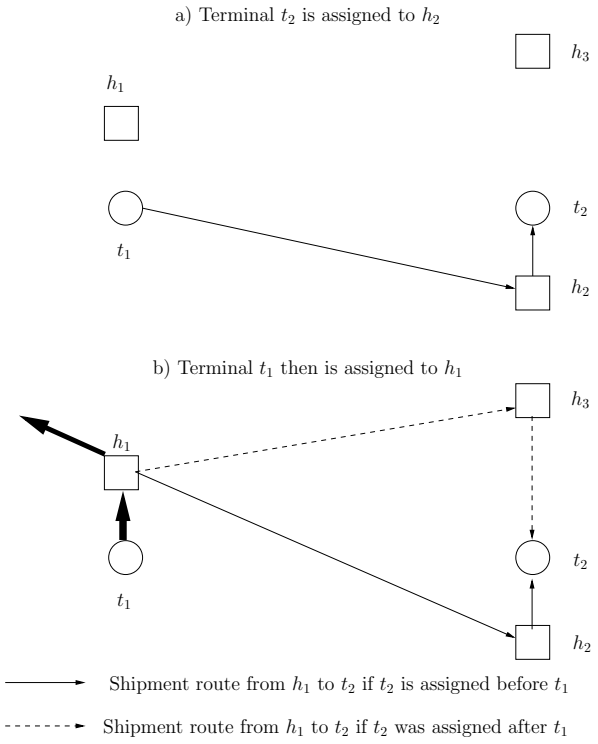


Figure 10: Sequence of terminal assignments can influence the shipment routes and hence the total costs in a network. This example shows how if a less influential terminal is assigned first it may yield costly shipment routes later on.

In the above mentioned approach a terminal is considered more influential based on the less-than-truckload shipment amount. In this approach the transportation cost is given more importance than sorting cost. Another approach is to consider a terminal influential if it has greater less-than-trailerload shipment amount originating and terminating at it. This is applicable when capacity constraints are enforced, in which case the sorting capacity is the bin to be packed. Typically, for a LTL carrier the number of truckload shipments is very small and hence we use the decreasing less-than-trailerload shipment size rule to select a terminal for assignment.

2.3.7 Candidate Hubs

For each terminal we choose hubs to which it might reasonably be assigned, estimate the cost of each assignment, and choose the least cost assignment.

Two business rules limit the choice of hubs to which a particular terminal may be asssigned. Most LTL carriers may have a time limit between freight leaving a terminal and arriving at its assigned hub for sorting. For example, FedEx Ground requires all of its shipments to be sorted within 10 hours, if possible, once it leaves the terminal.

For safety reasons, laws limit the number of hours a driver can drive continuously on the road. Sleeper teams cost twice as much per hour and so are to be used only when necessary. Driver sleeper teams cannot be justified economically on shuttle segments, because of reasonably shorter distances (less than 400 miles) traveled on shuttle segments. FedEx Ground requires all of the terminals to be within 8 hours of driving distance from the assigned hub, if possible.

Let us call any hub a *feasible hub* if it satisfies these two requirements for a terminal and the corresponding assignment be called a *feasible assignment*. Using this information within the heuristic, we limit the number of hubs a terminal is assigned to for cost estimation. For example, a hub in California need not be considered for assignment for a terminal in Florida.

By restricting the neighborhood over which a terminal looks for hubs we reduce the solution time of the heuristic. By preprocessing the data we can determine the maximum number of closest hubs that terminals should consider for assignment. If k_t is the maximum number of hubs that are feasible for terminal t then let $k = max_t\{k_t\}$. Then the heuristic considers only the k-closest hubs for assignment to the terminals.

It is interesting to note that the k^{th} closest hub may not be feasible for a terminal but $(k+1)^{st}$ closest hub can be feasible (see example below).

Example 2.4 *Consider terminal Lexington, KY and the hubs Atlanta, GA and Memphis, TN at a distance of 415 miles and 424 miles respectively from Lexington. Lexington and Atlanta are within the eastern standard timezone whereas Memphis falls within the central standard timezone. Assuming a industry standard speed of 50 miles/hour, a shipment leaving Lexington arrives at either of the hubs in a little over 8 hours. However, the truck going to Memphis crosses a timezone and gains an hour. Hence effectively, shipments leaving Lexington arrive at Memphis within 8 hours but arrive in Atlanta in a little over 8 hours. Memphis is thus a feasible hub for Lexington whereas Atlanta is not in spite of it being slightly closer to Lexington than Memphis.*

Certain terminals are located such that no hub is feasible. In such a case, the terminal selects

the cheapest infeasible hub. We illustrate an example from the FedEX data set.

Example 2.5 *Consider the terminal, El Paso, TX. The closest hub is Phoenix, AZ located 425 miles west of El Paso and the next closest hub is Fort Worth located 620 miles east. Since both of these hubs take over 8 hours from El Paso they are both infeasible. El Paso sends most (over 80%) of the shipments to the central and eastern USA. Though Phoenix is closer to El Paso, if El Paso is assigned to Phoenix most of the shipments have to travel 425 miles west and then to the respective destinations. However, by assigning El Paso to Fort Worth most of the shipments are routed through a hub which is en route to the respective destinations.*

2.3.8 Number of Closest Hubs to Consider

One of the parameters input to the heuristic is the number of closest hubs to consider for the assignment of each terminal. This parameter is dependent on the problem data. A terminal is assigned only to the least-cost feasible hub and if no hub is feasible then it is assigned to the least-cost hub. So the maximum number of feasible hubs for any particular terminal is an upper bound for the number of hubs to consider. So the next question that arises is: "Should all the possible feasible hubs be considered for assignment?". This is an important question especially when all the hubs are feasible, which is the case when the LTL carrier does not impose time limits on shipments being sorted at a hub once it leaves the terminal. Consider the following example.

Consider a terminal in Jacksonville, FL. Suppose Jacksonville does not receive any shipments but sends out shipments only to terminals located in north-west US. Also, if the shipments are such that it can fill a truck almost completely then Jacksonville can be assigned to the farthest hub, say Portland, OR if Portland is a feasible hub for Jacksonville. In that case, all the shipments will be sorted just once at Portland and the shuttle cost will replace the longhaul cost since the truck is almost full.

So essentially, assigning Jacksonville to Portland is equivalent to sending a direct truck from Jacksonville to Portland. Since the truck is almost full the average cost per package is almost the least.

Also, if Worcester, MA (assigned to Hartford, CT) sends shipments to Jacksonville then these shipments will then be routed as follows:

$$\text{Worcester} \rightarrow \text{Hartford} \rightarrow \text{Portland} \rightarrow \text{Jacksonville}$$

This is not the cheapest way to route freight from Worcestor to Jacksonville.

Now consider the case where Jacksonville cannot sufficiently fill a truck to Portland. In that case,

the average cost per package is exorbitant which can be reduced by moving the shipment from the shuttle segment to the longhaul segment where it can be consolidated with other north-west bound shipments, thus reducing the costs.

One way of understanding these dynamics is to consider the terminal-hub assignment problem as a weighted single facility location problem [Francis, McGinnis, and White, 1992]. In a weighted single facility Euclidean location problem there are weights associated with the distance of a "customer" from the facility. And the goal is to locate the facility so as to minimize the sum of weighted distances of the facility from the customers. In our case, the facility is the hub h to which t is being assigned and the customers can be categorized as:

1. One of the customers is the terminal t
2. All the hubs except h
3. All terminals assigned to h besides t

The weights can be considered as the number of trucks on the shuttle and longhaul segments. Once a terminal is assigned to it the number of inbound and outbound trucks are known. Since we are considering the continuous costing model note that the total number of trucks on the shuttle segments is at least equal to the total number of trucks on the longhaul segments. Hence, the new facility will be located nearer to the terminals assigned to it since their weights constitute at least 50% of the total weights. In the case of locating a hub in Euclidean distances setting and zero sorting costs the hub would be located *at* the terminal.

In the case where Jacksonville sends a full truck only to the north-west, the optimal hub location, ignoring the sorting costs, is anywhere on the straight line connecting Jacksonville and Portland (see figure 11(a)). However, to reduce the sorting costs the hub location coincides with Portland. But in the case where Jacksonville sends and receives shipments from other locations besides the north-west the optimal hub location shifts as close as possible to Jacksonville. Due to sorting cost considerations and non-Euclidean distances the terminal may not get assigned to the closest hub (see figure 11(b)).

2.3.9 Initial Approximations

Consider a terminal t_1 to be assigned to hub h_1. To estimate the longhaul costs we use the continuous cost model. For each shipment we estimate the longhaul cost and then add it for all the shipments associated with the terminal. Consider the shipment, say $f_{t_1t_2}$, from terminal t_1 to terminal t_2. Also, let at this point, terminal t_2 not be assigned to any hub. Now when estimating the transportation cost for $f_{t_1t_2}$, t_2 is not assigned to any hub, we just assume that the shipment will be sent directly

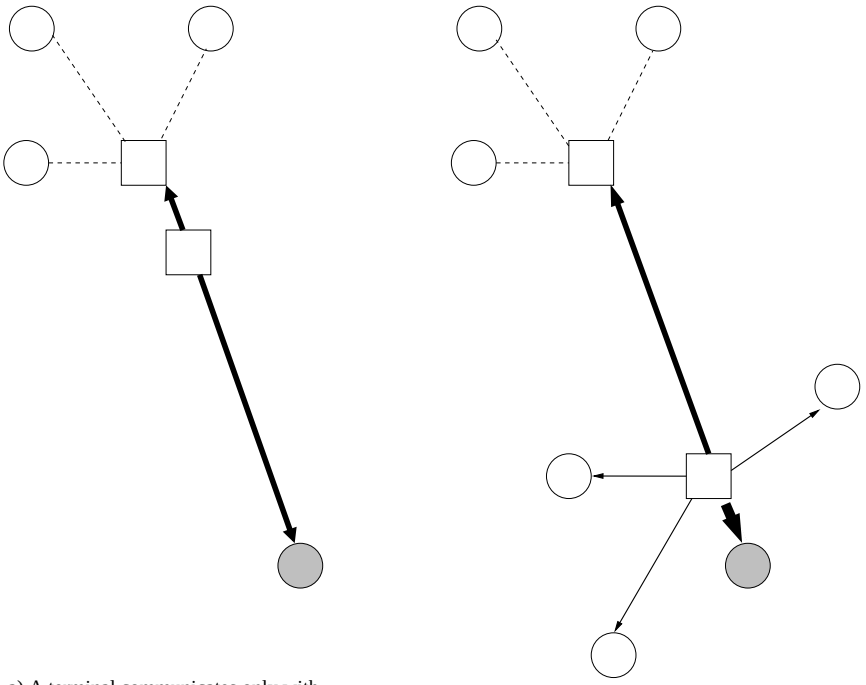


Figure 11: Freight patterns dictate terminal assignments. With the single facility location model, figure (a) shows how a terminal can be assigned to the farthest hub. Whereas, in figure (b) the terminal is assigned to the nearest hub.

from hub h_1 to terminal t_2. By this approximation the transportation cost is underestimated. This effect of this approximation becomes trivial as more and more terminals get assigned to hubs. As a result, the inaccuracy of the assignment is greater during the initial assignments and decreases as the heuristic progresses.

The inaccuracy is not as severe as it seems because the terminals which get assigned initially are the terminals that are more *influential*. So due to lack of adequate information for all the shipment routes, these influential terminals get assigned to the best possible hubs. Their assignments are not influenced by any terminal with a small shipment size. However, the assignment of terminals with small shipment sizes is determined by the influential terminals (see example 2.3).

In order to account for the inaccuracy of the heuristic one approach is to re-assign the terminals once all of the terminals are assigned. Let us call the step in the heuristic of assigning all the terminals once as *optimization pass*. During the second optimization pass, we re-assign the terminals with the additional assignment information available from the first pass. The assignment information available from the first pass may change the assignments of certain terminals during the second pass. In fact, the heuristic stops only when the assignments stabilize, that is, the assignments in

two consequent passes remain unchanged for all the terminals.

2.3.10 Number of Passes

When assigning a terminal to a hub, we estimate costs by adding up all the fractional trucks required for each shipment on the long haul. Consider the terminal t_o which is the origin of a shipment destined to t_d. Let h_o be the hub under consideration as the parent hub for t_o. If t_d is assigned to h_d then the estimate is based on the shipment route,

$$t_o \rightarrow h_o \rightarrow h_d \rightarrow t_d$$

However, if t_d is not yet assigned to any hub the estimate is based on the shipment being sent directly from h_o to t_d. The shipment route is,

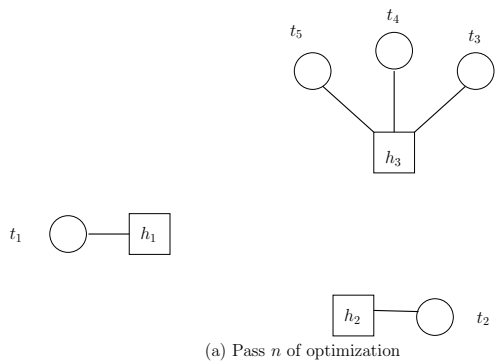
$$t_o \rightarrow h_o \rightarrow t_d$$

This, of course, is likely *not* the shipment route. This means that, when t_o is assigned it is assigned based on incomplete information about the routes of all its shipments. If t_o is the first terminal to be assigned, the shipment from t_o is sent directly from the hub, h_o to the destination terminal, t_d. However, at the end of the first pass of optimization, more information is available for the routes of all the shipments. Hence, to verify that h_o is still the least-cost hub for t_o based on the additional information of shipment routes, the heuristic repeats.

The following example gives a scenario where a terminal changes assignment in the subsequent pass.

Example 2.6 *Consider terminals $t_1, \ldots, t_5$ to be assigned, in that order. Let h_1 be the only hub that is feasible for t_1. Let hubs h_2 and h_3 be feasible for t_2 and let h_3 be the only feasible hub for terminals $t_3, \ldots, t_5$. When t_1 is to be assigned it gets assigned to h_1. Terminal t_2 can get assigned to either h_2 or h_3. With only t_1 assigned, the estimate of t_2 is based on the direct loading the shipments from h_2 to $t_3, \ldots, t_5$ respectively. Since the continuous cost model is used for longhaul and direct trucks t_2 gets assigned to h_2 instead of h_3. The only sorting costs that are to be considered are those for shipments to t_1 which have to be sorted twice irrespective of whether t_2 gets assigned to h_2 or h_3. Finally, terminals, $t_3, \ldots, t_5$ get assigned to h_3 (see figure 12(a)).*

During the second pass of optimization in the heuristic, assignment of t_1 remains unchanged since h_1 is the only feasible hub it can get assigned to. Now when terminal t_2 is being assigned, we have information about assignments of terminals $t_3, \ldots, t_5$ which was unavailable during the first pass of the heuristic. During the second pass, if t_2 gets assigned to h_2 all the shipments to $t_3, \ldots, t_5$ have to



(a) Pass n of optimization

(a) Pass $(n+1)$ of optimization

Figure 12: Assignments can change during re-optimization as more accurate shipment route information is available after all the terminals are assigned.

be sorted twice. However, if it gets assigned to h_3 the shipments to $t_3, \ldots, t_5$ have to be sorted just once. If the savings in sorting costs compensate for an increase, if any, in estimated transportation costs then t_2 will change its assignment from hub h_2 to hub h_3 (see figure 12(b)).

Table 1: Cost comparison after re-optimization

Cost	1^{st} pass	2^{nd} pass	3^{rd} pass
Sorting Cost	965,486	964,667	961,536
Transportation cost:			
Longhaul	1,014,862	1,015,242	1,015,670
Shuttle	392,131	391,821	394,283
Total	2,372,479	2,371,730	2,371,488

Table 1 compares the overall costs during the subsequent passes as we repeat the heuristic for one of the data sets provided to us by FedEx Ground[2]. The assignments stabilized after 3 passes of reoptimization. In figure 13 we can see the dynamics of the individual cost components. The sorting cost is the largest in the first of the three passes. However, during the second pass as information

[2]The data set was a combination of FedEx Ground and FedEx Home Delivery networks with 710 terminals, 24 hubs and about 120,000 shipments.

for all the shipment routes is available some terminals change assignments. During the second pass, reduction in sorting and shuttle transportation costs comes at the expense of increased longhaul costs. During the third pass, though shuttle as well as longhaul costs increase, these increases are compensated for by the savings in sorting costs and longhaul costs.

So if a terminal can change assignments during re-optimization, can the heuristic ever run indefinitely? Can the heuristic, ever generate assignments that flip-flop during consequent passes?

The heuristic only considers k hubs that are closest to a terminal for assignment. For a network with $|T|$ terminals, there are $k^{|T|}$ possible assignments. A terminal will change assignments only if it ensures positive savings in the cost estimate. This implies that at most $k^{|T|}$ assignments will be considered.

Though theoretically the heuristic has exponential running time, practically it works very well. For the FedEx Ground network with about 700 terminals and 24 hubs, considering 6 closest hubs for assignment to a terminal we needed only 3 passes of optimization.

2.3.11 Capacity Constraints

So far we have not considered restrictions on sorting capacity at a hub, so a terminal can be assigned to a hub even if its sorting equipment is over-utilized. When limits on sorting capacity are enforced within the heuristic we consider a hub for assignment only if has sufficient capacity to sort shipments (both inbound and outbound) from the terminal.

However, one of the problems that can arise with this approach is that at some point within the heuristic for a certain terminal, all of the k-closest hubs could have insufficient capacity and it would not be assigned to any hub. In such a case, depending on the LTL carrier's policy the terminal may be assigned either to a hub which has minimum utilization to reduce congestion or to the least-cost hub which reduces the operating costs.

Consider the following scenario when capacity constraints are enforced. Hub h_1 is feasible with regard to time constraint in which shipments from terminal t have to be sorted. However, it does not have sufficient capacity. Consider another hub, h_2 which has sufficient capacity but cannot make the sort within the required time. Which hub should the terminal t be assigned to? Both the hubs are feasible in one aspect but are infeasible in another aspect. This again depends on the LTL carrier's policy. FedEx Ground preferred to enforce hub capacity restrictions more strictly than constraints on time to sort shipments at the hub.

By assigning terminals in the order of decreasing LTL shipment size we ensure that the terminals with greater flow have sufficient capacity initially when they are being assigned. It is easier to find

hubs with available capacity for terminals with smaller flow than for terminals with larger shipment sizes.

Typically, for a LTL carrier, most of the shipments are less-than-trailerload and therefore, sorting by less-than-truckload shipment size is almost equivalent to sorting my less-than-trailerload. However, there may be cases where these two rules are not equivalent.

Example 2.7 *Consider terminal i which sends out 1001 packages (just over a trailerload) to all other terminals in the network. Terminal i will be assigned first under the decreasing LTL shipment size rule but will be assigned last under the decreasing less-than-trailerload shipment size rule. Consider terminal j has the maximum less-than-trailerload shipments associated with it. Also assume that if limits on hub capacities are not enforced both will be assigned to a single hub, say h. However, if hub capacities are enforced then either i or j can be assigned to h.*

If capacity is severely constrained, assigning i at the end is preferred. However, in that case, over-utilization of hub h is prevented at the expense of possible increase in transportation costs because most trailers originating at i might now have to travel longer distances. If i is assigned initially, to avoid congestion at hub h, terminal j will be assigned to another hub, possibly leading to increased sorting cost.

So there exists a trade-off between transportation costs and sortation costs when shipments are not less-than-trailerload shipments. However, for a LTL carrier which delivers mostly less-than-trailerload shipments assigning terminals by LTL shipment size is reasonable.

2.4 Comparing the Cost Models

We also implemented the marginal cost model to get more insights into the strengths and weaknesses of the continuous cost model. Figure 14(b) shows the hub-and-spoke network generated by the marginal cost model. In table 2 we compare the individual and total costs for the continuous and marginal costing models. As explained in sections 2.3.3 and 2.3.4 we make the following observations.

Observation 3 *Average utilization of longhaul truck is higher when marginal costing model is considered.*

The marginal cost model tries to send shipments with the least marginal cost. Trailers having additional capacity will attract shipments until it is full.

Observation 4 *Under the marginal costing model, fewer hubs are utilized and their utilization is increased.*

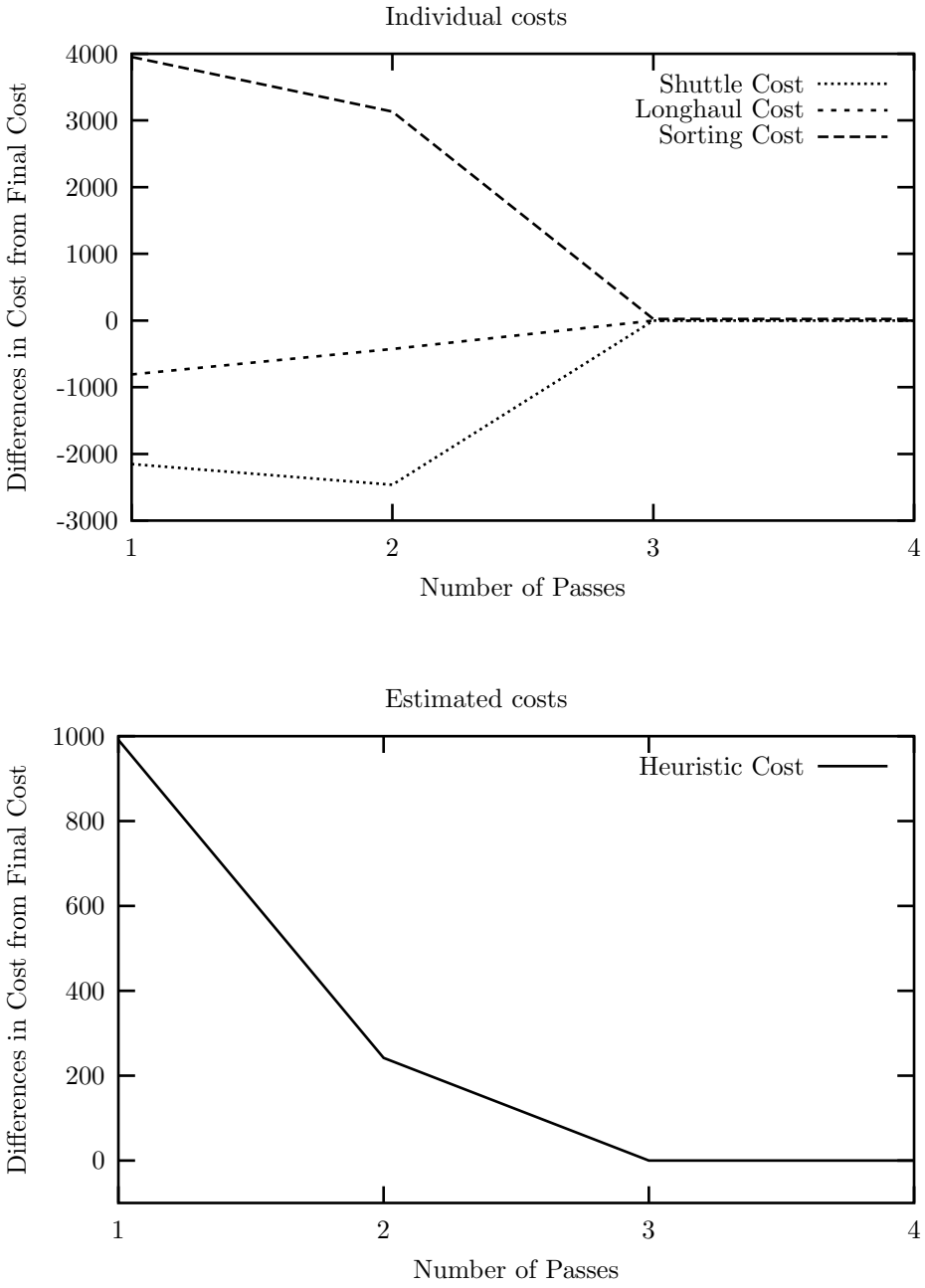
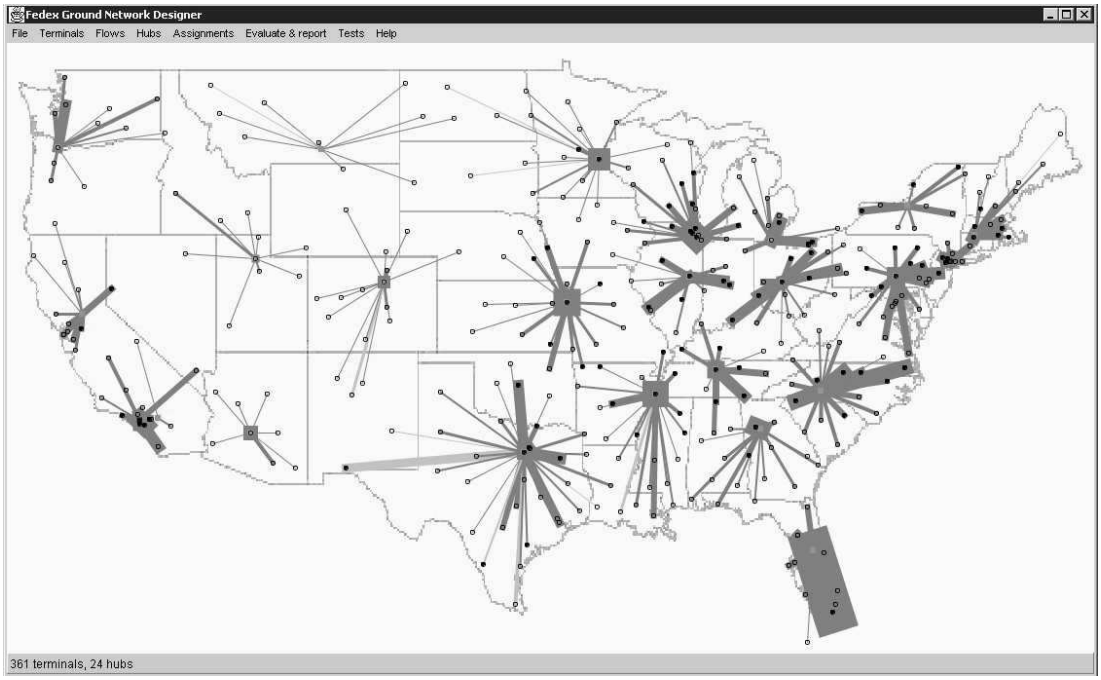
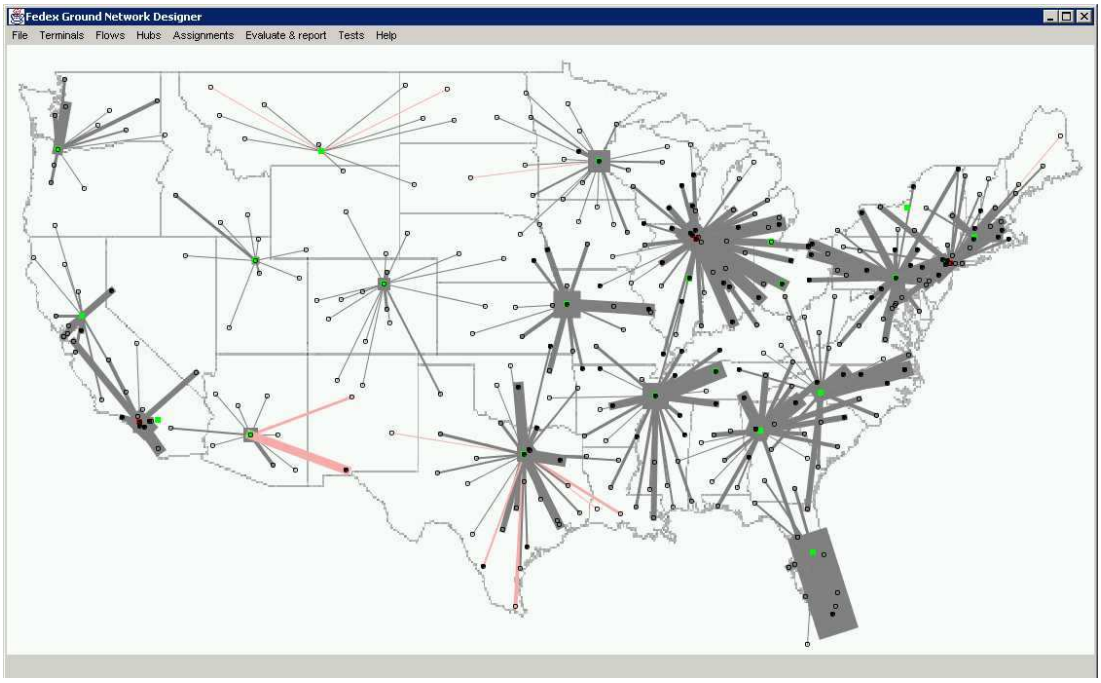


Figure 13: Changes in cost versus number of reoptimizations



(a) Continuous cost model



(b) Marginal cost model

Figure 14: Hub-and-Spoke network generated by the heuristic for FedEx Ground in mainland USA for single assignment policy – a terminal can be assigned to only one hub.

Table 2: Cost comparison: continuous cost model versus marginal cost model

Cost	Marginal cost model	Continuous cost model
Sorting Cost	704,096	718,271
Transportation cost:		
Longhaul	890,868	973,418
Shuttle	393,228	297,980
Total	1,993,791	1,995,267

Since, on average, a terminal communicates with 70% of other terminals in the network, its parent hub will send out longhaul trucks to most of the hubs in the network. So assigning a terminal to a hub which has no terminals assigned yet implies a very high average cost per packages originating or terminating at the terminal. Essentially, every hub repels terminals as long as it can be avoided, beyond which it attracts as many terminals as it can. This results not only in higher utilizations of trucks on longhaul segments but also higher utilization of sorting equipment at the hubs. When a hub attracts more terminals, more packages will avoid sorting the second time.

Observation 5 *Under the marginal costing model, shipments travel greater shuttle distances to avoid sorting costs and longhaul transportation costs.*

This is a consequence of the previous two observations. Though shipments from (to) a terminal travel larger shuttle distances, the savings in longhaul transportation costs and sorting costs offset the increases.

2.5 The Greedy Heuristic versus existing FedEx Ground Network

For the FedEx Ground data set the heuristic generated the hub-and-spoke network shown in figure 15(a). It is strikingly similar to the hub-and-spoke network currently existing at FedEx Ground, shown in figure 15(b). Less than 12% of the terminals differed in assignments in the two networks. The FedEx Ground network has evolved over a period of time with extremely sophisticated decision support tools and years of expertise. In some cases, our heuristic, which is entirely cost based may prefer an unintuitive assignment that is slightly cheaper over an intuitive one. We try to explain few of the assignment differences in the two networks.

Table 3: Comparison of continuous cost model versus marginal cost model

	Marginal cost model	Continuous cost model
Average utilization of		
sorting capacity at hubs	71.3%	55.2%
trucks on longhaul	71.8%	65.3%
Average distance (miles)		
Shuttle segment	209	170
Longhaul	1421	1333
Number of packages sorted only once	222,426	166,458
Number of hubs not utilized	6	0
Number number of terminals assigned to a hub	20	15
Solution time of the heuristic (minutes)	>>>5	3-5

2.5.1 Differences in Terminal Assignments

2.5.1.1 Assignment policies

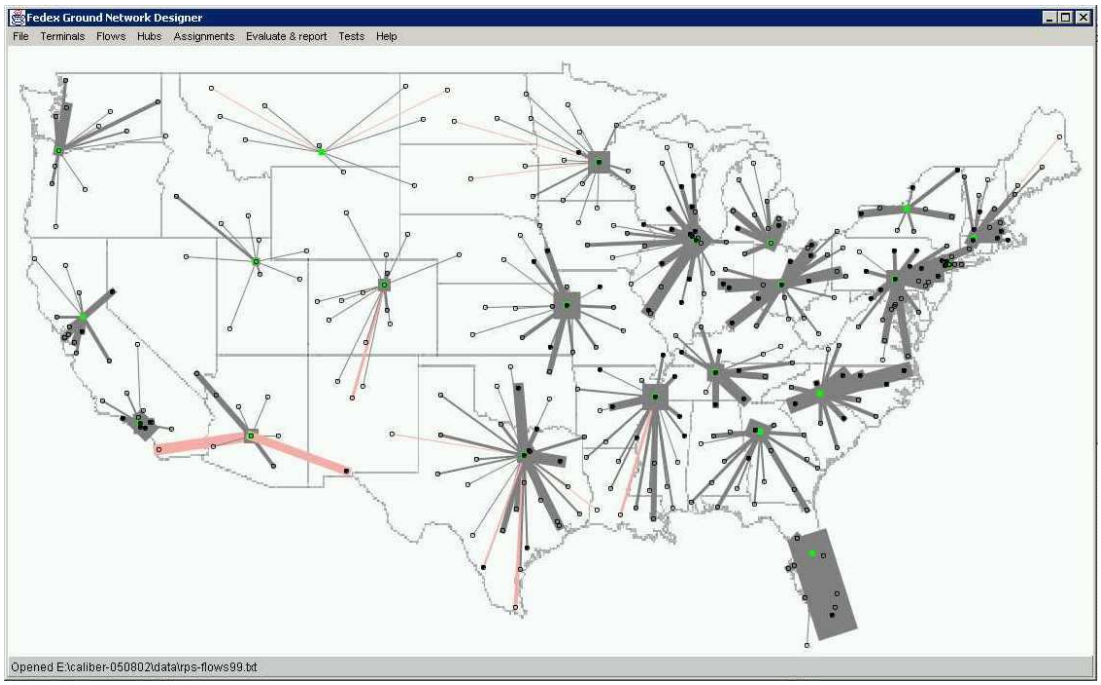
In cases, where no hub is feasible a terminal can be assigned either to minimize the infeasibility or to minimize the costs.

Example 2.8 *El Paso, TX: Shipments from El Paso to two of its closest hubs, Phoenix, AZ and Fort Worth, TX cannot be sorted within 8 hours, so both of these hubs are infeasible. Our heuristic makes the cheaper assignment and chooses Fort Worth, TX as the parent hub of El Paso. On the contrary, in the FedEx Ground network El Paso is assigned to Phoenix, AZ. FedEx Ground has a policy to assign a terminal to the closest hub in case no feasible hub exists. However, such policies can easily be implemented within the heuristic.*

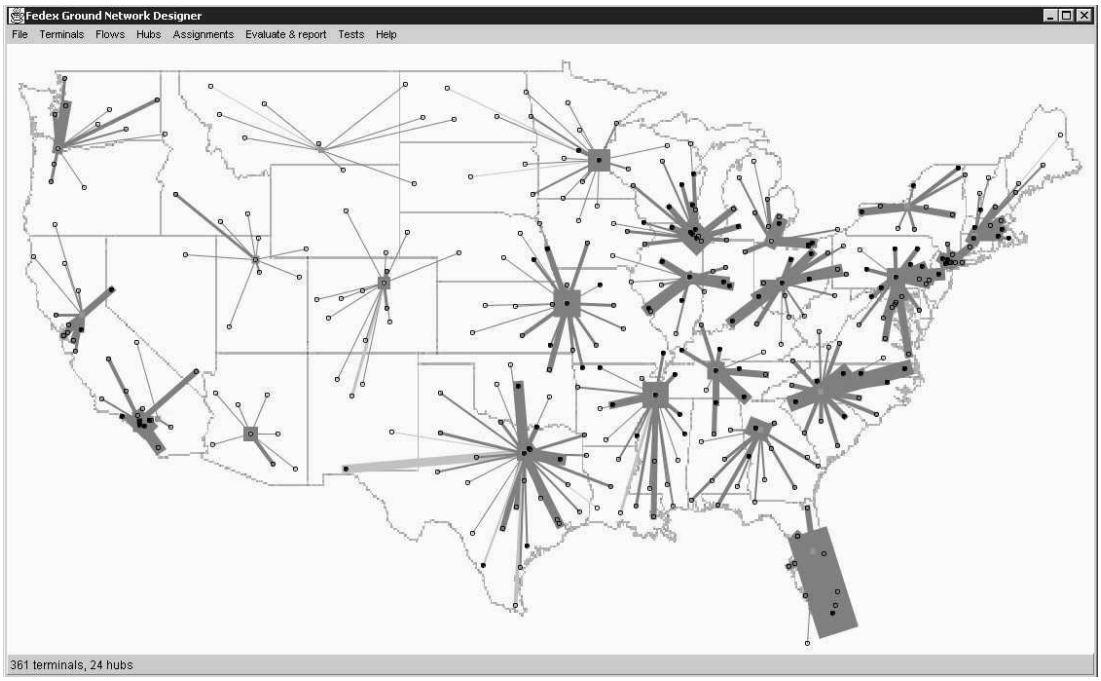
2.5.1.2 Insufficient data

Since not all distances between terminals and hubs are available in the heuristic, approximates the distance by the great-circle-mileage distance and a road factor (= 1.2) to account for actual over-the-road distances.

Example 2.9 *Grand Rapids, MI: Our heuristic assigned a Grand Rapids and a few other terminals on the east of Lake Michigan to Chicago. These are assigned to Toledo in the actual FedEx Ground network. Because of Lake Michigan the actual road distances between some terminals in that region and Chicago are longer than the estimated over-the-road distances.*



(a) Hub-and-spoke network in practice at FedEx Ground



(b) Hub-and-spoke network generated by NetworkDesigner

Figure 15: NetworkDesigner generates a hub-and-spoke network that very closely resembles the one FedEx Ground has in practice

2.5.1.3 Soft assignment policies

Typically, a terminal is assigned to a feasible hub, if one is available. However, sometimes such constraints maybe relaxed locally for a terminal either for operational convenience or operational efficiency. It is not possible to implement such soft constraints in the heuristic in a general scheme.

Example 2.10 *San Diego, CA: Our heuristic assigned San Diego, CA to a feasible hub, Sacramento, CA where as in the FedEx Ground network it was assigned to Phoenix, AZ though it was an infeasible hub.*

2.5.1.4 Negligible cost differences

Our heuristic may yield assignments different from the FedEx Ground network for certain terminals to obtain negligible estimated cost savings.

Example 2.11 *Alexandria, AL: Alexandria, AL is assigned to Memphis, TN in the FedEx Ground network. Our heuristic estimates that assigning Alexandria to Fort Worth, TX is $10 cheaper than assigning it to Memphis.*

2.5.1.5 Factors beyond economic consideration

The heuristic is strictly based on an economic model. However, in a real life scenario miscellaneous factors governed by daily operations and beyond the scope of economics play a significant role in designing the network. Our heuristic forms the backbone of a vital decision support. However, it does not aim to substitute for an analyst.

Figure 16 shows the terminals that were assigned differently by NetworkDesigner compared to the actual FedEx Ground hub-and-spoke network.

2.5.2 Comparing Operating Costs

Table 4: Operating cost estimate for the greedy heuristic is slightly lower than that of the FedEx Ground

Cost	Greedy Heuristic	FedEx Ground
Sorting Cost	704,445	709,066
Transportation cost:		
Longhaul	943,314	947,057
Shuttle	316,882	322,836
Total	1,969,641	1,978,959

Table 4 compares the individual costs along with the total cost for the hub-and-spoke networks generated by the greedy heuristic and the existing FedEx Ground network. The greedy assignment

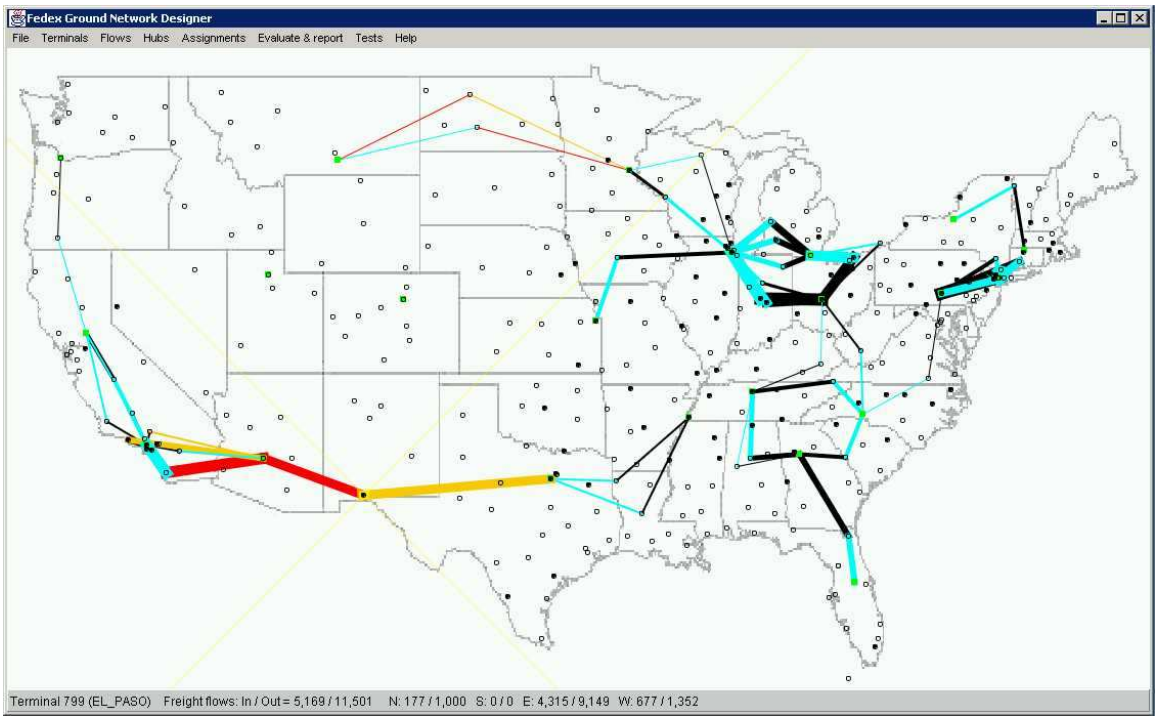


Figure 16: This figure shows the assignments that are different in NetworkDesigner and the FedEx Ground solution.

heuristic generates yields an daily operating cost which is about 0.5% cheaper than the existing FedEx Ground. In absolute terms, this results in about $2.8 million in annual savings.

2.6 Dual Assignments

So far we have focused on assigning a terminal to just one hub. The advantages include:

1. The spokes feed into the hub increasing shipment concentrations on the longhaul segments
2. By assigning a terminal to just a single hub we also maximize the shipment concentrations on shuttle segments.

However, some times it is possible that by taking advantage of the patterns of the shipment flows in and out of a terminal, a load planner can achieve lower transportation costs by assigning the terminal to more than one hub.

Under the policy of dual assignments a terminal can be assigned to one or two hubs. Consider a terminal t assigned to hub h_1 under the single assignment policy. This terminal sends trucks to h_1 and receives trucks from h_1. Now, if this terminal t is assigned to hubs h_1 and h_2 then terminal t sends trucks to and receives trucks from both the hubs. This may seem reasonable if the terminal

sends (receives) more than one truck to (from) the hub. In cases where a terminal sends (receives) only one truck to (from) the hub it may be difficult to economically justify assigning the terminal to two hubs wherein it sends one truck with reduced utilization to each of the hubs. However, a more reasonable approach is to assign a hub for outbound freight and another hub for inbound freight if justified. By doing so, we reduce costs and improve service without reducing shuttle truck utilizations.

For example, consider a terminal which receives most of its freight from the east but sends out most of the freight to the west. If the terminal is assigned to only one hub, say the one on its east, then all the bound freight has to travel east before going to the east. If, instead, the terminal is assigned to the hub on the east for all the inbound freight and to the hub on the west for all the outbound freight then most of the shipments are routed without significant deviation from the route to the final destination. In this case, there is no reduction in truck utilizations on the shuttle routes. This is because, the inbound and outbound shipments will be loaded onto different trucks in either case.

Besides maintaining the truck utilizations there is another advantage of dual assignments based on having inbound and outbound hubs. By specifying an inbound and an outbound hub (which may be the same) for a terminal, all the shipment routes are known,

$$t_o \rightarrow h_o^{out} \rightarrow h_d^{in} \rightarrow t_d$$

where,

t_o	origin terminal
t_d	destination terminal
h_o^{out}	outbound hub for the origin terminal
h_d^{in}	inbound hub for destination terminal

When a terminal is assigned to two hubs both of which can be used of inbound and outbound trucks, then the assignment does not portray any information about the routes of the shipments. Let terminal t_o, assigned to hubs h_1 and h_2, send a shipment to terminal t_2 assigned to h_3 and h_4. Then there are four possible ways in which this shipment can be routed.

With hub h_1 as the first hub we have the following two routes,

$$t_1 \rightarrow h_1 \rightarrow h_3 \rightarrow t_2 \quad \text{and} \quad t_1 \rightarrow h_1 \rightarrow h_4 \rightarrow t_2$$

Similarly with h_2 as the first hub we have the remaining two routes,

$$t_1 \rightarrow h_2 \rightarrow h_3 \rightarrow t_2 \quad \text{and} \quad t_1 \rightarrow h_2 \rightarrow h_4 \rightarrow t_2$$

A pair of hubs will be selected in order to reduce costs for each shipment, which specifies the route associated with each shipment. Without this information a shipment may be routed in a more expensive way increasing the overall costs. However, having a path based heuristic would increase the computational time exorbitantly since now each shipment route has to be determined and "remembered". Hence, we implemented a dual assignment policy by restricting a terminal to route all of its inbound (outbound) shipments through exactly one hub.

The heuristic is very much similar the that for single assignments except that now we determine inbound and outbound hubs separately.

The cost estimate for assigning terminal t_1 to hub h_1 (inbound) is,

$$c_{in}(t_1, h_1) = \overbrace{c \cdot d_{t_1,h_1} \cdot \left\lceil \frac{\sum_{j \in \mathcal{T}} f_{jt_1}}{C} \right\rceil}^{\text{shuttle cost}} + \overbrace{\sum_{j \in \mathcal{T}} c \cdot \frac{f_{jt_1}}{C} \cdot (d_{h_1,h_{out,j}} + d_{h_{out,j},j})}^{\text{longhaul cost}} + \underbrace{\sum_{j:h_{out,j} \neq h_1} (s_{h_1} + s_{hub(j)}) \cdot f^{ltl}_{jt_1} + \sum_{j:h_{out,j}=h_1} s_{h_1} \cdot f^{ltl}_{jt_1}}_{\text{sorting cost}}$$

and the estimated cost for assigning terminal t_1 to hub h_1 (outbound) is,

$$c_{out}(t_1, h_1) = \overbrace{c \cdot d_{t_1,h_1} \cdot \left\lceil \frac{\sum_{j \in \mathcal{T}} f_{t_1 j}}{C} \right\rceil}^{\text{shuttle cost}} + \overbrace{\sum_{j \in \mathcal{T}} c \cdot \frac{f_{t_1 j}}{C} \cdot (d_{h_1,h_{in,j}} + d_{h_{in,j},j})}^{\text{longhaul cost}} + \underbrace{\sum_{j:h_{in,j} \neq h_1} (s_{h_1} + s_{hub(j)}) \cdot f^{ltl}_{t_1 j} + \sum_{j:h_{in,j}=h_1} s_{h_1} \cdot f^{ltl}_{t_1 j}}_{\text{sorting cost}}$$

where, $h_{in,j}$ and $h_{out,j}$ are the inbound and outbound hubs to which terminal j is assigned.

Algorithm 3 Dual Assign terminals

```
1: Sort terminals by decreasing LTL flow
2: repeat
3:    for all t in T do
4:       minCostOutboundHub(t)
5:       minCostInboundHub(t)
6:    end for
7: until Assignments remain unchanged
```

The trade-offs existing are exactly similar to those in a single assignment model. Except now these trade-offs apply to the inbound and outbound shipments independently.

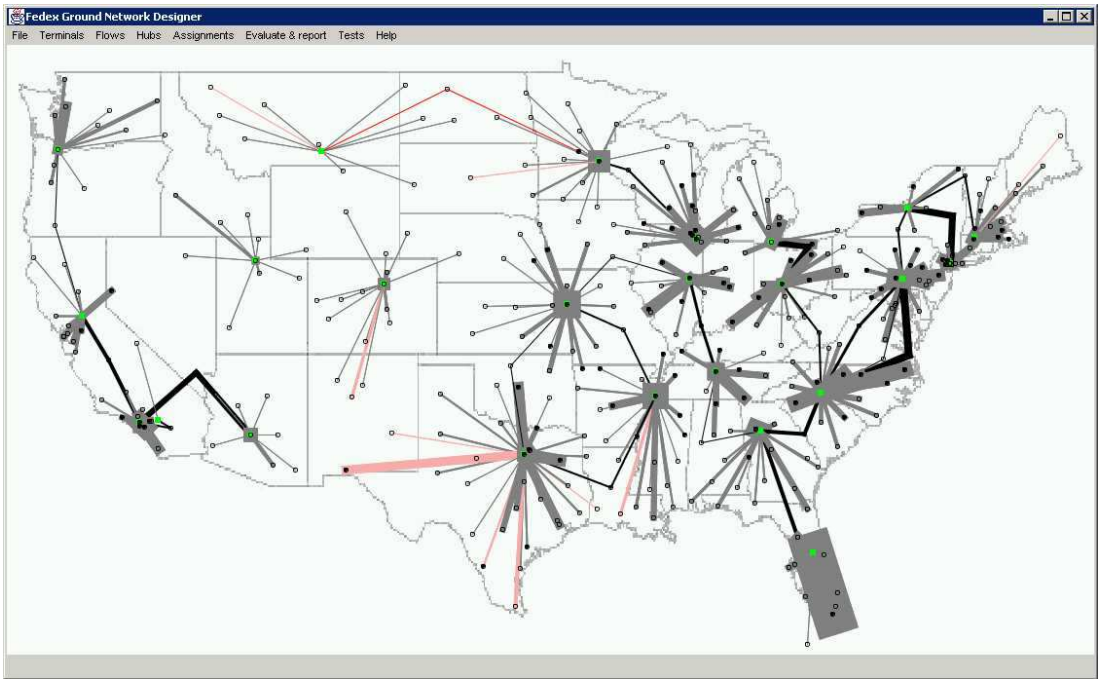
Figure 17(a) shows the hub-and-spoke network for the dual assignments. Figure 17(b) shows only those terminals that have been assigned to two different hubs for inbound and outbound shipments.

Algorithm 4 minCostOutboundHub(Terminal t): Determine the "least cost" assignment hub for terminal t for outbound shipments

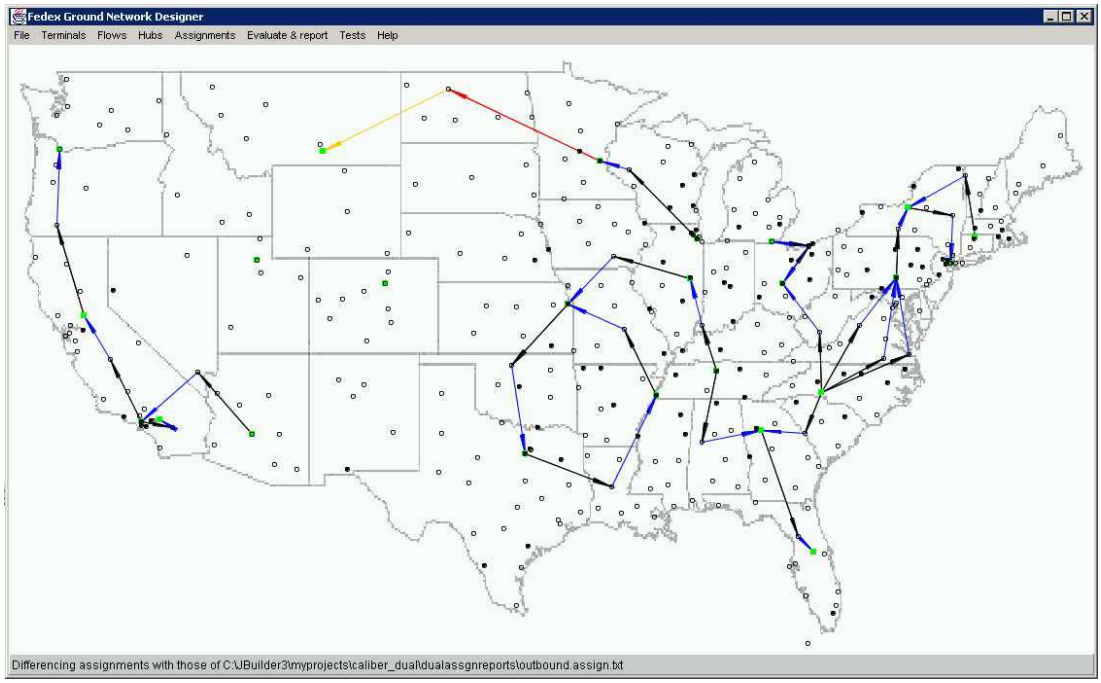
```
if H(t) = ∅ then
  hub(t) = closest(t)
  return hub(t)
else
  Sort hubs in H(t) by increasing distance from t
  hub(t) ← null
  for all h in H(t) do
    if hub(t) = null then
      hub(t) ← h
    else
      if c_out(t, h) < c_out(t, hub(t)) then
        hub(t) ← h
      end if
    end if
  end for
  return hub(t)
end if
```

Algorithm 5 minCostInboundHub(Terminal t): Determine the "least cost" assignment hub for terminal t for inbound shipments

```
if H(t) = ∅ then
  hub(t) = closest(t)
  return hub(t)
else
  Sort hubs in H(t) by increasing distance from t
  hub(t) ← null
  for all h in H(t) do
    if hub(t) = null then
      hub(t) ← h
    else
      if c_in(t, h) < c_in(t, hub(t)) then
        hub(t) ← h
      end if
    end if
  end for
  return hub(t)
end if
```



(a) The hub-and-spoke network



(b) Terminals assigned to two hubs - one for inbound and one for outbound shipments.

Figure 17: Hub-and-Spoke network generated by the heuristic for FedEx Ground in mainland USA for dual assignment policy – a terminal can be assigned to one or two hubs.

Based on the case study for FedEx Ground we made a few interesting observations that are listed below.

Observation 6 *The freight flows within the network are balanced.*

From figure 17(b) we observe that very few terminals are assigned to different hubs for inbound and outbound shipments. In fact, for FedEx Ground only 24 of the 361 terminals (less than 7%) got assigned to two distinct hubs. This means that very few terminals have significantly different shipment patterns for inbound and outbound shipments. This implies that for most of the terminals the shipments are such that the proportionality of flows in a particular direction is the same for inbound and outbound shipments.

Observation 7 *The volume of shipments on shuttle segments influences dual assignments.*

We observed that the terminals that were assigned to two hubs had significant differences in inbound and outbound shipment volume. As explained in single assignments, most terminals get assigned to the closest hub unless the utilization on shuttle segments was sufficient for it to travel farther to the next closest hub thereby reducing longhaul and/or sorting costs.

Example 2.12 *Ocala, FL sends out 2 trailers-load (1 truckload) of shipments and receives 5 trailer-load (3 truckloads of shipments). Atlanta is almost thrice as far from Ocala as Orlando. Instead of receiving the shipments from Orlando, FL if it receives all the shipments from Atlanta, GA the shuttle costs have almost tripled but have been compensated for by reductions in longhaul costs simply because Atlanta is on the way for most shipment coming into Ocala. The sorting costs are approximately constant for either hub.*

Table 5: Hub analysis for shipments from Ocala, FL

Hub	Distance	Shuttle cost	longhaul cost	sorting cost	total cost
Orlando	124	372	2615	2236	5223
Atlanta	326	977	1936	2243	5157

(a) Inbound shipments

Hub	Distance	Shuttle cost	longhaul cost	sorting cost	total cost
Orlando	124	124	825	721	1670
Atlanta	326	326	669	753	1747

(b) Outbound shipments

However, for the low volume outbound shipments the reduction in longhaul costs does not justify the increase in shuttle cost if assigned to Atlanta. Hence, Ocala receives all the shipments from Atlanta but sends out the shipments to Orlando.

Observation 8 *For most terminals the inbound and outbound hubs are not located angularly close to each other.*

If they are located angularly close to each other then only reason would be that the inbound and outbound shuttle have sufficient volume in shipments to justify going almost in the same direction but not to the same hub. The "small" savings per shipment are magnified by the volume of shipments.

2.7 Network Scaling and Robustness

2.7.1 Scaling

One of the tasks of the tactical planners is to have an insight of what the network will resemble over a short term period (3–5 years). Any approach would require a forecast of the freight flows. One of the basic concerns of tactical planners is the accuracy of the forecasts and any sensitivity analysis of freight flows can prove to be a useful tool.

One way to perform sensitivity analysis is to randomly perturb flow(s) and determine its effect on the network configuration. The inherent idea in this approach is simulation and it is difficult to conclude any specifics about the changes in the network due to freight flow perturbations. To get a better feel for perturbing the freight flows we introduce the concept of *scaled network*. A network scaled by a factor of k (> 0) means that all the freight flows are scaled by a factor of k. This is a reasonable practical scenario since the freight flows are dependent on the national economy and depending on the state of the economy the freight flows throughout the network can be scaled by an almost equal scaling factor. This idea is not applicable when terminals (customers) are added or deleted from the network. For example, adding a major customer which mostly ships out freight will possibly distort the existing freight pattern.

2.7.2 Robustness

To test the robustness of our heuristic, we recomputed the network based on re-scaled intensities of flow to see effect of uniform changes in nation economy. On FedEx Ground data, we evaluated scaling factors of 0.1 through 2.0 in steps on one-tenths. The most striking observation was the robustness of the network. As the flows were incremented, we found that the network did not change drastically. Over the 20 iterations of reconfiguring the network a terminal changed assignments at most 10 times between successive flow increases. During each iteration a maximum of 12 terminals of the 361 terminals (less than 4%) changed the assignments. Amongst all the possible 190 combination pairs of these assignments we found that at most 12 terminals changed assignments. 320 of the 361

terminals (about 89%) remained assigned to the same hub independent of the scaling factor, that is, never changed assignments.

About 98% of the shipments are less than 12% of the trailer size. Which means that if the shipments are scaled by a factor of 2, these shipments will still remain less-than-trailerload shipments. The LTL shipment size characteristics hedge the network against perturbations in flow.

2.8 Returns to Scale

The LTL trucking industry exhibits *constant* returns to scale [Thomas and Callan, 1989], which implies that there is not added benefit to a firm only because of its size. This is contrary to the observed trend towards a more concentrated industry in the post-deregulation period. However, this sector is very competitive and larger shippers provide intense competition based on "service" to smaller carriers. Emerson, Grimm, and Corsi [1992] find a mildly positive relationship between firm size (as measured by revenue ton-miles in a base period) and commercial success (as measured by the probability of market share growth of a given firm-size class over time). To explain this observation McMullen and Tanaka [1995] suggest that as long as a LTL carrier has ability to take advantage of unexploited network economies it will continue to grow. Network economies result from the efficient use of a network system to increase load factors and simultaneously maintain service levels.

A regression analysis of the scaled cost (k_c) with the scaling factor (k_f) yielded the following linear fit:

$$k_c = 0.1663 + 0.8244 k_f \qquad (k_f > 0) \tag{5}$$

with R-squared value (adjusted) of 99.9%.

The linear coefficient is less than 1 and so for an incremental unit increase in the flows the cost increases by less than an unit. This strongly supports the observation by Emerson et al.. Moreover, it also supports the suggestion proposed by McMullen and Tanaka that network economies will be exploited by increasing the firm size.

The industry estimates the coefficient k_f to be even smaller, typically in the range of 0.6 to 0.7. [pers. com. Trussel, 2002]. Our model yields a higher coefficient value of k_f essentially because we have implemented a continuous cost model for the longhaul costs.

We do not allow any direct loads in the heuristic. About 98% of the shipments are less than 12% of the trailer size. This means that if the shipments are scaled by a factor of 2, these shipments will still remain less-than-trailerload shipments and will be sorted at the hubs. If direct loads we allowed, increasing the shipment size may generate more direct loads either bypassing a hub entirely or by avoiding sorting at a hub. This may yield a coefficient in the industry expected range.

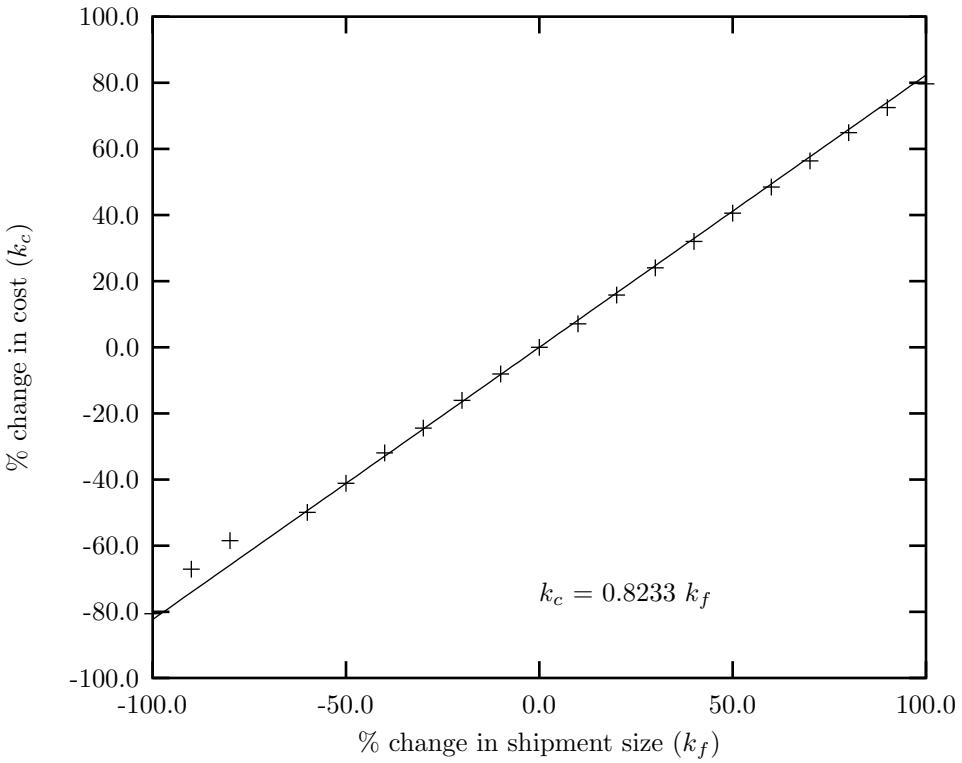


Figure 18: Increasing the shipment size by 1% increases the cost by about 0.82%

2.9 Marginal Cost of a Package

Another approach to determine the sensitivity of cost is to estimate the marginal cost of adding a package to a shipment. Since there are about 100,000 shipments in the network the marginal cost may significantly vary depending on the shipment to which the package is added. For example a shipment which has no excess capacity in a trailer will have significantly high marginal cost compared to a shipment that has excess trailer capacity. Instead of dealing with the entire input vector of shipments our focus will be on estimating the *average* marginal cost of adding a package to a shipment.

As we add more packages to the shipments the costs (transportaion and sorting) increase. A regression analysis of the increase in cost (δ_c) with the increase in total number of packages in the network (δ_f) yields the following linear fit:

$$\delta_c = c + 1.1864\delta_f \tag{6}$$

where,

δ_c = incremental increase in cost for adding δ_f packages to the network

c = a constant depending on the number of packages added to each shipment in the network.

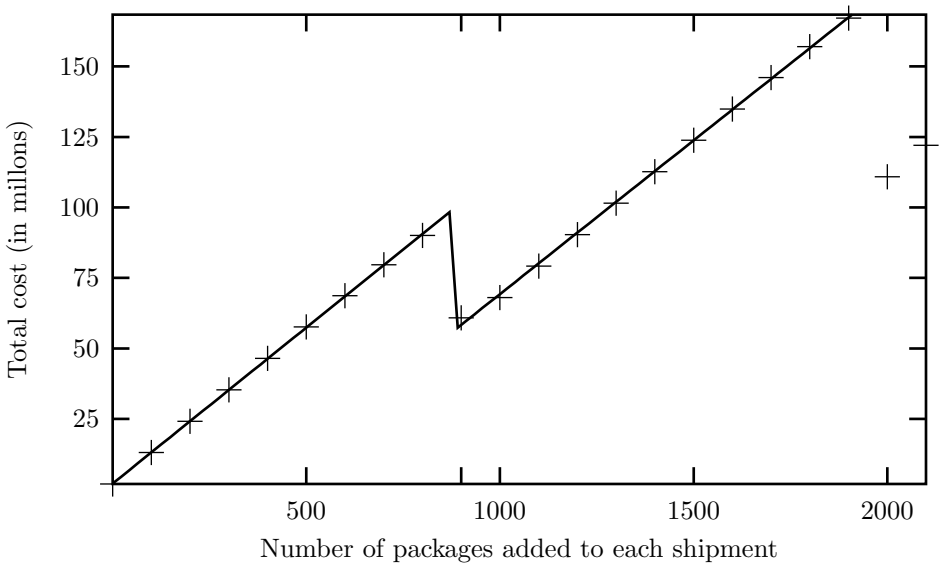


Figure 19: An additional package costs about \$1.19 on an average to the network.

Every package contributes \$1.1864 to the network cost.

The average distance between terminals is an estimated 1243 miles. Hence, the transportation cost per package, assuming a continuous cost model, is \$0.6210 (= 1243 miles $\cdot$ 1\$/mile/truck $\cdot$ $\frac{1}{2000}$ trucks/package). This is an underestimate for the transportation cost because of the continuous cost model on the shuttle and longhaul movements. About 89% of the packages are routed through two hubs and cost \$0.50 whereas 11% of the pacakges are routed through one hub and cost \$0.25 to sort. The expected sorting for a package is \$0.4725 (= $0.89 \cdot 0.50 + 0.11 \cdot 0.25$). The estimate of \$1.0935/package is slightly lower than the observed marginal cost per package of \$1.1864.

2.10 Sensitivity to Cost Parameters

There are two cost parameters based on which the terminal assignments are decided: sorting cost per package at the hubs (s_h) and the transportation cost per mile per truck (c). For the FedEx Ground data set, the sorting cost per package is the same at all the hubs, say s. We will assume that any changes in sorting cost per package will be identical for all the hubs.

For a fixed value of sorting cost, transportation cost increases, transportation costs dominate sorting costs and it becomes more important to route freight to reduce the average transportation cost per package. This results in a terminal being assigned to its nearest hub to reduce the distance traveled by partially empty trucks and feeding the longhaul trucks which are generally more utilized than trucks doing shuttle movements.

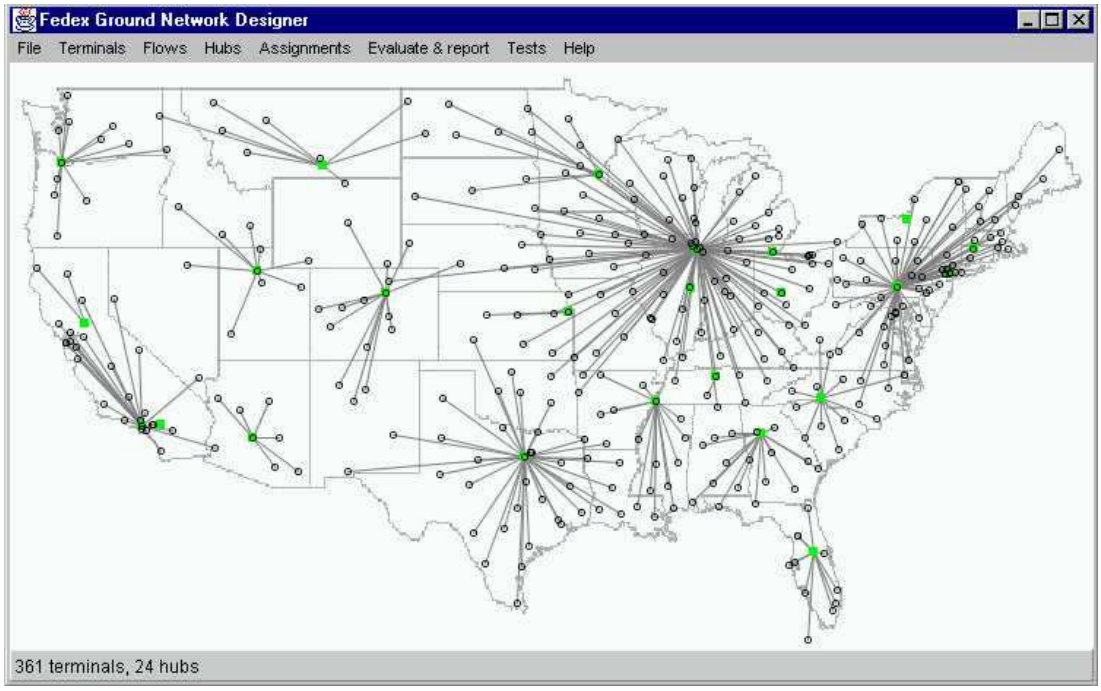
On the contrary, for a given value of c as s increases, sorting costs dominate transportation

costs. It may be economical to drive the lower utilized shuttle trucks over longer distances to save an additional sort for some shipments.

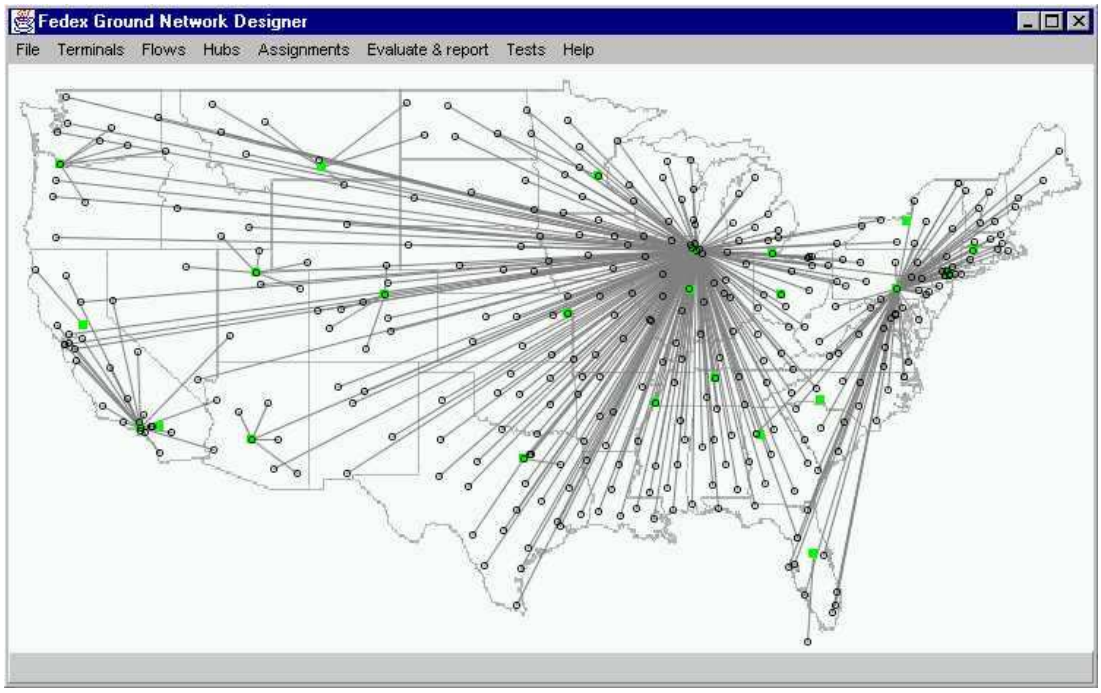
2.10.1 Speed Networks

One application of studying the sensitivity of cost parameters is designing *high-speed networks*. Most LTL carriers provide different levels of service to the customers. Some packages have to be expedited and typically this is done by avoiding sorting at one or both the hubs. One way to design such a high-speed network is to impose high penalty on sorting.

Figure 20 shows how regional *super-hubs* are created. To avoid the expensive second sort a terminal may be willing to send the partly utilized truck over longer distances and yet be economically justified.



(a) Sorting cost: 2.50$/package



(b) Sorting cost: 25.00$/package

Figure 20: High-Speed Networks: Increasing the penalty on sorting, a terminal may be willing to send lower utilized trucks over longer distances to avoid the second sort.

CHAPTER 3

SHIPMENT ROUTING

3.1 Problem Description

In the previous chapter we addressed the hub-and-spoke network design problem and a heuristic was proposed to assign each terminal to a hub. However, one of the simplifying assumptions made in the heuristic is that no direct loads were allowed. Because of this assumption, load planners cannot benefit from underlying patterns of freight flow. When direct loads are not permitted, under a single assignment policy, routing of shipments is trivial and all the shipments will be sent to and received from its assigned hub. All the less-than-trailerload shipments will be sorted at the hubs through which it passes. The simplicity in operations comes at the expense of inability to reduce sorting and transportation costs.

Despite variations in daily demands, definite patterns exist, within a season, in the flow of shipments. As a result direct loads can be identified and run on a regular basis. An objective of the shipment-routing problem is to identify pairs of locations (hubs and/or terminals) that have sufficient freight to justify regular direct runs. This is a step that assists *á posteriori* in the design of the service network.

Load planners seek opportunities to consolidate shipments to reduce transportation and sorting costs. Typically, load planners are required to identify direct loads. In some LTL carriers, the performance of terminal managers is also evaluated on the trailer utilizations, that is, the average number of pounds (packages) they put on trailers. In absence of a centralized load planning effort, terminal managers try to route shipments simply based on available capacity in a trailer rather than considering the more complex objective of minimizing cost. Since the decisions made by a terminal manager can affect decisions made at the other terminals and/or hubs in the network, it is essential to coordinate the efforts of all the terminal managers under the umbrella of a single loadplan. Routing shipments may seem to be operational level decisions, especially with a stochastic shipment demand. Due to the variations in demands it may seem less useful to generate a single loadplan that optimally routes shipments and satisfies the service constraints. But the idea is not to solve the model for everyday operations but to propose a loadplan which provides a guideline to the load planners for routing with the entire network in perspective; that is a centralized loadplan.

Another objective is to gain insights into the structure of routings generated so that fast heuristics maybe developed for routing shipments.

Another important use of identifying direct loads is to estimate hub utilization. If sorting capacities at a hub are based on a loadplan that routes all the shipments through the hubs and its spokes then it will result in a expensive under-utilization at the hubs when the terminal managers generate direct loads.

Centralized load planning also streamlines the decisions made by the load planners throughout the network. Though each load planner tries to reduce operating cost by building direct loads, unfortunately each load planner has a local perspective and a greedy operating policy may conflict with the other load planners within the network. For example, the load planner at origin terminal may want to hold on to shipments so that he can collect sufficient freight to fill up a truck and send it directly to the second hub, bypassing the first hub. However, the load planner at the hub to which the terminal is assigned may be planning to use that freight to fill a truck to send directly to the destination terminal, bypassing the second hub. Thus a decision taken by a load planner may have a cascading effect on load building throughout the network. Therefore, decentralized load planning may result in expensive global solutions.

Our focus will primarily be on building models to route freight and solving the models under a centralized setting for load planning. Since the shipment data is based on a five year forecast it will not be our emphasis to generate optimal solutions but rather near-optimal solutions as quickly as possible.

Powell [1986] proposes a local improvement heuristic to re-route freight throughout the network when links (direct services) are added or dropped from the network. Lamar and Sheffi [1987] approximate the inter-city costs as linear costs with fixed charge for determining the service frequency and shipment routes. They used the model to generate a series of heuristic solutions that also provide a lower bound to the total carrier costs. Powell and Sheffi [1989] and Powell and Koskosidis [1992] consider the load planning problem with tree constraints wherein all the shipments at terminal i headed for terminal s must move next to terminal j. In this formulation a shipment cannot be split over alternative paths. Roy and Delorme [1989] propose a non-linear mixed integer programing model to determine the service frequencies as well as the shipment routes. They reported results for networks with about 35 terminals. Leung, Magnanti, and Singhal [1990] formulate the routing problem as a mixed integer quadratic programing model. They treat the problem as consisting of two stages - an "assignment" subproblem and a mixed integer multi-commodity flow subproblem. They implemented Lagrangian relaxation-based techniques to solve each of the subproblems. Though

their work is similar to our research, they assume that a shipment cannot be split, that is, all parcels between the same origin-destination pair must use the same route. Rather than identifying direct loads, their focus is on identifying minimum cost route for a shipment. They assign shipments to hub pairs rather than the more conventional assignment of a terminal to hub in a hub and spoke framework. Akyilmaz [1994] provides an algorithmic framework to consolidate and route LTL shipments. The algorithm is aimed at minimizing the "net empty ton-miles" of a trailer. However, he neglects sorting costs at hubs. This algorithm can be viewed as a constructive tool to identify and generate spider-legs, which we ignore in this research. Lin [2001] addresses the freight routing problem for time-definite LTL carriers. Besides reducing the overall cost there are constraints which dictate the maximum between local sorts at origin and destination terminals. However, a shipment can be sent only by a single route. Kuby and Gray [1989] explore the trade-offs and savings involved in stop-overs and feeders in a air network. Based on their proposed mathematical models and computational experience they suggest that by considering stop-overs, substantial cost savings are achieved. Spider-legs in LTL networks are equivalent to stop-overs and feeders in air networks. We do not consider the possibility of spider-legs.

3.2 Assumptions

Assumption 1. We assume that a shipment from terminal i to terminal j, assigned to hubs k and l respectively, must be routed

$$i \rightarrow j \rightarrow k \rightarrow l$$

or a subset of this path.

If we do not restrict the shipment to be routed through the assigned hub then the shipment may be routed from i to its assigned hub k and then to j via hub k' which is not the assigned hub of j. For example, a shipment from Macon, GA to Pensacola, FL may first be routed to Atlanta, GA and then to Sacramento, CA before being sent to Pensacola (see figure 21(a)).

For example, a shipment from Macon, GA to Salem, OR may first be routed to Portland, OR and then to Atlanta, GA before being finally sent to Salem, as shown in figure 21(b)).

The reason that such unintuitive routings may be produced is because the marginal transportation cost of a shipment on a trailer with capacity is zero. As long as the trailer has capacity the shipment can be transported at no cost.

Assumption 3. We do not consider overnight transfer points in the formulation. We discard overnight freight that has to be sent through an OTP. Spider legs are not very common and

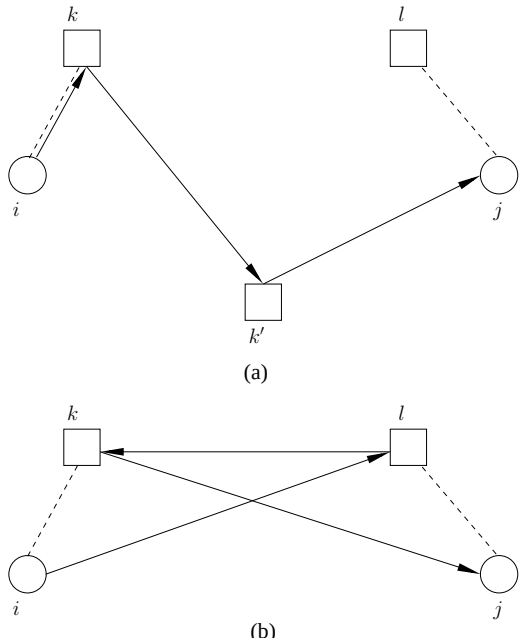


Figure 21: The routing model can generate very unintuitive routing to save costs.

are ignored in the formulation since they tend to make the formulation larger and complicated.

Assumption 4. We ignore head loads since these MIP models are used to assist load planners in freight routing and loading/unloading is a operational issue. Once the freight routes have been decided head loads can be further identified.

Assumptions 1 and 2 require us to provide a hub-and-spoke network to the model. That is each terminal will be *á priori* assigned to a one hub. The output from the heuristic described in Chapter 2 is the input to the model.

3.3 Hub-and-Spoke Based Shipment Routing

In this section we propose a freight routing model based on a hub-and-spoke infrastructure. Each terminal is pre-assigned to a single hub but an IP model is used to extract the cheapest direct loads.

Based on direct loads, direct runs and the assumptions listed in the previous section, we identified the following seven ways in which a shipment can be routed from the origin to the destination. Sorting costs are not negligible because of the volume of shipments sorted. Hence, deciding whether or not a shipment is sorted at a hub is important in addition to determining its route.

Consider terminals i and j assigned to hubs k and l respectively. A shipment from origin terminal i to destination terminal j can be routed in *at least* one of the following seven ways.

1. Directly from i to j, with no sorting cost.

2. Through hubs k and l respectively. It is not sorted at either hub. It travels from i to j in the same trailer which is coupled with other trailers bound to the destination.

3. Via a single hub k. It is not sorted at k. It travels in a trailer from i to k and this trailer is coupled with another trailer from k to j.

4. Through hubs k and l respectively. It is sorted at hub l but not at hub k. The package travels in a single trailer from i to l (via k) and possibly in a different trailer from l to j.

5. Through hubs k and l respectively. It is sorted at hub k but not at hub l. The package travels in a single trailer from k to j (via l) , possibly a different trailer from i to k.

6. Via a single hub k where it is sorted.

7. Through hubs k and l respectively. The package is sorted at both the hubs.

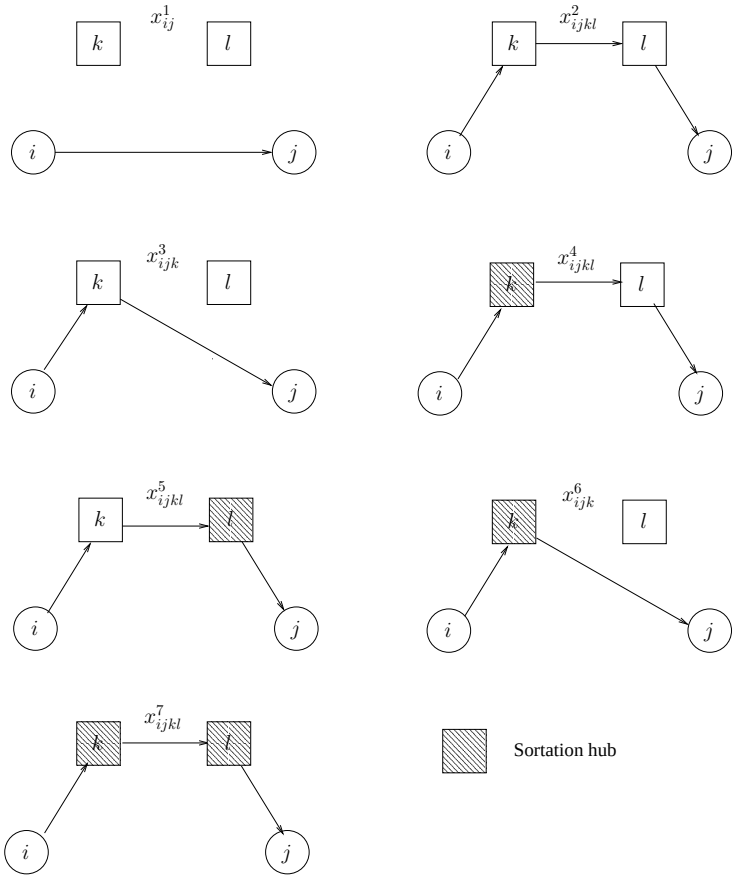


Figure 22: Various paths for routing a flow from terminal i to j.

Figure 22 shows the various ways in which a package can be routed from one terminal to another.

To ensure good service we restricted the routing of freight from i to j only through their assigned hubs. If this was not enforced, very unintuitive routings were generated which could save costs but provide poor service.

A single shipment can be sent by more than one of the above mentioned ways. For example, a shipment of 3200 packages may be sent from terminal i to terminal j by the following three ways:

1. 2000 packages (two trailers/one truck) sent directly from i to j.
2. 1000 packages (one trailer) from terminal i to j via hubs k and l. It is not sorted at either hub.
3. remaining 200 packages (partial trailer) from i to j via hubs k and l where its is sorted at both the hubs.

Though, the model captures the intricacies of the freight routing operation (providing freight route for each shipment) we believe that freight patterns can be identified. The primary objective of this paper is to provide mathematical tools to assist load planners in extracting patterns in freight flows by identifying direct loads and direct runs from the *pure* hub-and-spoke network.

3.4 Mixed Integer Programing Formulations

3.4.1 Notation

C: trailer capacity

We assume that each trailer has a fixed capacity based on the number of packages that can be loaded on it. Because of the strict restrictions on the size of the parcel this is a reasonable assumption.

D: minimum direct load size ($0 < D \leq C$)

We assume that a trailer can be sent directly only if it carries at least D packages.

c: cost per mile per truck

For convenience we assume this is the same on all routes.

f_{ij}: size of shipment (total number of packages) from terminal i to terminal j

d_{ij}: distance between location i and location j

s_k: estimated sorting cost at hub k

x^1_{ij} : packages shipped directly from terminal i to terminal j

x^2_{ijkl} : packages shipped from terminal i to j hubs k and l

There is no sorting cost incurred at hub. The trailer carrying these packages will be coupled with other trailer going from k to l and another trailer from l to j.

x^3_{ijk} : packages that are sent in a trailer from terminal i to hub k where they are not sorted but coupled with other trailer going from hub k to terminal j

x^4_{ijkl} : packages sent in a trailer from terminal i to hub k where they are sorted. The trailer is coupled with another trailer from k to hub l (no sorting takes place here) and sent to the final destination (terminal j) with another trailer.

x^5_{ijkl} : packages sent in a trailer from terminal i to hub k where they are not sorted

The trailer is coupled with another trailer from hub k to hub l where these packages are sorted and sent to the final destination (terminal j).

x^6_{ijk} : packages shipped from terminal i to hub k where they are sorted and sent to the final destination j

x^7_{ijkl} : packages sent from terminal i to terminal j via hubs k and l

Sorting takes place at both the hubs.

t^1_{ij} : number of trailers traveling directly from terminal i to terminal j

t^2_{ijkl} : trailer(s) shipped from terminal i to j via hubs k and l carrying flow x^2_{ijkl}

t^3_{ijk} : trailer(s) sent from terminal i to hub k carrying flow x^3_{ijk} where they are not sorted but coupled with other trailer going from hub k to terminal j

t^4_{klj} : trailer(s) from hub k to hub l (no sorting takes place here) sent to the final destination (terminal j) with another trailer.

t^5_{ikl} : trailer(s) carrying packages x^5_{i*kl} from terminal i to hub k where they are not sorted

The trailer is coupled with another trailer from hub k to hub l where the packages are sorted and sent to the final destination (terminal j).

t^6_{kj} : trailer(s) from hub k to terminal j carrying flow $x^6_{\cdot jk}$

t^6_{il} : trailer(s) from terminal i to hub l carrying flow $x^6_{i\cdot l}$

t_{ij} : total trailers from location i to location j

T_{ij} : total trucks from location i to location j

y_{ij} : binary variable which indicates if arc ij is utilized (1) or not (0)

If arc ij is not utilized then no flow route utilizing this arc is permitted.

Refer figure 22 for the x variables.

3.4.2 Original Formulation

3.4.2.1 Model (F_O)

$$\text{Minimize} \quad \overbrace{c\{\sum_i \sum_j d_{ij}T_{ij} + \sum_i \sum_k d_{ik}(T_{ik}+T_{ki}) + \sum_k \sum_l d_{kl}T_{kl}\}}^{\text{truck transportation costs}} + \underbrace{\sum_i \sum_j \sum_{k:k=h(i)} \sum_{l:l=h(j)} \{(x^4_{ijkl}s_l + x^5_{ijkl}s_k + x^7_{ijkl})(s_k+s_l)\}}_{\text{sorting costs for packages sorted twice}} + \underbrace{\sum_i \sum_j \sum_{k:k=h(i),h(j)} x^6_{ijk}s_k}_{\text{sorting costs for packages sorted only once}}$$

Subject to:

$$f_{ij} = x^1_{ij} + \sum_{k:k=h(i),h(j)} (x^3_{ijk} + x^6_{ijk}) + \sum_{k:k=h(i)} \sum_{l:l=h(j),l\neq k} (x^2_{ijkl} + x^4_{ijkl} + x^5_{ijkl} + x^7_{ijkl}) \tag{7}$$

$$x^1_{ij} \leq Ct^1_{ij} \qquad \forall i,j \in \mathcal{T} \tag{8}$$

$$x^2_{ijkl} \leq Ct^2_{ijkl} \qquad \forall i,j \in \mathcal{T}, k = h(i), l = h(j) \tag{9}$$

$$x^3_{ijk} \leq Ct^3_{ijk} \qquad \forall i,j \in \mathcal{T}, k \in \{h(i), h(j)\} \tag{10}$$

$$\sum_i x^4_{ijkl} \leq Ct^4_{klj} \qquad \forall i,j \in \mathcal{T}, k = h(i), l = h(j) \tag{11}$$

$$\sum_j x^5_{ijkl} \leq Ct^5_{ikl} \qquad \forall i,j \in \mathcal{T}, k = h(i), l = h(j) \tag{12}$$

$$Ct^6_{il} \geq \sum_{j:h(j)=l} x^6_{ijk} \qquad \text{if } l \neq h(i) \tag{13}$$

$$Ct^6_{kj} \geq \sum_{i:h(i)=k} x^6_{ijk} \qquad \text{if } k \neq h(j) \tag{14}$$

$$Ct_{kl} \geq \sum_{i:h(i)=k} \sum_{j:h(j)=l} (t^2_{ijkl} + t^4_{ikl} + t^5_{klj})C + \sum_{i:h(i)=k} \sum_{j:h(j)=l} x^7_{ijkl} \qquad \forall k,l \in \mathcal{H} \tag{15}$$

$$Ct_{ik} \geq \begin{cases} \sum_{j:h(j)=k} \{Ct^3_{ijk} + x^6_{ijk}\} & \text{if } k \neq h(i) \\ C \sum_j \{t^3_{ijk} + \sum_{l:l=h(j)} \{t^2_{ijkl} + t^5_{klj}\}\} + & \\ \sum_j \{x^6_{ijk} + \sum_{l:l=h(j)} \{x^4_{ijkl} + x^7_{ijkl}\}\} & \text{if } k = h(i) \end{cases} \tag{16}$$

$$Ct_{lj} \geq \begin{cases} \sum_{i:h(i)=l} \{Ct^3_{ijk} + x^6_{ijk}\} & \text{if } l \neq h(j) \\ C \sum_i \{t^3_{ijl} + \sum_{k:k=h(i)} \{t^2_{ijkl} + t^4_{ikj}\}\} + & \\ \sum_i \{x^6_{ijk} + \sum_{k:k=h(i)} \{x^5_{ijkl} + x^7_{ijkl}\}\} & \text{if } l = h(j) \end{cases} \tag{17}$$

$$t_{ij} \leq 2T_{ij} \qquad \forall i,j \in \mathcal{T} \cup \mathcal{H} \tag{18}$$

$$0 \leq t^1_{ij}, t^2_{ijkl}, t^3_{ijk}, t^4_{ijkl}, t^5_{ijkl} \qquad \text{integer} \tag{19}$$

$$0 \leq t_{ij}, T_{ij} \qquad \text{integer} \qquad \forall i,j \in \mathcal{T} \cup \mathcal{H} \tag{20}$$

$$0 \leq x^1_{ij}, x^3_{ijk}, x^6_{ijk}, \qquad \text{integer} \tag{21}$$

$$0 \leq x^2_{ijkl}, x^4_{ijkl}, x^5_{ijkl}, x^7_{ijkl} \qquad \text{integer} \tag{22}$$

3.4.2.2 Constraints

Constraint 7 ensures that all packages are delivered from the origin to the destination. Note that this constraint routes flow through the hubs to which the origin and/or destination terminal are assigned.

Constraints 8 – 12 are accounting constraints that determine the number of direct loaded trailers for each route type.

Constraints 15, 16 and 17 sum up the the total number of trailers between each location pair.

Constraint 18 determines the actual number of trucks (from the number of trailers) between a pair of locations. A tractor can pull two trailers.

Constraints 19 – 22 are the non-negativity and integrality constraints.

3.4.2.3 Model Size

If the network has $|T|$ terminals and $|H|$ hubs, the total number of constraints are $6|T|^2 + 2|H|^2 + 6|T||H| - 8|T| - 2|H|$ where as the model has $17|T|(|T|-1)+2|H|(|H|-1)+4|T||H|$ integer variables.

These values serve as an upper bound on the model size. The exact values depend on the number of terminals assigned to each hub which is instance specific. A formulation for 388 terminals and 24 hubs yields an upper bound of approximately 2.59 million variables and 0.96 million constraints. For our formulation the model had 1,841,342 variables and 820,178 constraints. Presolving reduced the model to 1,207,494 variables and 581,567 constraints.

3.4.3 Tightening Constraints

To tighten formulation F_O we add constraints 23 – 39 yielding formulation F_T. These additional constraints ensure that no flow is routed over any arc that is not available.

$$x^1_{ij} \leq f_{ij} \cdot y_{ij} \tag{23}$$

$$x^2_{ijkl} \leq f_{ij} \cdot y_{ik} \tag{24}$$

$$x^2_{ijkl} \leq f_{ij} \cdot y_{kl} \tag{25}$$

$$x^2_{ijkl} \leq f_{ij} \cdot y_{lj} \tag{26}$$

$$x^3_{ijk} \leq f_{ij} \cdot y_{ik} \tag{27}$$

$$x^3_{ijk} \leq f_{ij} \cdot y_{kj} \tag{28}$$

$$x^4_{ijkl} \leq f_{ij} \cdot y_{ik} \tag{29}$$

$$x^4_{ijkl} \leq f_{ij} \cdot y_{kl} \tag{30}$$

$$x^4_{ijkl} \leq f_{ij} \cdot y_{lj} \tag{31}$$

$$x^5_{ijkl} \leq f_{ij} \cdot y_{ij} \tag{32}$$

$$x^5_{ijkl} \leq f_{ij} \cdot y_{kl} \tag{33}$$

$$x^5_{ijkl} \leq f_{ij} \cdot y_{lj} \tag{34}$$

$$x^6_{ijk} \leq f_{ij} \cdot y_{ik} \tag{35}$$

$$x^6_{ijk} \leq f_{ij} \cdot y_{kj} \tag{36}$$

$$x^7_{ijkl} \leq f_{ij} \cdot y_{ik} \tag{37}$$

$$x^7_{ijkl} \leq f_{ij} \cdot y_{kl} \tag{38}$$

$$x^7_{ijkl} \leq f_{ij} \cdot y_{lj} \tag{39}$$

Constraint 40 ensures that no unused arcs are available. The positive cost coefficient for T variables ensures that if y_{ij} is 0 then T_{ij} is also 0.

$$y_{ij} \leq T_{ij} \tag{40}$$

Constraint 41 restricts the y variables to be binaries.

$$y_{ij} \in \{0,1\} \qquad \forall i,j \in \mathcal{T} \cup \mathcal{H} \tag{41}$$

Compared to formulation F_O this model has an additional $t(t-1)+h(h-1)+2th$ binary variables.

Observation 9 *Every solution feasible to F_T is also feasible to F_O but not every solution feasible to F_O is feasible to F_T.*

Formulation F_O without the tightening constraints is a valid formulation. However, it is a standard modeling practice to always include such constraints, as in formulation F_T, as it is well-known that F_T offers improved computational time than formulation F_O. We consider formulation F_O because preliminary computational results showed that the primal heuristic, discussed in section 3.7, generates more integer feasible solutions for formulation F_O than for F_T.

3.5 Direct Load Factor

In Chapter 2 we designed the heuristic based on the fact that no direct loads can be built. Hence, all shipments were routed through the hubs to which their origin and destination hubs were assigned. Against the industry notion to maximize direct trailer utilization, it is possible that sending a very lowly utilized trailer may be a cheaper alternative (see example 4.4).

3.5.1 Restricting Size of Direct Load

In the proposed model, we neglect truck utilizations on direct routes. Typically, LTL carriers impose minimum utilization requirements for direct trailers. FedEx Ground requires its direct trailers to be at least 75% utilized.

Constraints 42–49 ensure that at least D packages are sent in a direct trailer.

$$Dt^1_{ij} \leq x^1_{ij} \leq Ct^1_{ij} \tag{42}$$

$$Dt^2_{ijkl} \leq x^2_{ijkl} \leq Ct^2_{ijkl} \tag{43}$$

$$Dt^3_{ijl} \leq x^3_{ijl} \leq Ct^3_{ijl} \tag{44}$$

$$Dt^3_{ijk} \leq x^3_{ijk} \leq Ct^3_{ijk} \tag{45}$$

$$Dt^4_{jkl} \leq \sum_i x^4_{ijkl} \leq Ct^4_{jkl} \tag{46}$$

$$Dt^5_{ijk} \leq \sum_j x^5_{ijkl} \leq Ct^5_{ijk} \tag{47}$$

$$Dt^6_{il} \leq \sum_j x^6_{ijl} \leq Ct^6_{il} \tag{48}$$

$$Dt^6_{kj} \leq \sum_i x^6_{ijk} \leq Ct^6_{kj} \tag{49}$$

3.6 Computational Strategies

Hardware and OS

The computations were performed on Sun Ultra 80 Model 2450, 2x450-MHz UltraSPARC-II, 4-MB L2 Cache, 1-GB Memory running on Solaris 8.

Software

We used CPLEX 7.5 to solve the MIP formulations.[1]

All the computational experiments were performed on a network with 9 terminals and 3 hubs. This is a tiny network compared to the actual LTL network instances in mainland USA. Instead of relying on the various default strategies provided by CPLEX for growing the branch-and-bound tree we experimented with the various strategies for node and variable selection as well as cut generation and implemented the strategy best suited for the structure of our problem. The selected strategy is italicized in Table 6.

To measure the quality of a integer feasible solution we use the following terms:

z^{opt}_{IP} = least cost feasible routing solution

z^{opt}_{LP} = best lower bound on the cost for routing shipment within the sub-network.

z^{best}_{IP} = cost associated with the current feasible routing solution and

[1]We would like to express our thanks to Ilog for their software license support of CPLEX.

z_{LP}^{best} = best available lower bound on the cost for routing shipment within the sub-network.

One way to measure the quality of a integer feasible solution is by *integrality gap* given by,

$$\text{integrality gap} = z_{IP}^{opt} - z_{LP}^{opt}$$

This is an absolute measure and its significance depends on the cost parameters. The quality of a feasible solution can also be measured by the *fractional integrality gap remaining* defined by,

$$\text{fractional integrality gap remaining} = \frac{z_{IP}^{opt} - z_{LP}^{best}}{z_{IP}^{opt} - z_{LP}^{opt}}$$

Since the values for z_{IP}^{opt} is not always known, the fractional integrality gap remaining is approximated by,

$$\text{approximate fractional integrality gap remaining} = \frac{z_{IP}^{best} - z_{LP}^{best}}{z_{IP}^{best} - z_{LP}^{opt}}$$

This is a reasonable approximation and is well-defined. As the branch-and-bound progresses, the accuracy of the approximation increases. CPLEX uses *relative gap* to measure the quality of the ineteger solutions. The relative gap is calculated as,

$$\text{relative gap} = \frac{z_{IP}^{best} - z_{LP}^{opt}}{z_{IP}^{best}}$$

We will use relative gap to measure the quality of our integer solutions.

3.6.1 Node Selection

Depth first search is a good strategy to find the first integer solution very quickly. But it is a very expensive strategy to find a good solution if it keeps searching the part of the tree with bad solutions. The best bound node selection strategy provides the best lower bound. But the best estimate strategy selects the node with the best estimate of the integer objective value that would be obtained from a node once all the integer infeasibilities are removed. The results are summarized in Table 6(a).

3.6.2 Branching Variable Selection

The variable selection strategy based on minimum integer infeasibility performs better than maximum integer infeasibility because in the maximum integer feasibility strategy, we round-up (or round-down) a trailer or truck which is half full. Rounding up this trailer generates excess capacity on the trailer and hence does not generate cheaper solutions. On the contrary rounding down the trailers forces approximately half of the trailerload shipments to be re-routed in an alternative more expensive route. Strong branching generates good solutions but is computationally very expensive

especially when the LP relaxations are not easy to solve. Table 6(b) tabulates the branch-and-bound progress for each variable selection strategy. For details on each strategy the interested reader may refer Lee [2001].

3.6.3 Cuts

Our computational experiments suggested that mixed integer rounding cuts (MIR) were the most effective in improving the objective at the root node. They also yielded the best LP bound of all the cut strategies and were used in the branch-and-bound tree. Disjunctive and Gomory fractional cuts improved the root objective substantially but were not effective in general. However, since Gomory fractional cuts were the least computationally expensive, they were also used in the branch-and-bound optimization. The effect of each type of cut is provided in Table 6(c).

For details on each of these cuts, an interested reader may refer Wolsey [1998].

Table 6: Comparison of various strategies on the performance of branch and bound tree for an instance with tightening constraints.

(a) Node selection strategy

Strategy	Nodes solved	% Gap	Best IP objective	LP bound	Solution time
Best estimate	*500,000*	*3.61*	*24,020*	*23,152*	*10,345*
Best bound	500,000	8.45	25,425	23,278	12,586
Alternate best estimate	500,000	9.31	24,319	22,055	7,522
Depth first search	500,000	20.83	27,858	22,055	3,748

(b) Branching variable selection strategy

Strategy	Nodes solved	% Gap	Best IP objective	LP bound	Solution time
Pseudo costs	*500,000*	*3.61*	*24,020.5*	*23,152.3*	*10,703*
Pseudo Red costs	500,000	4.25	24,073.9	23,049.8	7,595
Strong branching	61,500	4.38	24,257.5	23,195.3	21,600
Min integer infeasible	500,000	21.55	28,280.9	22,186.5	5,117
Max integer infeasible	500,000	66.24	65,697.3	22,177.0	5,634

(c) Cut strategy

Cuts	Number of cuts	Improvement in root objective	Nodes solved	% Gap	Best IP objective	LP bound	Solution time
Mixed Integer Rounding	*910*	*507*	*15,700*	*3.88*	*24,109*	*23,173*	*16,164*
GUB covers	0	0	97,500	6.43	24,185	22,629	21,600
Flow path	0	0	92,200	6.48	24,185	22,617	15,326
Cliques	0	0	86,000	6.54	24,185	22,602	21,600
Covers	7	0	80,900	6.95	24,304	22,615	3,669
Flow covers	143	92	86,600	7.00	24,515	22,800	20,595
Disjunctive	49	400	81,100	8.22	24,734	22,702	14,467
Gomory Fractional	*44*	*409*	*70,000*	*9.25*	*25,013*	*22,699*	*7,983*

3.6.4 Primal Heuristic

To obtain integer solutions from a node LP solution we implemented a primal node heuristic which exploits the structure of the problem. This heuristic is described in the next section.

3.6.5 Termination Criteria

In our computational results we constrained the tree size to 3 GB and optimization time to 6 hours. For evaluating the performance of various strategies we limited the search in branch-and-bound tree to a 500,000 solved nodes. However, to evaluate the performance of the heuristic and the effectiveness of the cuts on the two formulations (F_O and F_T), we restricted the tree search to 250,000 nodes. In most cases, time was a stopping criterion rather than the number of nodes solved.

3.7 Primal Heuristic

When solving the MIP using CPLEX the most striking observation was the fact that very few integer feasible solutions were generated. In fact, at any node we can very easily generate an integer feasible solution, if one exists. Consider a node in the branch-and-bound tree where an integer feasible solution exists. If the node solution is integer infeasible then the following can be said about the solution:

- All the flow variables are integer since the shipment size is integer
- One or more of the truck or trailer variables are non-integer

This primal node heuristic is an extension of the heuristic for 0/1 IP problem introduced by Bixby, Cook, Cook, and Lee [1999]. Since an integer feasible solution exists and all flow variables are integer, each package is shipped unsplit from its origin to its destination. To generate an integer feasible solution we fix the flow variables and then calculate the number of trailers and trucks required to route the packages. It has to be observed, that this procedure yields an integer feasible solution for any feasible routing of the shipments. The rounding-up procedure involves the following three steps:

1. Fixing the flow variables at the node LP values.
2. Calculating the trailer variables and rounding them up.
3. Calculating the truck variables and then rounding them up.

If the new integer feasible solution has to be useful, it has to be cheaper than the existing best integer feasible solution. The simple rounding up procedure may generate none or very few useful

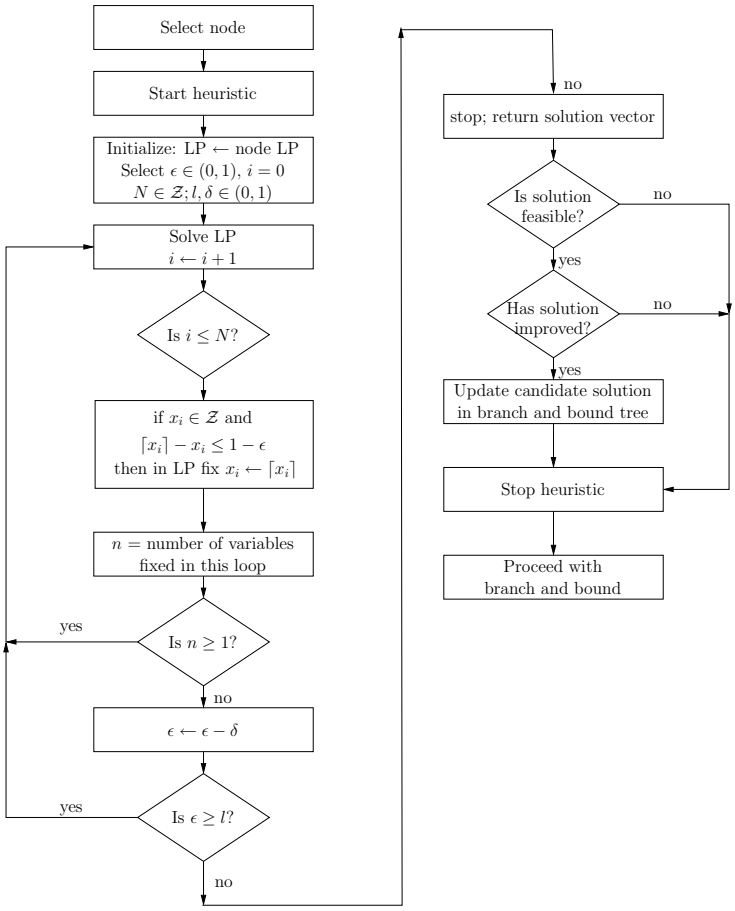


Figure 23: Flow chart for the primal heuristic implemented for branch and bound

integer feasible solutions as rounding up the trailers and trucks decreases the utilization of the trucks in the new solution. To utilize the trucks better the rounding up is limited to those trucks that are almost full. Once the few almost full trucks are fixed, we try to re-route the packages from the nearly empty trucks to the nearly full trucks trying to eliminate nearly empty trucks. If no more packages can be rerouted then the rounding up procedure is applied.

When we initialize the heuristic we clone the node LP and the original IP. If the truck utilization is greater than ϵ then the truck is considered to be almost full. Initially, ϵ is set to ϵ_0. We solve the node LP. All trailers and trucks that are almost full are fixed to their rounded up values. If there are no trailers or trucks that are almost full then ϵ is decremented by a small amount, δ. With new bounds on the fixed trailer and truck variables, the node LP is resolved. This is continued until either the node LP is solved a prespecified maximum number of times or ϵ falls below the lower threshold, ϵ_l, that defines an almost full truck. For example, $\epsilon_l = 0.70$ means that any truck has

Algorithm 6 Node Heuristic

Parameters: nodeLP, rootIP, ϵ_0, N, ϵ_l, δ
Require: $0 < \delta < \epsilon_l < \epsilon_0 < 1 \wedge N > 1$
Initialize:
LP $\leftarrow$ nodeLP; IP $\leftarrow$ rootIP
$\epsilon \leftarrow \epsilon_0$; $i \leftarrow 0$
V_T: set of trailer and truck variables
V_F: set of flow variables
repeat
Solve LP
$i \leftarrow i + 1$
repeat
$n \leftarrow 0$
for all variables, $x \in V_T$ **do**
if $(\lceil x \rceil - x \leq 1 - \epsilon)$ **then**
Fix $x \leftarrow \lceil x \rceil$ in LP
$n \leftarrow n + 1$
end if
end for
if $n = 0$ **then**
$\epsilon \leftarrow \epsilon - \delta$
end if
until $n > 0$
until $(i > N \vee \epsilon < \epsilon_l)$
for all variables, $x \in V_F$ **do**
Fix x in IP
end for
Solve IP
if IP objective is lesser than the best integer solution **then**
replace incumbent best solution with current solution;
end if
continue with branch and bound

to be utilized at least 70% for it to be considered sufficiently full. Then the flow variables are all fixed in a clone of the original problem and solved as an IP. This essentially is just an accounting step, calculating the number of trailers and trucks required. But solving it as an IP and getting the solution is much easier to implement than accounting.

Note that the rounding-up loop may make the problem infeasible since rounding up and fixing the truck and trailer variables may violate some pre-existing bounds or constraints.

3.8 Computational Results

We performed preliminary computations on smaller sub-networks extracted from the FedEx data set. We selected 9 sub-network; with 9, 14, and 25 terminals and 3, 4 and 5 hubs respectively, to understand how increasing the size of the network by increasing the number of terminals and hubs influences the computational performance of shipment-routing problems. We consider both

the models, F_O and F_T, for each of the sub-networks. In table 7 we provide the size of each of these instances along with the best available upper and lower bounds for each of the instances. The termination criteria are discussed in section 3.6.5.

It is clear that adding tightening constraints increases the problem size considerably. However, the increase in the number of variables is less significant than the increase in the number of constraints. The number of constraints are approximately tripled.

Table 7: Problem statistics of the instances (presolved)

name	Terminals	Hubs	Cols	Rows	Continuous	0/1	General	IP objective	LP objective	Gap
9t3ho	9	3	480	373	54	0	426	17,214.6	7,259.7	9,954.9
9t3ht	9	3	556	987	54	76	368	17,214.6	14,034.2	3,180.4
9t4ho	9	4	666	522	72	0	594	34,497.1	19,008.9	15,488.2
9t4ht	9	4	768	1,398	72	102	594	34,497.1	27,784.1	6,713.0
9t5ho	9	5	692	–	70	0	622	29,955.0	23,438.5	6,516.5
9t5ht	9	5	800	1,429	70	108	622	29,955.0	28,763.1	1191.9
14t3ho	14	3	1,390	1,011	171	0	1,219	35,118.8	14,705.3	20,413.5
14t3ht	14	3	1,595	2,880	171	205	1,219	35,118.8	32,173.7	2,945.1
14t4ho	14	4	1,460	1,094	171	0	1,289	38,787.4	16,540.7	22,246.7
14t4ht	14	4	1,671	3,065	171	211	1,289	38,787.4	34,163.5	4,623.9
14t5ho	14	5	1,527	1,172	171	0	1,356	42,618.4	15,717.2	26,901.2
14t5ht	14	5	1,746	3,210	171	219	1,356	42,618.4	35,970.8	2,816.6
25t3ho	25	3	4,444	3,050	572	0	3,872	191,902.1	121,814.7	70,087.4
25t3ht	25	3	5,072	9,458	572	628	3,872	191,902.1	177,691.1	14,211.0
25t4ho	25	4	4,590	3,204	572	0	4,018	179,332.3	102,505.6	76,826.7
25t4ht	25	4	5,222	9,741	572	632	4,018	179,332.3	158,854.9	20,477.4
25t5ho	25	5	4,716	3,350	572	0	4,144	181,265.1	105,174.4	76,090.7
25t5ht	25	5	5,358	10,236	572	642	4,144	181,265.1	162,824.6	18,440.5

We present our computational results in tables 8–10. For each problem instance we have the following associated values:

Cuts used: To indicate whether cuts were used while solving this instance.

Heuristic used: To indicate whether the primal heuristic was used while solving this instance.

MIP gap: The fractional integrality gap remaining when termination criteria is reached.

Nodes solved: The number of nodes explored in the branch-and-bound tree until the termination citeria is reached.

Nodes remaining: The number of unsolved nodes termination criteria is reached.

Time: The time until the termination criterion is reached for the optimization.

z_{LP}^{best}*:* The best available lower bound when optimization is stopped.

Time until best solution: Time taken to obtain the first least cost integer solution available when optimization terminates.

Number of integer solutions: The total number of integer solutions found for the problem during the course of the optimization.

Table 8: Computational Results: 9 terminals

Instance	Cuts used	Heuristic used	MIP Gap	Nodes Solved	Nodes Remaining	Time	Best Solution	z_{LP}^{best}	Time until best solution	Number of integer solutions
9t3ht	√	√	3.93%	14,900	6,932	21,600	24,109	23,161	18,227	10
9t3ht	√	×	3.68%	17,800	6,933	21,600	24,052	23,167	14,516	10
9t3ht	×	×	6.74%	89,600	47,099	21,600	24,250	22,614	18,697	8
9t3ht	×	√	6.51%	90,000	43,855	21,600	24,185	22,611	16,674	14
9t3ho	√	√	14.55%	84,100	47,428	21,600	24,719	21,118	18,976	37
9t3ho	√	×	15.21%	84,000	53,685	21,600	24,788	21,018	19,402	22
9t3ho	×	×	29.59%	250,000	164,566	9,559	26,736	18,824	3,380	15
9t3ho	×	√	28.14%	250,000	175,993	9,387	26,184	18,817	5,916	18
9t4ht	√	√	14.77%	36,600	31,365	21,600	34,893	29,739	3,688	9
9t4ht	√	×	14.67%	37,200	30,311	21,600	34,866	29,750	20,683	10
9t4ht	×	×	20.22%	130,400	101,397	21,600	35,672	28,460	16,010	11
9t4ht	×	√	20.49%	128,600	99,129	21,600	35,794	28,459	20,105	13
9t4ho	√	√	18.08%	173,400	101,801	21,600	34,170	27,992	10,873	27
9t4ho	√	×	17.39%	153,300	98,697	21,600	33,765	27,893	11,673	16
9t4ho	×	×	31.59%	250,000	191,852	9,157	36,551	25,004	8,869	42
9t4ho	×	√	32.82%	250,000	221,006	11,101	37,175	24,976	364	6
9t5ht	√	√	4.24%	24,600	6,181	21,600	30,037	28,763	17,745	23
9t5ht	√	×	4.09%	23,100	6,474	21,600	29,955	28,731	7,519	20
9t5ht	×	×	7.75%	136,900	51,920	21,600	30,326	27,974	20,746	37
9t5ht	×	√	7.44%	124,500	35,094	21,600	30,086	27,849	20,670	33
9t5ho	√	√	17.56%	244,100	94,562	21,600	31,380	25,868	8,881	47
9t5ho	√	×	19.16%	229,400	87,437	21,600	32,013	25,878	20,428	65
9t5ho	×	×	32.44%	250,000	240,436	10,419	31,930	21,570	5,595	35
9t5ho	×	√	32.80%	250,000	241,165	11,736	32,091	21,565	2,047	6

Table 9: Computational results: 14 terminals

Instance	Cuts used	Heuristic used	MIP Gap	Nodes Solved	Nodes Remaining	Time	Best Solution	z_{LP}^{best}	Time until best solution	Number of integer solutions
14t3ht	√	√	5.44%	6400	3112	21,600	34667	32782	19142	16
14t3ht	√	×	6.20%	8700	4558	21,600	34908	32744	20500	20
14t3ht	×	×	16.68%	40700	33719	21,600	38452	32040	2076	3
14t3ht	×	√	10.05%	33600	23350	21,600	35646	32065	8107	5
14t3ho	√	√	19.32%	26200	21384	21,600	35662	28773	2902	4
14t3ho	√	×	21.81%	27400	19023	21,600	36527	28560	8621	7
14t3ho	×	×	65.02%	250000	227089	18583	70222	24542	5347	3
14t3ho	×	√	56.62%	250000	221706	20922	56631	24566	9698	3
14t4ht	√	√	11.56%	6900	5939	21,600	47192	41736	3796	3
14t4ht	√	×	14.45%	9200	7343	21,600	48696	41659	9522	3
14t4ht	×	×	25.11%	43900	36015	21,600	54408	40747	11711	5
14t4ht	×	√	27.65%	37300	33513	21,600	56269	40711	2134	2
14t4ho	√	√	30.17%	31900	26187	21,600	55875	39019	4679	5
14t4ho	√	×	22.60%	29200	22980	21,600	50402	39013	5131	4
14t4ho	×	×	36.89%	207500	165809	21,600	54855	34618	14545	4
14t4ho	×	√	31.66%	220500	190991	21,600	50598	34579	6654	12
14t5ht	√	√	13.61%	6100	5026	21,600	44732	38643	11082	2
14t5ht	√	×	9.83%	7300	4730	21,600	42775	38569	15871	5
14t5ht	×	×	24.39%	39800	36908	21,600	50476	38164	14762	3
14t5ht	×	√	16.45%	36300	28152	21,600	45554	38061	19615	3
14t5ho	√	√	35.32%	31700	30094	21,600	47192	30524	1781	5
14t5ho	√	×	36.13%	33800	31759	21,600	47115	30094	15351	7
14t5ho	×	×	67.91%	239900	233241	21,600	79591	25538	1930	7
14t5ho	×	√	67.63%	221800	207328	21,600	79234	25651	12442	7

Using the heuristic helps produce greater number of integer solutions. Typically, the heuristic is more effective as the size of the network decreases. For most cases, the heuristic generates more integer solutions for the formulation F_O than F_T. This is because the constraints 23–40 may result in infeasibility in the rounding up loop. Since the LP relaxation for formulation F_O is easier to solve than that of F_T, more nodes were solved for formulation F_O than for formulation F_T.

As the size of the network increases the problems become more difficult to solve. For a network with t terminals and h hubs, consider the addition of another terminal. For simplicity assume that this terminal communicates with all the t terminals. Each of these $2t$ shipments have about 3 to 9 possible flow routes as shown in figure 22 each with its set of constraints to account for the number of trailers and trucks. There are additional $2t + 2h$ truck and trailer variables. However, instead if a hub is added to this network, the existing assignments may be changed and some of the terminals which were assigned to the same hub will now be possibly assigned to different hubs. So for some shipments which had only 3 possible routes in the original network now have 9 possible shipment

Table 10: Computational Results: 25 terminals

Instance	Cuts used	Heuristic used	MIP Gap	Nodes Solved	Nodes Remaining	Time	Best Solution	z_{LP}^{best}	Time until best solution	Number of integer solutions
25t3ht	√	√	19.46%	8600	8262	21,600	200208	161246	3944	1
25t3ht	√	×	25.57%	9200	8930	21,600	216546	161169	3608	1
25t3ht	×	×	33.19%	14000	13349	21,600	237169	159749	2883	1
25t3ht	×	√	32.04%	13900	13192	21,600	233481	158670	3137	1
25t3ho	√	√	51.36%	22800	21052	21,600	299328	145600	715	2
25t3ho	√	×	53.79%	24600	22793	21,600	315077	145608	608	1
25t3ho	×	×	55.17%	70400	63310	21,600	312456	140066	375	1
25t3ho	×	√	55.19%	70100	63077	21,600	312376	139961	662	1
25t4ht	√	√	38.02%	8300	8086	21,600	277565	172051	4229	1
25t4ht	√	×	34.85%	7400	7185	21,600	264150	172096	4373	1
25t4ht	×	×	33.20%	12000	11532	21,600	251289	167862	3281	1
25t4ht	×	√	35.10%	12200	11672	21,600	258508	167765	3450	1
25t4ho	√	√	52.19%	23500	21556	21,600	332301	158863	729	2
25t4ho	√	×	52.89%	23300	21549	21,600	337276	158886	692	1
25t4ho	×	×	58.00%	68300	62738	21,600	363047	152472	394	1
25t4ho	×	√	58.79%	59600	53556	21,600	369902	152431	1185	2
25t5ht	√	√	32.38%	9500	9227	21,600	236669	160033	3444	1
25t5ht	√	×	31.59%	8800	8486	21,600	233898	160018	3682	1
25t5ht	×	×	37.98%	14200	13659	21,600	251656	156066	3105	1
25t5ht	×	√	34.77%	14400	13875	21,600	239270	156072	3328	1
25t5ho	√	√	60.98%	23900	21844	21,600	375346	146460	698	2
25t5ho	√	×	58.78%	22500	20785	21,600	355255	146451	705	1
25t5ho	×	×	60.86%	61300	56338	21,600	354117	138609	436	1
25t5ho	×	√	61.38%	57200	51729	21,600	358783	138579	877	3

routes. The number of truck and trailer variables only increases by $2h$. Moreover, adding any LTL shipment implies that the LP relaxation uses fractional trucks yielding very poor lower bounds. The computational results presented show that adding a terminal to a network makes the problem much more difficult than adding a hub. However, since the set of assignments are provided for a terminal, keeping the size of the network small is equivalent to selecting fewer hubs in the networks.

It can be seen that for some instances by using the heuristic more nodes were solved than the case when no heuristic was implemented. This may be because once the heuristic has found an integer feasible solution, the search has been moved to the part of the branch-and-bound tree where solving the LPs are slightly easier. Even a slight decrease in solving LPs may result in solving a significant number of additional nodes.

We highlight the following key observations from our computational results:

1. Using the tightening constraints improves the branch-and-bound performance by improving lower bounds and proving quality of of solutions.

2. As the size of the network increases, the heuristic almost always helps generate provably better solutions and yields the best solution early on in the the search tree.

3. Implementing MIR and Gomory fractional cuts substantially helps to improve the lower bound.

CHAPTER 4

SHIPMENT ROUTING – NETWORK DECOMPOSITION AND PARALLELIZATION

4.1 Problem Description

Empirical results presented in the previous chapter suggest that freight routing is an extremely difficult problem even for networks with only 25 terminals and 5 hubs. The mixed integer programing model for the entire network has about 1.2 million variables and 0.6 million constraints. Currently available commercial solvers are not able to solve the LP relaxation (root relaxation) after 8 hours. Thus it is not possible to solve the entire problem with existing branch-and-bound techniques. Two factors regarding the solution approach for routing shipments through a network that is of the same size as of FedEx Ground network (361 terminals and 25 hubs) are:

1. Determining optimal routing of freight through the entire network using current optimization techniques is intractable.

2. Near-optimal solutions for routing of shipments through small size networks are readily available.

Based on this information we consider decomposing the network into small sub-networks within each of which shipments can be routed to near-optimality.

Although there are several ways to decompose the network, we focus on the decomposition technique which allows for efficient reconstruction of a global feasible solution from the solutions of the sub-networks.

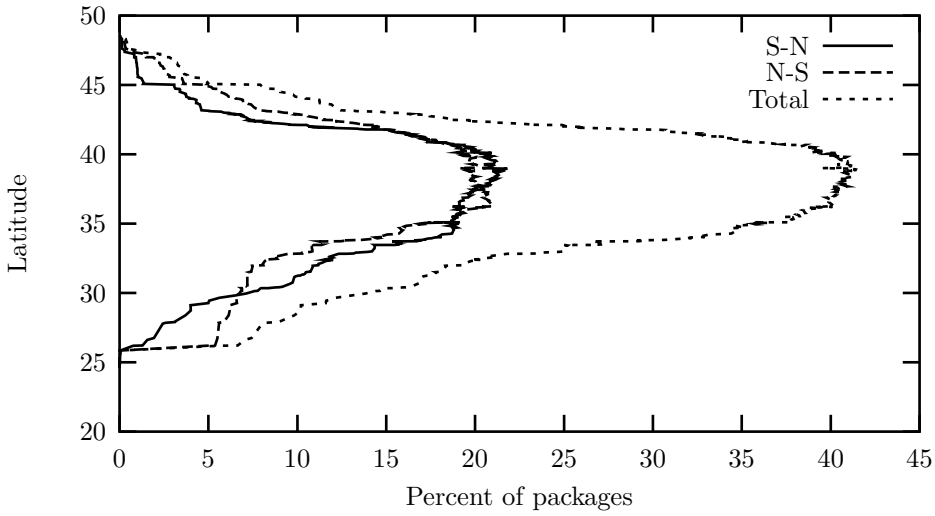
4.2 Difficulties in Network Decomposition

One way to decompose the network is based on geographic information. Figure 24 shows the percentage of shipments that travel across a particular latitude and longitude. The plots are unimodal. Let us consider the following example.

Example 4.1 *Approximately 5% of shipments cross the 43° latitude north to south and 10% of shipments cross the 42° latitude north to south. This means that between 42° and 43° latitudes*

approximately 5% of the total packages originate more than those delivered. As we move towards the mode latitude, the latitude corresponding to the peak in the graph, more packages originate than terminate. And as we move away from the mode latitude more packages are delivered than picked up.

(a) Flow across latitudes



(b) Flow across longitude

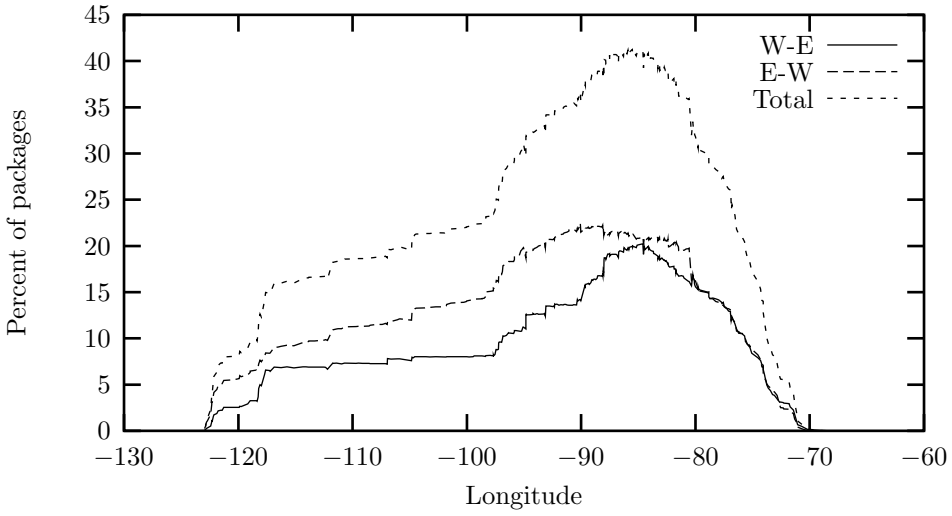


Figure 24: Flow of shipments across latitudes and longitudes

The unimodularity enforces the idea that it is not possible to decompose the network based on simple geography, such as north-south or east-west.

Even if the plot (figure 24) was not unimodal but multi-modular, it is a non-trivial task to decompose the network. For example, one can divide the network into four distinct sub-networks, such as – south, east, north and west. However, because every terminal communicates with almost every other terminal it becomes harder to extract freight flow information that is important from the perspective of the entire network but is less useful for the smaller sub-network. Let us consider again the situation in which we divide the network into four sub-networks. It is possible to route the shipments for all the terminals within a sub-network, for example east zone. However, terminals in the east zone also communicate with the terminals in the other three zones, north, west and south. Routing the shipments for the terminals in east zone only gives us information about shipment routes that originate and terminate within the same zone, however, it does not provide information about the routes of shipments that originate *or* terminate between terminals that are in different zones. So one approach is to discard direct loads for the terminals that are not in the same zone which defeats the purpose of the proposed routing model.

Another approach is to shrink all the hubs within another zone into a *pseudo hub* and locate this pseudo hub. By shrinking the hubs into a pseudo hub one loses information about the distances between the hubs that are shrunk. Mapping these distances can be difficult and hence locating this pseudo hub is a non-trivial task. Moreover, the concept of the pseudo hub is not particularly useful since by shrinking all the hubs into a single hub, the shipments are consolidated onto a single *pseudo truck*. Since all the direct load shipments (if any are detected) are sent to the pseudo hub, the problem of extracting information about individual shipments from the pseudo truck still needs to be resolved .

In order to efficiently construct a solution to the entire network from the sub-networks, one needs to specify a route for every shipment in the network. Thus a shipment has to be part of at least one sub-network.

Decomposition Requirement 1 *Every terminal origin – destination pair has to be covered in at least one sub-network.*

Note that if an origin – destination pair is covered by more than one sub-network then it may lead to inconsistent routing during the reconstruction phase of a global feasible solution for the entire network.

Example 4.2 *Let terminals i and j be assigned to hubs k and l respectively. Assume 1000 packages (a trailer-load) were sent from i to j. If k and l are both covered in two sub-networks, N_1 and N_2, then the following scenarios are possible:*

- *In N_1, the trailer is sent from i to j with flow x^2_{ijkl} (see figure 22, page 62); that is, via hubs k and l without being sorted at either hubs*
- *In N_2, the trailer is sent from i to j as flow x^3_{ijk} (figure 22); that is, via hub k bypassing hub l without being sorted at hub k*

When one attempts to construct a feasible solution for the entire network, it is unclear which of the two routings is cheaper when the entire network is considered. Although one can re-optimize it is easy to see that it is possible to obtain many such pairs of solution which will lead to a difficult problem on its own. The example presented above stresses that more considerations have to be included when decomposing the network. Not only that all the origin – destination pairs need to be covered, the overlap of origin – destination pairs should also be minimized.

Decomposition Requirement 2 *The overlap of shipments amongst the sub-networks should be minimal.*

While it is not possible to avoid all the overlaps of origin – destination pairs that are assigned to the same hub, it is desirable to maintain the minimum amount of intersection of the shipments amongst the various sub-networks.

4.3 Mathematical Formulation

To decompose the network we propose a set covering problem. Each hub pair consists of two distinct hubs, that is, hub pairs of the form (h, h) are not permitted.

Notation

a_{ij} : is a constant. $a_{ij} = 1$ if hub pair[1] i is covered by sub-network j, 0 otherwise.

J : set of all sub-networks.

P : set of all hub pairs and $|P| = \binom{N}{2}$.

y_j : is a binary decision variable. $y_j = 1$ if sub-network j is used, 0 otherwise.

Mathematically, the set-covering problem is formulated as:

$$\text{Min} \quad z = \sum_{i \in P} \sum_{j \in J} a_{ij} y_j \tag{50}$$

$$\sum_{j \in J} a_{ij} y_j \geq 1 \tag{51}$$

$$y_j \in \{0, 1\} \tag{52}$$

[1] If overlap of origin-destination pairs was considered then i is a terminal pair rather than a hub pair

The set J is the power set of all the hubs excluding the empty set. Constraint 51 ensures that a hub pair is covered by at least one of the sub-networks and the objective 50 ensures that the overlap of the hub pairs is minimized.

Since $\sum_{j \in J} a_{ij} y_j \geq 1$:

$$z = \sum_{i \in P} \sum_{j \in J} a_{ij} y_j \geq \sum_{i \in P} 1 = |P| = \binom{N}{2}$$

There are several ways in which the network can be decomposed into sub-networks that yield this lower bound. Clearly, one solution is where the entire network is selected.

4.4 Decomposition Techniques

One can restrict the size of the sub-networks by imposing limits on the maximum number of hubs in a sub-network. Let k be the maximum number of hubs permitted in a sub-network. If $k = 1$ we route shipments between terminals assigned to the same hub. 1-hub sub-networks do not cover shipments between terminals not assigned to the same hub and violate the first decomposition requirement. Hence, we require $k > 1$. For all sub-networks j which have more than k hubs we set $y_j = 0$. Clearly, this can be achieved by restricting the elements in the set J.

Consider a network with N hubs. We restrict the size of each sub-network to a maximum of k (≥ 2) hubs. Then we can classify the set of decomposition solutions into the following three categories:

1. *Minimal Decomposition*: The decomposition scheme yields as many sub-networks as possible with exactly k hubs which gives us the minimum total number of decomposed sub-networks.

2. *Maximal Decomposition*: The decomposition scheme yields only 2-hub sub-networks, independent of the value of k. This scheme gives us maximum number of decomposed sub-networks, namely $\binom{N}{2}$

3. *Hybrid Decomposition*: The decomposition scheme is a hybrid of the minimal and maximal decomposition techniques.

We shall explain these three classes of decomposition schemes with the help of an example. Consider a network with 9 hubs with the constraint that at most 3 hubs can be in any sub-network.

4.4.1 Minimal Decomposition

We try to use as many sub-networks with 3 hubs as possible. There are several possible decompositions possible in this technique. Figure 25 shows how the sub-networks are generated that contain

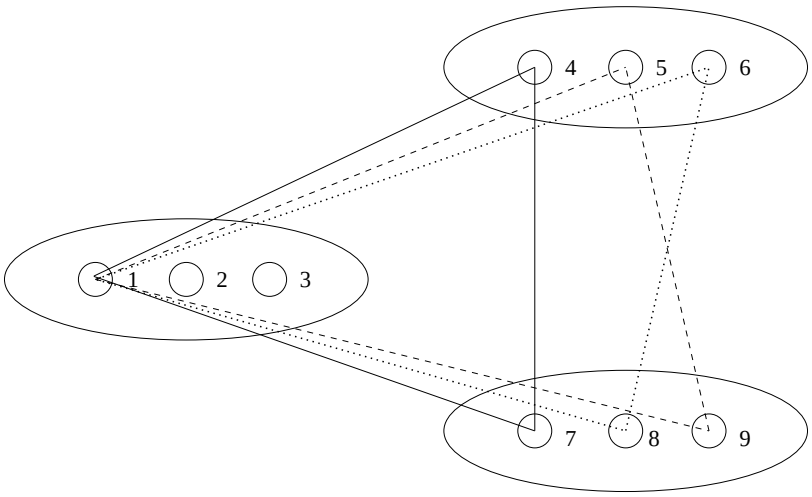


Figure 25: Minimal Decomposition Technique: All sub-networks covering hub 1

hub 1. Each of the triangles denotes a sub-network. This decomposition scheme yields the following sub-networks:

(1,2,3) (1,4,7) (2,4,8) (3,4,9)

(4,5,6) (1,5,8) (2,5,9) (3,5,7)

(7,8,9) (1,6,9) (2,6,7) (3,6,8)

This example is simple as the maximum number of hubs per sub-network (k) is equal to the maximum number of distinct sub-networks. That is, the total number of hubs in the network is k^2. In this case, we get $k(k+1)$ sub-networks.

4.4.2 Maximal Decomposition

A simpler technique is to generate sub-networks with two hubs, independent of the maximum number of hubs allowed in a sub-network.

For a network with N hubs this technique generates $\binom{N}{2}$ sub-networks (see figure 26). The

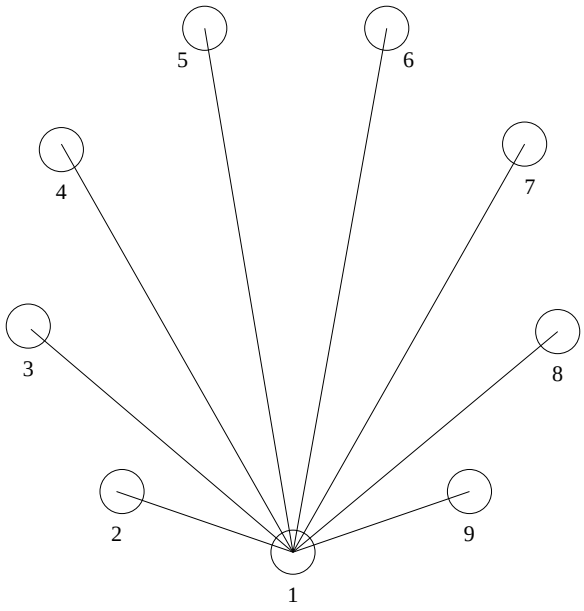


Figure 26: Maximal Decomposition Technique: All sub-networks covering hub 1

sub-networks are listed below:

(1,2)							
(1,3)	(2,3)						
(1,4)	(2,4)	(3,4)					
(1,5)	(2,5)	(3,5)	(4,5)				
(1,6)	(2,6)	(3,6)	(4,6)	(5,6)			
(1,7)	(2,7)	(3,7)	(4,7)	(5,7)	(6,7)		
(1,8)	(2,8)	(3,8)	(4,8)	(5,8)	(6,8)	(7,8)	
(1,9)	(2,9)	(3,9)	(4,9)	(5,9)	(6,9)	(7,9)	(8,9)

4.4.3 Hybrid Decomposition

In the case where the network does not have k^2 hubs, decomposing the network using the minimal decomposition technique is complex. One way is to divide the network into as many disjoint sub-networks with size k as possible. This leaves some shipments that are not covered which are then covered by hub-pairs. So for an N-hub network we get:

1. $s = \lfloor \frac{N}{k} \rfloor$ distinct (non-overlapping) sub-networks each with k hubs
2. 1 sub-network with $k' = N - ks$ hubs, if $N \bmod k \neq 0$

3. $k^2 \cdot \frac{(s-1)s}{2} + kk's$ sub-networks each with two hubs.

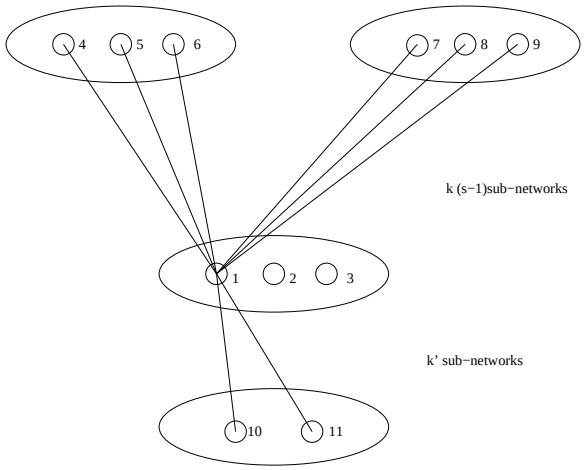


Figure 27: Hybrid Decomposition Technique: All sub-networks covering hub 1

Figure 27 shows the generation of such set of sub-networks. In this example the following sub-networks are generated:

(1,2,3)	(1,4)	(1,5)	(2,4)	(2,5)	(3,4)	(3,5)	(4,7)	(5,7)	(6,7)
(4,5,6)	(1,6)	(1,7)	(2,6)	(2,7)	(3,6)	(3,7)	(4,8)	(5,8)	(6,8)
(7,8,9)	(1,8)	(1,9)	(2,8)	(2,9)	(3,8)	(3,9)	(4,9)	(5,9)	(6,9)

As the name suggests, this decomposition technique is a hybrid of the minimal and maximal decomposition techniques.

In fact, shipments that are not overlapping are covered by exactly 1 sub-network. Recall that non-overlapping shipments cannot be eliminated if we decompose the network.

4.5 Selection of Decomposition Technique

We shall focus on analyzing the minimal and maximal network decomposition techniques. The hybrid decomposition technique retains some of the advantages of each and gets rid of some of the disadvantages of each. We can make the following observations:

1. Minimal decomposition yields the least number of sub-networks. The reduced number of sub-networks comes at the expense of the degree of difficulty in generating them. The combinatorial aspect and high level of communication between terminals makes it difficult to decompose the network, especially in the case where the network does not have exactly k^2 hubs.

2. Maximal decomposition provides an extremely easy algorithm to decompose the network. However, the number of sub-networks generated increases.

If the minimal decomposition technique is implemented, there are different ways in which k (>2) hubs can be grouped together to form a sub-network. Clearly, the difficulty associated with routing the shipments within each sub-network depends on the hubs and the terminals assigned to those hubs within that sub-network. Since the terminals are assigned *á priori* the number of hubs in a sub-network decides the size of a sub-network. Numerical experiments presented in the previous chapter show that sub-networks with fewer hubs tend to be more tractable computationally. Moreover, the difficulty also depends on the intensity of shipment flows and the distances involved. Clearly, it is not trivial to determine which is the best combination.

To overcome this difficulty we select $k = 2$. With this choice of k, minimal decomposition technique is equivalent to the maximal decomposition technique. Since each sub-network contains exactly two hubs this yields a unique decomposition of the network into $\binom{N}{2}$ sub-networks. Also for each sub-network this is the smallest number of hubs required for the model to remain valid. Computationally, 2-hub sub-networks should be the easiest to solve since adding more hubs will add more terminals with LTL shipments thereby increasing the difficulty in solving the problems. Furthermore, our computational experience shows that solving some of the 2-hub sub-networks is already a challenge and solving them to optimality can be challenging.

We select the minimal decomposition technique with $k = 2$ (or equivalently, the maximal decomposition technique) for the following reasons:

1. The decomposition technique is simple.
2. It yields a unique decomposition of the network.
3. From our computational experience (section 3) we know that the MIPs for the sub-networks, though still difficult, are the most tractable computationally.

4.6 Overlapping Origin – Destination Terminal Pairs in Sub-Networks

The ease of solving the routing sub-problems comes at the expense of larger number of sub-networks to be solved. Furthermore, there is one additional hurdle to overcome. In our formulation, we are interested in generating sub-networks which minimize the overlap between origin and destination terminals that are assigned to different hubs. However, if this overlap amongst terminals assigned to same hubs is also minimized, then the decomposition scheme will also try to minimize the total number of sub-networks that are generated for each hub since a greater number of sub-networks covering a hub means a greater overlap for the terminals assigned to that hub. Since every pair of

hub has to be covered, the decomposition scheme will try to find sub-networks in which hub-pairs are covered as few times as possible. This limits the decomposition scheme to minimal decomposition. However, some hubs have more terminals assigned to them than others. These hubs will be covered in as few sub-networks as possible. Minimal decomposition with $k > 2$ provides the least overlap when origin and destination terminals assigned to same hub are considered. However, maximal decomposition yields the most overlap.

Example 4.3 *For the 9-hub network considered previously, terminals assigned to hub 1 are covered in 4 sub-networks when minimal decomposition is used but are covered in 8 sub-networks when maximal decomposition is used.*

In the maximal decomposition scheme which we chose, shipments whose origin and destination terminals are assigned to different hubs are covered in exactly one sub-network. However, the shipments whose origin and destination terminals are assigned to same hubs are covered in $(N-1)$ sub-networks.

To maintain consistency in routing of overlapping shipments, one approach is to solve one sub-network and then fix the route of the overlapping shipments in all other sub-networks which are not yet solved and repeat the process until all the sub-networks are solved. This means that until the first sub-network has been solved the constraints for the second sub-network cannot be written. This approach requires all of the problems to be solved consecutively.

Recollect that for a origin (t_1) and destination terminal (t_2) assigned to hub h a shipment can be routed in one or more of the following 3 ways (see figure 22):

1. Directly from t_1 to t_2 (x^1)
2. From t_1 to t_2 via hub h without sorting (x^3)
3. From t_1 to t_2 via hub h with sorting at h (x^6)

Since most of the shipments (99,538 of 99,541) are less-than-trailerload shipments, it is likely that most of the shipments between terminals that are assigned to the same hub will not be split (sent by more than one route) in any of the sub-networks.[2] In fact, over 99% of the shipments occupy less than 10% of the truck capacity. It is reasonable to assume that these shipments will be sorted at the hub common to the origin and destination terminal. In a theoretical sense, it is quite possible that a shipment consisting of only a few packages may be sent as direct trailer from terminal t_1 to t_2. Consider the following example:

[2]It is possible that a shipment between terminals not assigned to the same hub may be split. (see example(cite relevant example here)).

Example 4.4 *Suppose terminal t_1 sends 40 packages to terminal t_2. At a cost of \$0.25/package, sorting cost for these packages is estimated at \$10. However, if these two terminals are located less than 10 miles apart, at a transportation cost of \$1/mile/truck, it is cheaper to send the 40 packages directly from t_1 to t_2.*

This example not only presents a scenario where sending an almost empty truck directly maybe be optimal but also suggests why freight routing is an extremely difficult problem to tackle.

4.7 Assumptions

In current practice, routings as suggested in example 4.4 are not practical because load planners are typically appraised on the average number of packages they put on trailers [Braklow et al., 1992]. Since we are trying to redesign the network we do not enforce any existing restrictions on the size of direct loads but we make the following assumption.

Assumption 1 *All shipments between terminals assigned to the same hub will be routed identically in each of the sub-networks.*

This assumption may not be valid for the shipments that are more-than-trailerload. In the previous assumption, a shipment between terminals assigned to the same hub will be sent by exactly one of the three routes. Based on the shipment size statistics, another approach is to avoid the overlapping altogether. Since over 99% of the shipments are less than 10% of the truck capacity it would be reasonable to assume that no shipment will be sent as a trailerload or truckload.

To synchronize the shipment routings in all sub-networks it can also be assumed that less-than-trailerload shipments between terminals assigned to the same hub will be sorted at the common hub in each of the sub-networks. Disallowing truckloads and trailerloads essentially implies that all the shipments will be sorted at the hub. Hence, besides tackling the problem of non-unique shipment routes in different sub-networks, it also simplifies the problem and may improve the performance of branch-and-bound algorithm by providing tighter lower bounds. However, since this is a very strong assumption, instead of enforcing this assumption *á priori* in the model we enforce these constraints *á posteriori*. When we superimpose the solutions from all the sub-networks to generate a global feasible solution for the whole network, we randomly select one of the shipment routings in case of overlapping origin – destination pairs. There is no way to measure the effect of this selection on the solution of different sub-networks without re-solving the routing sub-problems.

4.8 Routing Shipments in the Sub-Networks

Decomposing the network results in $\binom{N}{2}$ sub-networks. To route shipments through each of sub-networks we need to solve the MIP model associated with each sub-network. For the FedEx Ground data set we need to solve 276 MIP instances. Mixed integer programs for shipment routing are NP-hard problems, there are no efficient algorithms to route shipments through the network.

Two-hub sub-networks are the smallest size networks for which all the proposed 7 shipment routes (figure 22) are valid. Routing shipments through 2-hub sub-networks may seem to be easy. However, some of these sub-networks are extremely difficult. In some cases a good solution can be found early on in the branch-and-bound tree but the algorithm spends substantial amount time to close the gap. We use relative gap (see page 70) to measure the quality of the integer solution.

Because of the extremely loose lower bounds that can be generated for the routing of less-than-trailerload shipments, shipment routing problems are notoriously difficult to solve in practice [Leung et al., 1990]. To account for this we stopped the optimization process when an integer solution was found that was provably within 10% over the optimal solution. Based on conversation with experts in the industry a 10% bound was considered to be a acceptable bound.[3] We limit the solution time for each of the sub-networks to 6 hours, which we thought was substantial time for sub-networks of this size to obtain solutions of desired quality. For sub-networks for which no feasible solution could be found in 6 hours optimization continued until an integer feasible solution was found. Figure 28 shows histograms for the number of sub-networks with the following:

1. Solution time to solve each of the associated MIPs to within 10% optimality.

2. Optimality gap achieved when optimization terminated.

Within 6 hours of allocated time, 47 of the 276 sub-problems (17%) could not be solved within 10% of optimality and no integer solution could be found for 10 of these 47 sub-networks. The average solution time per sub-problem was 5,962 seconds (1.65 hours) and the total processing time to solve all the 276 instances was 1, 645, 530 seconds (457 hours).

In figure 29 we show a correlation between the number of variables and constraints in the MIP associated with a sub-network and its associated solution time. As expected, the solution time increases as the variables and constraints increase.

Since the number of variables and constraints in a network depends on the number of terminals in the 2-hub sub-network there is also a correlation between solution times and the number of

[3]Extrapolated from the conversation with Sev Murtry when he mentioned that they were not interested in the optimality gap.

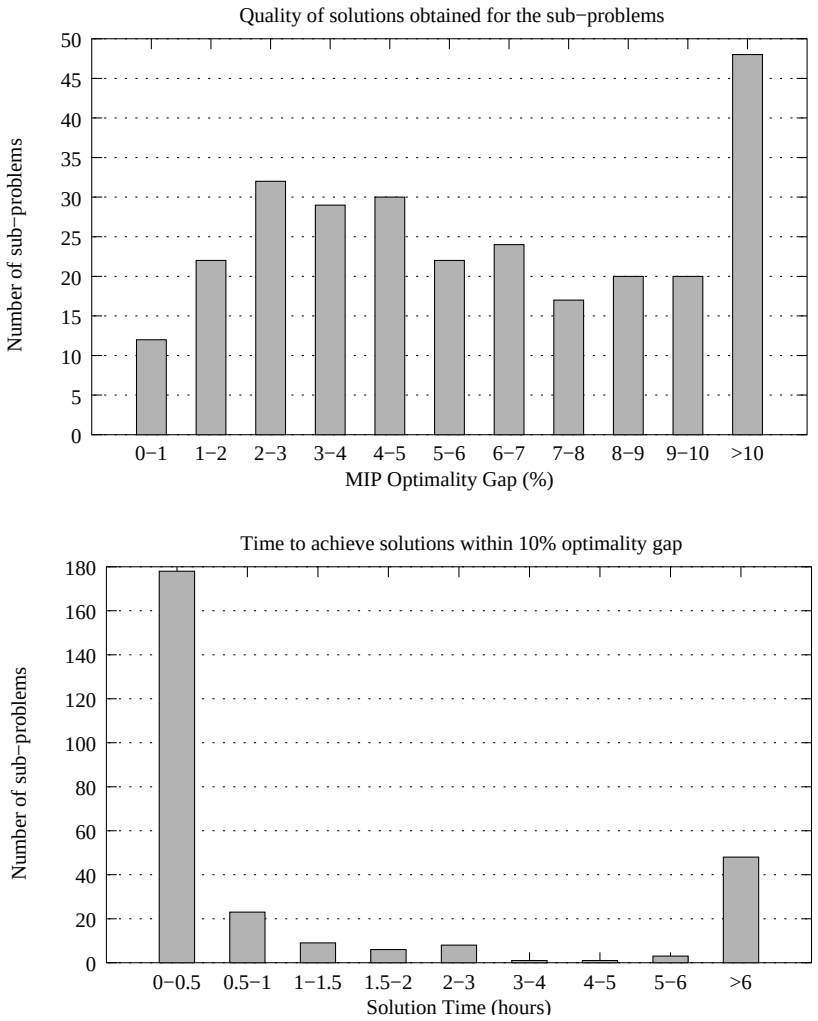


Figure 28: Histogram of the solution times and optimality gaps for a problem decomposed into 276 sub-problems.

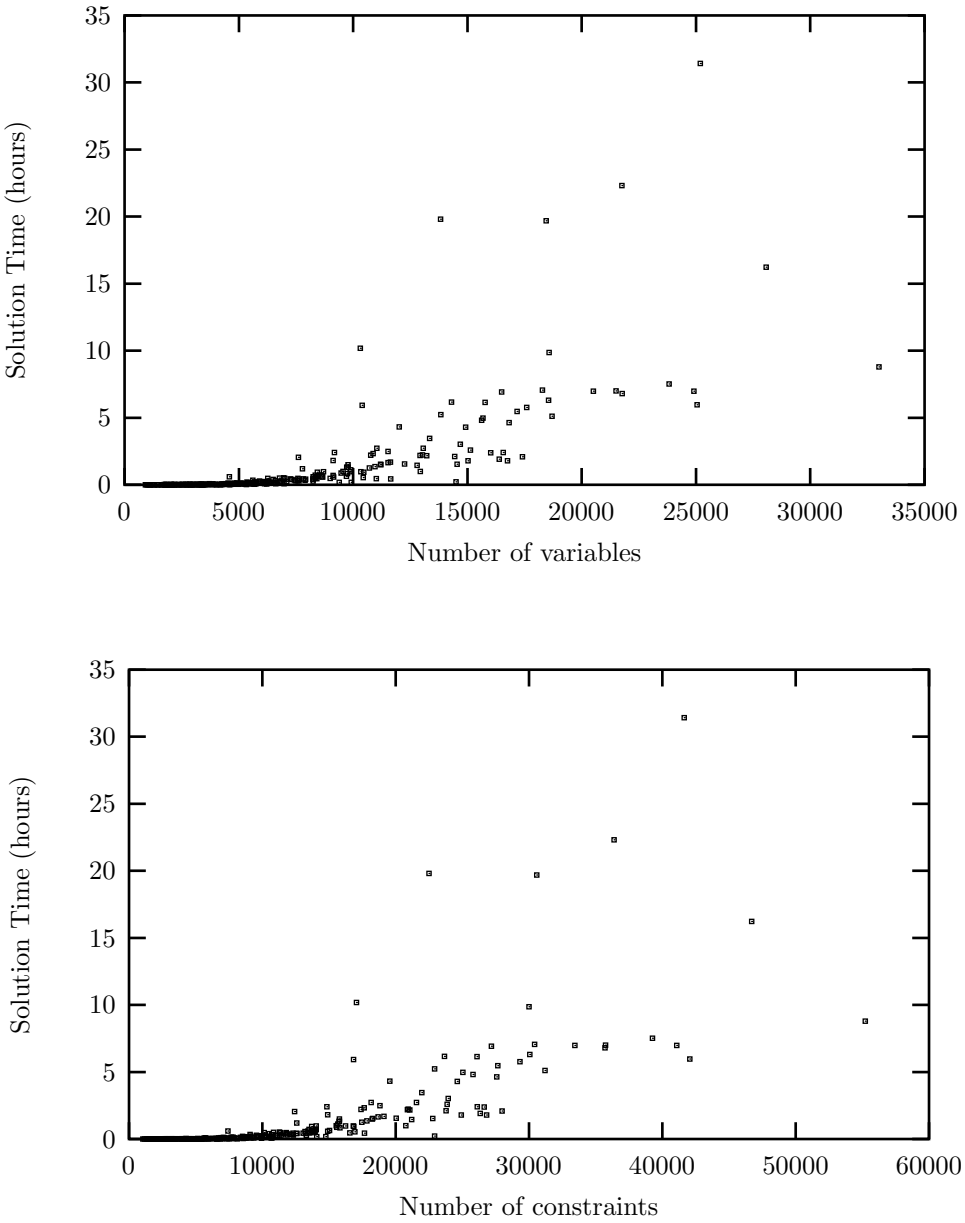


Figure 29: Solution time versus number of variables and constraints

terminals in the network. As expected, sub-networks with more terminals are difficult to solve than sub-networks with fewer terminals (figure 30).

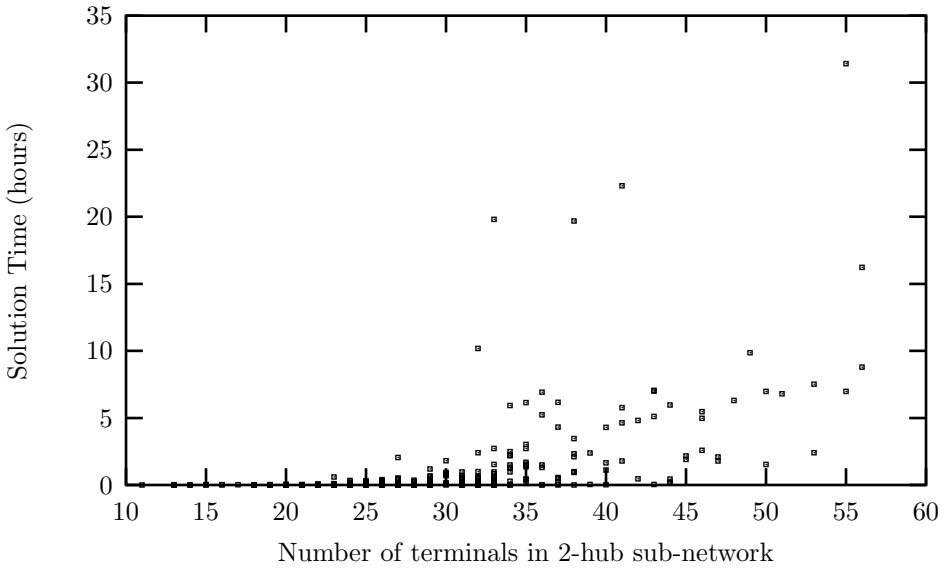


Figure 30: Solution time versus number of terminals in a 2-hub sub-network

The characteristics of the shipment routings are discussed in section 5.1.

4.9 Parallelization

As explained in previous section, without assuming that all the shipments will be uniquely routed in all the sub-networks, one way to synchronize shipments within all the sub-networks is to solve one sub-network and fix the routings of the shipments overlapping in other sub-networks. This approach is similar to a local search algorithm and hence not (provably) optimal. This approach forces all of the problems to be solved sequentially. For the FedEx data set solving 276 sub-networks (to 10% optimality) sequentially takes about 2 weeks.[4]

However, with the assumption that the overlapping shipments are uniquely routed in all the sub-networks, solving the sub-networks individually as disjoint sub-networks yields the same results as solving these sequentially. Hence, the synchronization has proved to be unnecessary. Under this assumption, the decomposition technique for the original network classifies as an *embarrassingly parallel* decomposition technique which is extremely suitable for parallelization. An embarrassingly parallel decomposition technique is one in which there is no knowledge-passing between any two

[4]This time may be reduced because some route variables that are fixed from the previously solved sub-network are no longer decision variables in other unsolved sub-networks.

sub-networks [Fox, Williams, and Messina, 1994].

The following analogy [Chalmers and Tidmus, 1996] clearly explains the need for parallel processing for our shipment routing problem.

The relevance of a 24 hour weather forecast may
be questioned if it requires 36 hours to calculate.

The main goals of parallel processing are:

1. Reduce "wall-clock" time and
2. To obtain good solutions for large-scale instances for daily use.

We use parallel processing to reduce wall-clock time to solve all of the $\binom{N}{2}$ sub-networks. A problem can be solved on a parallel cluster in either of the two ways:

Algorithmic decomposition The algorithm itself is analyzed to identify which of its features are capable of being solved in parallel. This is also known as *functional parallelism.*

Domain decomposition Instead of determining the parallelism inherent in the algorithm, domain decomposition, also known as *data parallelism*, examines the problem domain to ascertain the parallelism that may be exploited by solving the algorithm on distinct data items in parallel.

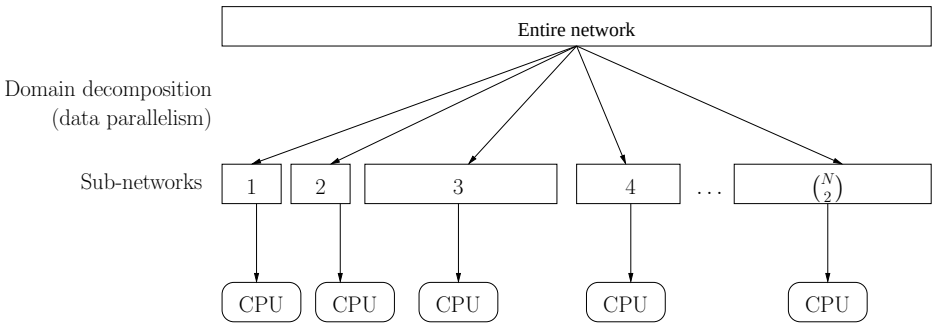


Figure 31: Domain decomposition for routing shipments in the network

The effectiveness of the decomposition depends on *granularity* of the parallelism. Domain decomposition is most effective when the entire problem is divided into sub-problems that require approximately the same amount of work so that all the processors are almost identically loaded. Load balancing is an important aspect for significant speed-ups, and is discussed in section 4.10.4.

The proposed network decomposition technique is categorized as domain decomposition (figure 31). The network is decomposed into sub-networks each of which is then solved on an individual

processor. Each processor uses the branch-and-bound technique (discussed in Chapter 3) to determine shipment routes within each sub-network. Though there is no data dependency because of assumption 1 there needs to be some coordination among processors while accessing the sub-networks so as to avoid collision (repetition).

4.9.1 Message Passing

One approach to achieve parallelism is by *message passing*. A message passing function is simply a function that explicitly transmits data from one process to another. A *library* of such standardized functions is specified. *M*essage *P*assing *I*nterface (MPI) is a standard specification for message passing libraries. Message passing is a powerful and very general method for expressing parallelism, and is currently one of the most widely used method of programing for many types of parallel computers. The principal drawback of message passing is that it is very difficult to design and develop programs using message passing. However, these sophisticated algorithms are encapsulated in portable MPI libraries which can then be called as functions into a program [Snir, Otto, Huss-Lederman, Walker, and Dongarra, 1996].

MPICH[5] is a freely available, portable implementation of the full MPI specification for a wide variety of parallel computing environments, including workstation clusters and massively parallel processors. MPICH contains, along with the MPI library itself, a programing environment for working with MPI programs. MPICH is the parallel implementation platform used in this research.

4.9.2 Round-Robin Scheme

A common approach to schedule jobs on parallel processors is the use of Round-Robin scheme. Let $1,\ldots,n$ be the sub-networks decomposed and let the number of processors be κ. Then processor j $(0 \leq j < \kappa)$ will solve the instance associated with sub-network i $(1 \leq i < n)$ if $(i \bmod k) = j$.

This scheme is decentralized and extremely easy to implement. Moreover, it generates (near-)optimal load balance schemes if it takes approximately the same time to route shipments within the sub-networks. However, a serious drawback of this scheme arises when this is not the case. It is possible that one processor gets significantly more difficult instances than others. All other processors will idle until this processor has routed shipments through all of its assigned sub-networks. This may lead to very poor speed-ups.

Though all of our sub-networks have 2 hubs, for the FedEx data set the number of terminals in

[5]The "CH" in MPICH stands for "Chameleon," symbol of adaptability to one's environment and thus of portability. Chameleons are fast, and from the beginning a secondary goal was to give up as little efficiency as possible for the portability.

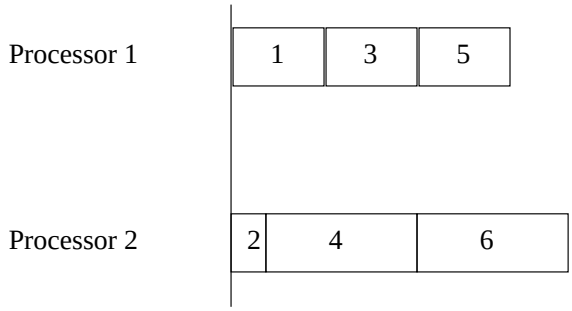


Figure 32: Round-robin scheme for 2 processors: Processor 1 solves all the even sub-networks and processor 2 solves all the odd sub-networks.

the sub-networks vary from 11 to 57 and it may be significantly difficult to route shipments in some sub-networks than others. Hence the Round-Robin scheme is not suitable in our case.

The maximum speed-up can be obtained theoretically when each routing sub-problem has equal solution time and all processors are identically loaded. Theoretically, the maximum speed-up possible for κ processors is κ.

4.9.3 Master-Slave Paradigm

In order to distribute jobs evenly the Master-Slave paradigm is commonly used. In this case, load balancing is centralized. There is a single processor, known as the *master processor* which is in control of all processors, and all other processors are collectively called as *slave processors*. The master processor keeps a list of unsolved tasks and hands over an unsolved task to an idle slave processor. In this case, the task is removed from the list of unsolved tasks. The tasks may be indexed and queued by a certain rule under consideration. For example, if we try to solve difficult MIP instances of the routing sub-networks first then the tasks (MIPs associated with the sub-networks) may be sorted and indexed by the decreasing number of variables or constraints. Figure 33 illustrates a flow diagram of this approach.

In an absence of predefined tasks as in the case of Round-Robin scheme, the master processor keeps a list of unsolved tasks and hands them to the slave processors to ensure no repetition of work. However, the centralization in task management comes at the expense of the master processor idling for most time while the slave processors solves the routing sub-problem in each sub-network.

Again, the maximum speed-up can be obtained theoretically when each routing sub-problem has equal solution time. However, since the master processor does not contribute to any processing, the maximum speed-up that can be obtained from κ processors is $\kappa - 1$.

In spite of the centralization, there is no guarantee to improve the load balance especially if there

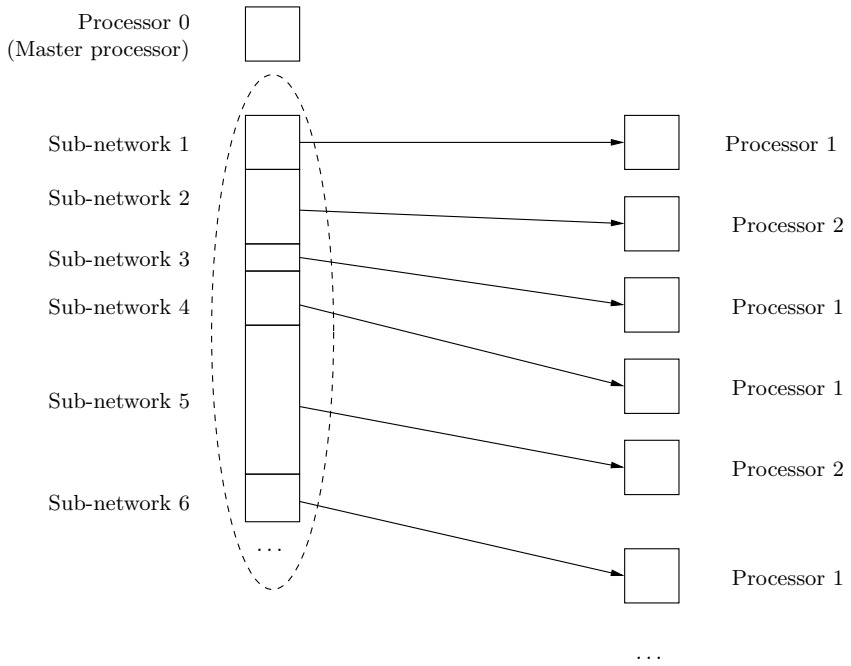


Figure 33: A master-slave scheme to solve sub-networks on 3 processors. The centralization in task management comes at the expense of the master processor idling for most time while the slave processors route shipments within the sub-network

is high variation in the time to solve the routing sub-problem for each sub-network.

4.9.4 Co-operative Decentralized Paradigm

In the Master-Slave paradigm, the idling of a single processor may seem insignificant especially when the total number of processors employed is high. However, when very few processors are employed, dedicating one entirely for centralization implies a significant under-utilization of available resources. In order to hedge our parallelization approach against the total number of processors employed we implemented a co-operative decentralized scheme.

Figure 34 illustrates idea of solving the routing sub-problems in a decentralized parallel computational environment [Lee, 1999, 2004]. Unlike the Round-Robin scheme there is no prior division of task among the processors and unlike the Master-Slave algorithm there is no processor which keeps a list of unsolved tasks and hands them to the idle processors.

This means this parallel paradigm is extremely susceptible to collision (repetition of tasks) unless there is a mechanism for co-operation among all the processors so that one processor does not grab a MIP, associated with a sub-network to route shipment, which is currently being solved by another processor.

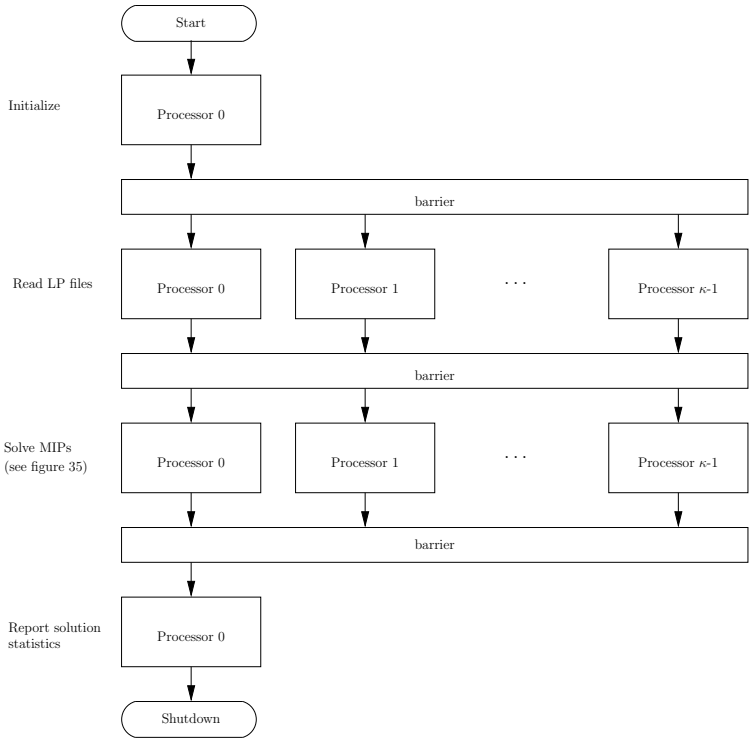


Figure 34: A higher-level flowchart for routing shipments in $\binom{N}{2}$ sub-networks using κ processors in a decentralized co-operative parallel computational environment.

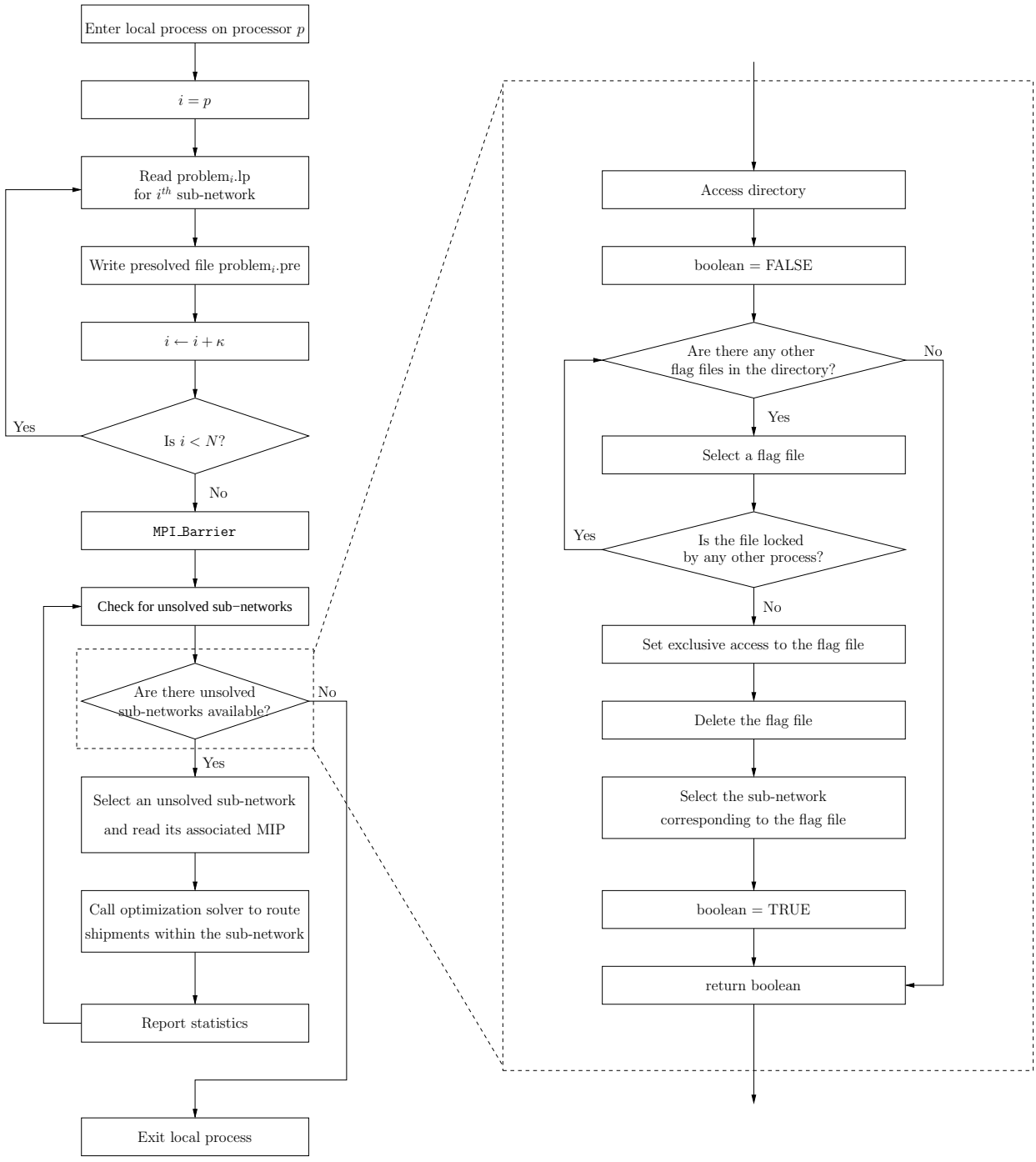


Figure 35: A detailed flowchart for each processor to read in a MIP for each sub-network and solve it

Once the parallel processes have been initialized, the root processor initializes and cleans up the necessary file structure so that the new log files can be written to gather and report solution statistics. All other processors wait until the root processor has finished initialization. This is accomplished using a MPI function, `MPI_Barrier` [Snir et al., 1996]. We use the same function for the processors to wait until all of them have solved the sub-networks.

After the barrier, all processors grab the MIP files associated with the sub-networks and solve the MIPs to route shipments within each of the sub-networks. To make sure that the MIP files are not erroneous, we check the MIP files by reading and presolving them. Since the time taken to read the MIP files is typically under 5 seconds, we use the Round-Robin scheme. However, since the solution times have high variability among different instances, we use the co-operative decentralized paradigm to solve the MIPs. Figure 35 represents the flow of tasks within each local processor to read the MIP files and grab sub-networks to route shipments.

To ensure that as few problems are solved repetitively as possible we use *file locking* as shown in figure 35. When a processor grabs a sub-network to route shipments it sets an exclusive lock on it so that no other processor can grab the same sub-network. In section 4.10.6 we discuss the robustness of our implementation of file locking mechanism.

4.10 Computational Results

4.10.1 Hardware and Software

All the computations were performed on a cluster of 17 eight-processor servers. Each processor is a 550MHz Pentium III Xeon with 4MB RAM and 18GB SCSI disk. The networking medium was dual Gigabit Ethernet, with etherchannel aggragation 1.2 Gb myrinet cards using OS Redhat Linux 7.1.

4.10.2 No Limitations on Size of Direct Load

In section 4.8 we presented the results for the instance where we do not enforce any direct load constraints. The instances are extremely hard and we could not solve 47 sub-networks to 10% optimality within 6 hours. For 10 of the 47 sub-networks no integer solution could be found in 6 hours. We report here the statistics for 266 MIP instances associated with the routing sup-problems for which a feasible solution could be found within 6 hours. On 80 processors we obtained a speed-up[6] of about 49.6. The speed-up is really lucrative since solving these instances sequentially would

[6]The speed-up can be calculated as the ratio T_1/T_p where, T_κ = elapsed time to solve $\binom{N}{2}$ sub-problems on κ processors.

have taken about 397 hours (about 16.5 days). However, we can now solve all the instances within 8 hours (less than half a day). In table 11 we see that the reduction in wall-clock time is rather significant.

Number of processors	80
Parallel processing time	28,824 sec
Sequential processing time	1,429,530 sec
Speed-up	49.59

Table 11: Speed-up statistics for the case where there are no restrictions on minimum size of direct loads

Since we do not take into account the extremely difficult 10 instances for which no feasible solution could be found in 6 hours, the "pseudo" speed-up reported here is overly optimistic. If the 10 instances were allowed to be solved until their first feasible solution, these 10 processors would be utilized while the remaining 70 are idling, resulting in a very poor speed-up. In this case, algorithmic decomposition would be extremely useful to improve the speed-up.

Solving the instances where direct loads of any size can be sent is extremely difficult and has been shown to be computationally intractable. As explained in section 5.4.1 solving this case is not useful to get good shipment-routing solution for the entire network. As will be seen in the rest of this chapter and the next chapter, enforcing direct load size not only makes the problems easier and quicker to solve but also yields reasonable good shipment-routing solution for the entire network. Hence, we focused on direct load size restrictions in steps of one-tenth of a trailer capacity, that is, steps of 100 packages. However, to maintain clarity in the discussion and graphs we only report the results for direct load size in multiples of 200 packages.

4.10.3 Elapsed Computational Time

Consider the constraints which restrict direct loaded trucks. From equation 7 we have

$$f_{ij} = x^1_{ij} + \sum_{k:k=h(i),h(j)} (x^3_{ijk} + x^6_{ijk}) + \sum_{k:k=h(i)} \sum_{l:l=h(j),l\neq k} (x^2_{ijkl} + x^4_{ijkl} + x^5_{ijkl} + x^7_{ijkl})$$

which implies,

$$x^1_{ij} \leq f_{ij}$$

From equation 42 we have

$$Dt^1_{ij} \leq x^1_{ij} \leq Ct^1_{ij}$$

which implies,

$$Dt_{ij}^1 \leq x_{ij}^1 \leq f_{ij}$$

Since number of truck cannot be negative,

$$D > f_{ij} \Longrightarrow x_{ij}^1 = 0 \tag{53}$$

Extending this argument to the remaining directly loaded trucks we get:

$$D > f_{ij} \Longrightarrow x_{ijkl}^2 = 0 \tag{54}$$

$$D > f_{ij} \Longrightarrow x_{ijk}^3 = 0 \tag{55}$$

$$D > f_{ij} \Longrightarrow x_{ijl}^3 = 0 \tag{56}$$

Since 90% of the shipments are under 20 packages, for any value of $D > 20$ bounds implied by equations 53–56 are valid for all of these shipments. For K shipments which consist of less than D packages, depending on the network configuration at least $2K$ and at most $4K$ variables will be fixed by these implied bounds.

$$D > \sum_{j:hub(j)=l} f_{ij} \Longrightarrow x_{ijkl}^4 = 0 \qquad \forall j : hub(j) = l \tag{57}$$

$$D > \sum_{j:hub(j)=l} f_{ij} \Longrightarrow x_{ijl}^6 = 0 \qquad \forall j : hub(j) = l \tag{58}$$

$$D > \sum_{i:hub(i)=k} f_{ij} \Longrightarrow x_{ijkl}^5 = 0 \qquad \forall i : hub(i) = k \tag{59}$$

$$D > \sum_{i:hub(i)=k} f_{ij} \Longrightarrow x_{ijk}^6 = 0 \qquad \forall i : hub(i) = k \tag{60}$$

It may be the case that the maximum number of packages that can be sent from a terminal to a hub (or from an hub to a terminal) directly may be less than the minimum requirement of D. This further implies bounds for some of the route variables as shown in equations 57–60.

As D increases, since the number of variables fixed to 0 increase, more variables are eliminated and the instances become easier to solve as shown in figure 36

Clearly, with no restriction on the size of direct load, none of these route variables are eliminated by pre-processing. This explains why instances with no restrictions on the size of direct load are extremely hard. By restricting trailer utilizations to 10% (direct load size of 200 packages) a substantial amount of route alternatives are eliminated resulting is drastically reduced solution times.

		Direct Load Size				
		200	400	600	800	1000
Number of Processors	1					
	8	832	415	196	134	86
	16	730	288	148	94	53
	24	631	233	99	84	35
	32	580	241	86	72	31
	40	541	211	84	73	25
	48	553	219	80	70	20
	56	525	213	76	68	21
	64	524	209	75	66	24
	72	534	223	81	66	21
	80	527	202	84	88	23

Solution time (seconds)

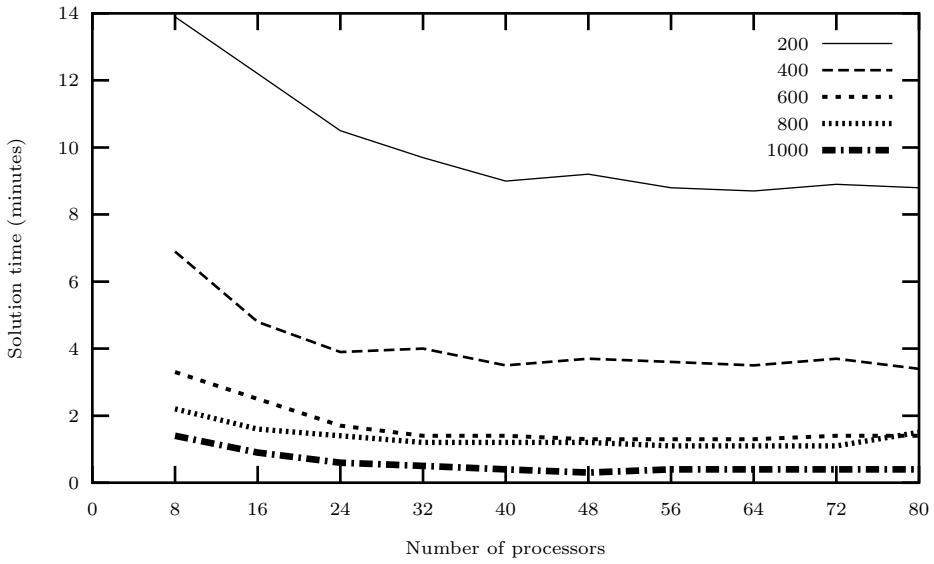


Figure 36: As we increase the direct load size requirements the MIP instances become easier to solve, thus decreasing the total wall-clock time.

		Direct Load Size				
		200	400	600	800	1000
Number of Processors	1	1.000	1.000	1.000	1.000	1.000
	8	0.791	0.655	0.742	0.744	0.898
	16	0.494	0.473	0.486	0.491	0.696
	24	0.353	0.382	0.481	0.370	0.710
	32	0.278	0.336	0.399	0.322	0.620
	40	0.244	0.258	0.345	0.270	0.611
	48	0.202	0.199	0.302	0.229	0.629
	56	0.173	0.177	0.271	0.199	0.532
	64	0.153	0.171	0.249	0.184	0.548
	72	0.137	0.181	0.219	0.165	0.483
	80	0.143	0.156	0.192	0.187	0.426

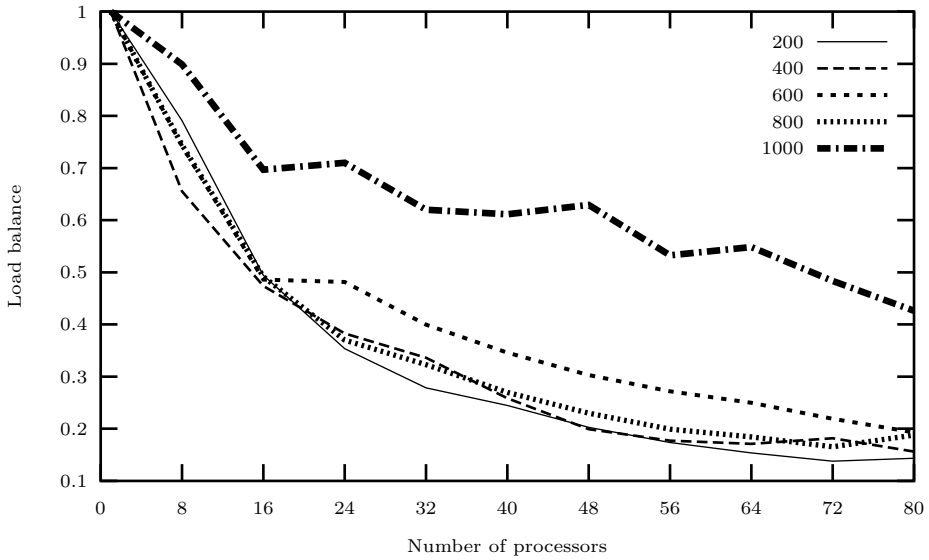


Figure 37: As we increase the direct load size requirements, the variability in times to solve the MIP instances decreases and the load balance on the processors improves.

4.10.4 Load Balancing

Figure 37 shows the efficiency (load-balance) achieved by using parallel processing. For a single parallel run, the efficiency is calculated as the ratio of the shortest active CPU time spent by a processor to the longest active CPU time spent by another processor.

We can make the following two observations:

1. Efficiency almost monotonically decreases with increase in the number of processors.
2. For the same number of processors, the efficiency typically decreases with the size of direct load.

The low efficiency observed is due to the variability in solution times to solve to 10% optimality. For example, for the case where direct load sizes are restricted to 100 packages per trailer we have the following statistics:

Statistic	Value
Maximum	1,821 sec
Average	61 sec
Variance	32,788 sec^2
Standard deviation	181 sec
Coefficient of variation	3

The sub-network which takes the longest to solve dictates the efficiency of the parallel process. Independent of the number of processors used, this sub-network will take the same time to solve unless an algorithmic decomposition scheme is used. In fact, the more processors are used the more CPUs/processors will idle until this MIP associated with this sub-network is solved by one single CPU/processor.

Figure 38 shows the distribution of the routing sub-problems on processors using the co-operative decentralized scheme. An idle processor can grab any unsolved sub-network for processing which result in longer time to finish solving all the MIP instances thereby resulting in slightly lower efficiency. For example, in figure 38 the longest job is solved after three other short jobs are solved. Instead, the three short jobs could be solved on some other processors which are idling while the longest job is still being solved. One heuristic rule to achieve this is the *longest processing time first rule* where the available processor grabs the problem which has the longest processing time. This heuristic rule can be implemented if the processing times were known *a priori* (see figure 39). However, the solution time for solving MIPs can not be predicted (or determined) before hand.

Figure 38: Load balancing on the processors when an unsolved sub-network is randomly selected.

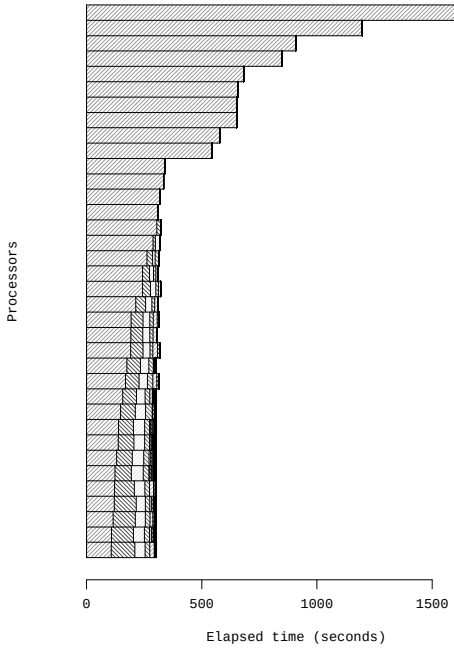


Figure 39: Schedule of job on the processors by *longest processing time first* rule if the solution times were known.

Since solution times are not known *a priori* to any rule which schedules the sub-networks on the processors has to be independent of the solution times. As explained earlier there is a correlation between the number of variables and constraints in the MIP associated with a sub-network and the time required to solve the instance (figure 29). Since the number of variables and constraints in a network depends on the number of terminals in the 2-hub sub-network there is also a correlation between solution times and the number of terminals in the network. As expected, sub-networks with more terminals are difficult to solve than sub-networks with fewer terminals (figure 30). The sub-networks may be sorted and ranked by decreasing number of terminals and solving the lowest ranked unsolved network is almost equivalent to the longest processing time first rule. However, in our experience the longest processing time first rule did not yield significant improvement in speed-ups.

In order to achieve significant speed-up, a more effective approach is to use algorithmic decomposition along with domain decomposition (see figure 40). Algorithmic decomposition for branch-and-bound optimization is a common approach [Gendron and Crainic, 1994]. We have not implemented algorithmic decomposition along with domain decomposition and is beyond the scope of current research.

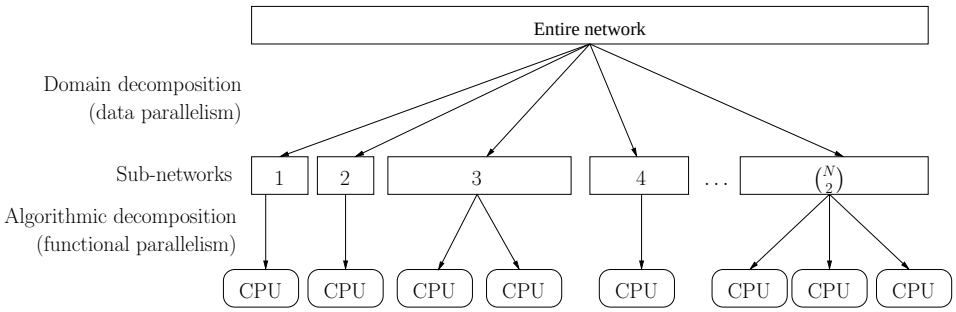


Figure 40: Algorithmic and domain decomposition for routing shipments in the network

4.10.5 Speed-ups

Amdahl's law [Dongarra, Duff, Sorensen, and van der Vorst, 1993] states that for an algorithm in which the proportion of time that needs to be spent on the purely sequential parts and parallel parts are s and p respectively, the maximum speed-up that can be obtained is:

$$\text{maximum speed-up} = \frac{(s+p)}{s+\frac{p}{n}}$$

The decomposition technique proposed is designed so that $s = 0$. So for our case, by Amdahl's law we can theoretically attain a maximum speed-up of n. However, the only way to achieve this

speed-up is if all the processors are perfectly load-balanced. This means that all of the instances should take the same time to solve. Figure 41 shows the speed-ups obtained for various direct load sizes. As we increasingly restrict the size of a direct load, the speed-up increases. As we increase the restriction on the size of the direct loads, a shipment has fewer options to be routed and the instances become easier to solve and as expected the variation in the solution time decreases.

		Direct Load Size				
		200	400	600	800	1000
Number of Processors	1	1.00	1.00	1.00	1.00	1.00
	8	6.32	5.24	5.93	5.95	7.18
	16	7.90	7.58	7.77	7.86	11.15
	24	8.48	9.19	11.55	8.88	17.04
	32	8.91	10.76	12.79	10.32	19.84
	40	9.79	10.36	13.83	10.80	24.45
	48	9.73	9.55	14.54	11.03	30.19
	56	9.70	9.93	15.22	11.15	29.82
	64	9.83	10.95	15.99	11.78	35.10
	72	9.92	13.09	15.77	11.92	34.80
	80	11.49	12.50	15.43	15.03	34.11

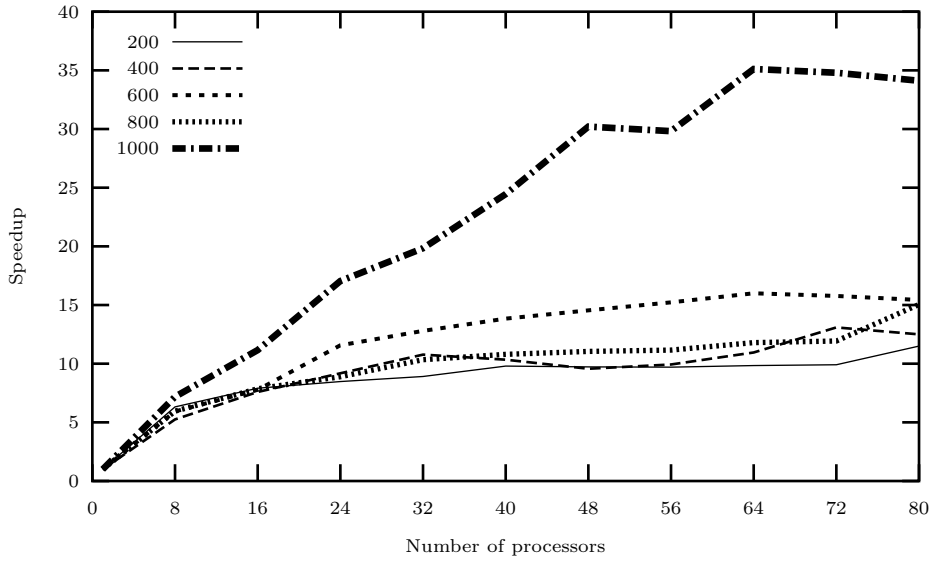


Figure 41: As we increase the direct load size requirements the instances become easier to solve and we achieve higher speed-ups.

The speed-up can be calculated as:

$$\text{speed-up} = \frac{T_1}{T_p}$$

where,

T_p = elapsed time to solve $\binom{N}{2}$ sub-problems on p processors and

T_1 = elapsed time to solve $\binom{N}{2}$ sub-problems on a single processor.

Since the decomposition technique is embarrassingly parallel we get

$$\text{speed-up} = \frac{T_{total}}{T_{max}} = \frac{\sum_{i=1}^{p} t_i}{max_i\{t_i\}}$$

where, t_i = total elapsed time to solve sub-networks on a processor p.

Though each sub-network consists of only 2 hubs and its assigned terminals, as discussed in section 4.5 the difficulty in solving the instance associated with the sub-network depends on the distances involved and the shipment size. Some sub-networks, especially those with greater number of terminals, are typically more difficult to solve than sub-networks with fewer terminals. As seen in figure 39, the sub-network that takes the longest to solve dominates the speed-up achieved. In the same figure, we can see that the time to solve all the sub-networks, and hence the speed-up, remains unchanged in spite of increasing the number of processors. As long as few routing sub-problems dominate the total processing time, increasing the number of processors does not improve speed-ups in a data decomposition parallelization scheme.

Our computational experience suggests that for direct load sizes under 800 packages per trailer the speed-up increases up to 24-32 processors beyond which the increase in speed-up is not substantial. However, for the case where only a completely utilized trailer can be sent directly, significant speed-ups were achieved up to 64 processors.

4.10.6 Parallel Collision

In spite of taking advantage of file locking as explained in section 4.9.4, some sub-networks are solved again. This is because just before a processor grabs a file and locks it, another processor may simultaneously grab the same file. Since the first processor has not yet completed locking the file, the second processor may grab the same file to route shipments in the associated sub-network.

For the FedEx Ground data, theoretically only 276 problems should have been solved irrespective of the number of processors used. Figure 42 shows the number of problems actually solved. The number of problems solved (and hence the number of collisions) increases with the number of processors. Moreover, as the direct load size increases the instances become easier and quicker to solve and for the same number of processors used there tend to be more collisions than for instances corresponding to smaller direct load sizes.

		Direct Load Size				
		200	400	600	800	1000
Number of Processors	1	276	276	276	276	276
	8	291	279	286	319	319
	16	311	288	293	293	289
	24	298	291	295	298	295
	32	296	294	299	303	304
	40	293	297	307	310	304
	48	301	298	306	309	304
	56	302	305	312	308	312
	64	304	320	321	319	322
	72	315	332	306	326	328
	80	326	336	331	340	338

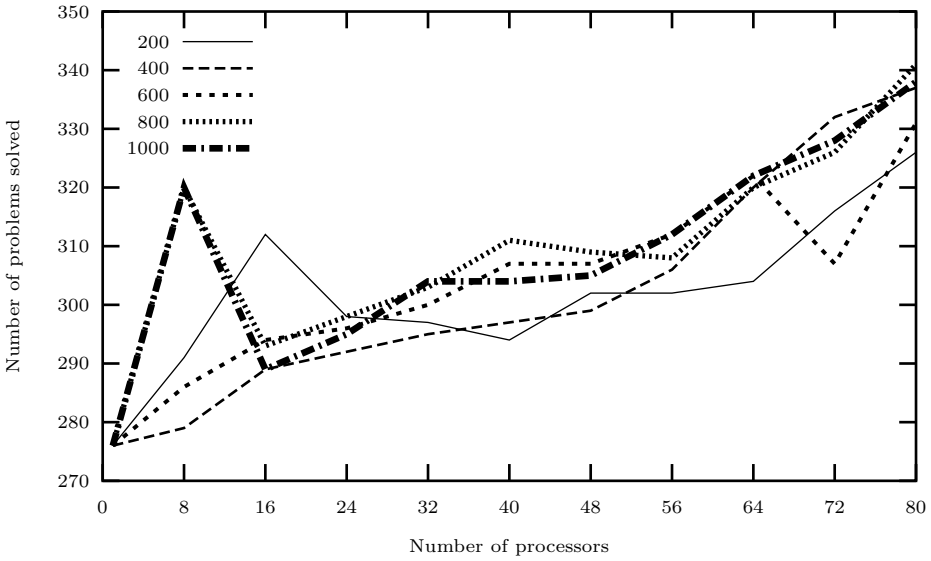


Figure 42: As we increase the direct load size requirements the instances become easier to solve and there is more collision in file locking.

CHAPTER 5

SHIPMENT ROUTING – NETWORK ANALYSIS

5.1 Analysis of Freight Routes for the Entire Network

Figure 43 shows how the shipments are routed within the hub-and-spoke network when there are no restrictions on the size of a direct load. For the given set of assignments for the FedEx data set, about 10.9% of the total packages are between terminals assigned to the same hub and approximately 89% of the packages are between terminals assigned to different hubs. Over 70% of the shipments between terminals assigned to the same hub are sorted at that hub.

Only 1.9% of the total number of packages are sent in direct trucks. Most of the trucks that are sent directly are poorly utilized. This may be partly due to the fact that some sub-problems could not be solved to within 10% of optimality. As seen in figure 28, within 6 hours of allocated time, 17 of the 276 sub-problems could not be solved within 10% of optimality.

Approximately 89% of the packages were routed and sorted through both the hubs in pure hub-and-spoke setting. After optimization, only about 35% are now routed and sorted through both the hubs. About 1.3% of the packages are shipped in direct trucks. About 14.5% of the shipments for which the origin and the destination terminals are not assigned to the same hub are loaded on to direct trailers bypassing a hub. Almost twice the number of packages are loaded directly from the terminal to the hub as from the hub to the terminal. By carefully loading the trailers about 34% of the shipments avoid sorting at one of the hubs though they are routed through both the hubs.

In figure 44 we provide a simplified analysis of the routing solution. The highlights of the solution are:

1. The most expensive routing for the packages, the double sort, reduced from about 89% to 34%.
2. About 35% of the packages are routed through both the hubs but sorted at only one hub or not at all.
3. About 22% of the packages were routed through and sorted at only a single hub. Of all the packages, 65% (14.5% of the total) were sent and received between terminals assigned to different hubs. The remaining 35% of these packages were sent and received between terminals assigned to the same hub.

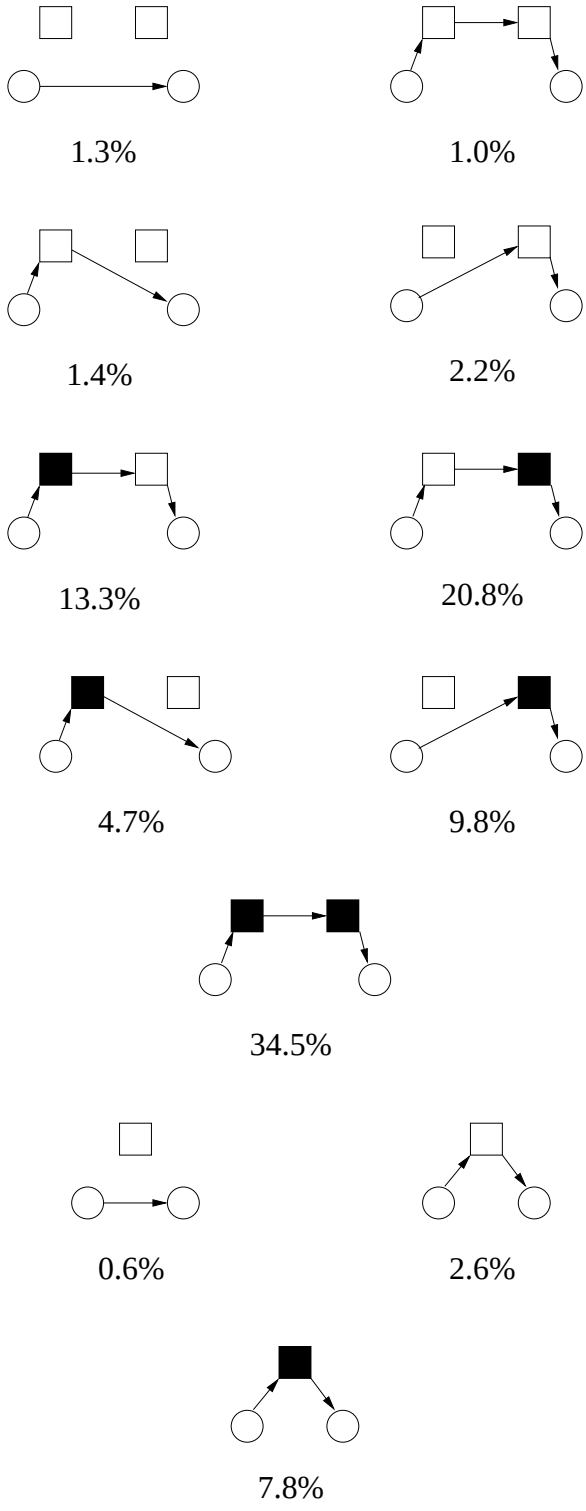


Figure 43: How the packages are routed within the hub-and-spoke network

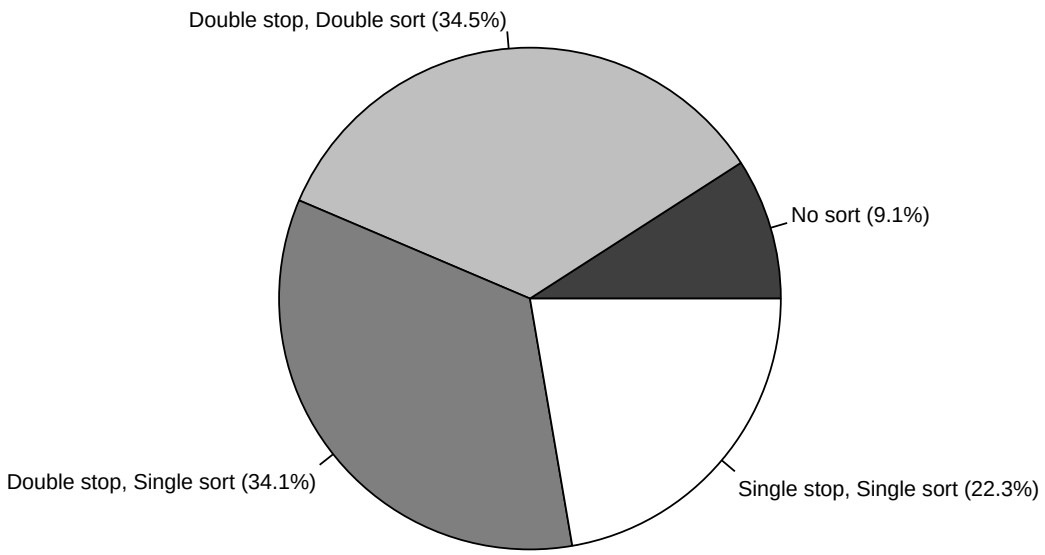


Figure 44: The most expensive (double sort) routing reduced to about 34% from about 89%

4. Approximately 9% of the packages were not sorted at either hub. Of these packages, 21% were sent directly, 68% were routed through a single hub and 11% were routed through both the hubs.

The prevalent notion in the industry is that there are ample opportunities to sort freight at a hub rather than a terminal [Braklow et al., 1992]. This fact is used in most heuristics which prefer to generate hub-to-terminal direct loads over terminal-to-hub direct loads. However, contrary to this industry trend, we found that more shipments are sent as terminal-to-hub direct loads than hub-to-terminal direct loads. About 10% of the packages are loaded directly from terminal to hub versus about 5% that are loaded directly from hub to terminal. For the packages that were routed through two hubs but were only sorted at one, over 50% more packages were consolidated at the second hub (20.8%) than at the first hub (13.3%).

This suggests relatively large producers of freight and relatively small consumers: perhaps manufacturers distributing products.

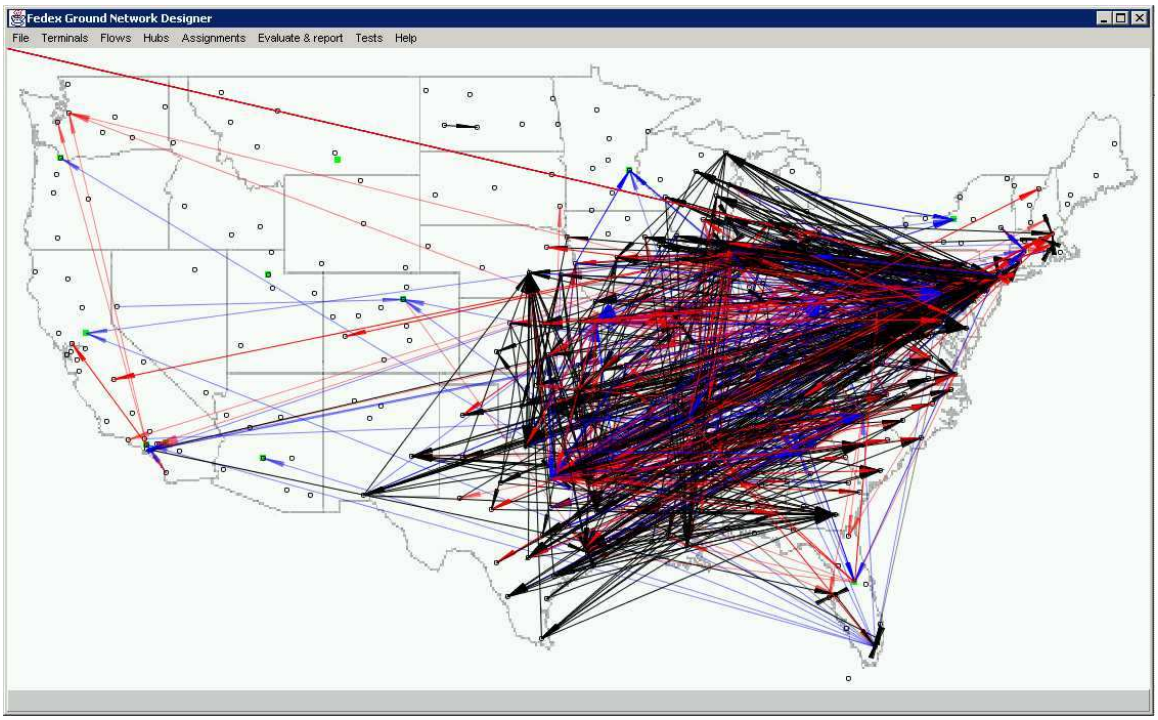


Figure 45: Load plan generated for direct load factors of 0.0. Clearly, this loadplan is extremely complicated compared to the pure hub-and-spoke network.

Figure 45 shows the hub-to-terminal and terminal-to-hub direct lanes. This loadplan is very complicated compared to the pure-hub-and-spoke network. We can now visualize the number of direct trucks originating from a single terminal or hub which could have been consolidated. This highly dense load plan results from the inability to centralize direct trucks at a hub or terminal within overlapping sub-networks and the inability to solve the routing sub-problems for each network to optimality.

5.2 Overlapping Origin – Destination Terminal Pairs in Sub-Networks

For the FedEx data set, the decomposition technique has 5,243 overlapping shipments because the origin and destination terminals are assigned to the same hub. Each of these 5,243 shipments overlap in $N-1(=23)$ sub-networks. 3,161 of the 5,243 shipments (about 60%) are routed uniquely in each of the 23 sub-networks. So overlapping due to the decomposition technique is not a issue for these shipments.

But 2,080 shipments are routed in more than one way in the 23 sub-networks. However, of these

2,080 about 75% (1,548 shipments) had identical routing in at least 19 of the 23 sub-networks.

Only 10% of the total shipments do not satisfy assumption 1 (see page 88). However, these routings are for the formulation where we impose no restriction the the size of a direct load. which makes the problem extremely hard because now every single package in a shipment can be sent as a direct load. Due to the increased difficulty of the problem these problems were solved only to 10% of optimality. So these solutions represent packages which are sent directly in the current solution but maybe be sorted in the optimal solution. Whether solving all the sub-networks to optimality would result in unique routings remains to be investigated but will require significantly greater computational power.

5.3 Approximate Solution for the Original Network

By decomposing the network we compromise on the quality of the solution. Solutions which may be optimal for the sub-networks may not necessarily be optimal for the whole network. In this section we discuss the cases in which there is a better solution than simply superimposing the solutions of the sub-networks.

5.3.1 Consolidating Direct Trailers

A possible improvement is in case of direct trailers. This is a simple improvement to visualize, especially in cases where triangle inequality holds for distance. Any time single trailers from two sub-networks bypass the same hub, it may be possible to consolidate them into a single truck at the hub.

Example 5.1 *Consider two sub-networks in which two terminals, t_1 and t_2, one in each sub-network send trailers directly to terminal t_3 in both the networks. Since the shipment routing for these two sub-networks was done independently there was no possibility to consolidate direct trailers to t_3 together. However, if both the sub-networks were optimized as a single network it may have been possible to consolidate the trailers at either the first or the second hub as shown in figure 46.*

Extending the example to a hub-to-terminal (terminal-to-hub) direct trailers, the second (first) hub can be a consolidation point for direct trailers as shown in figure 46.

Table 12 shows how flows in the sub-networks can be routed in a less expensive way when considered jointly with other sub-networks.

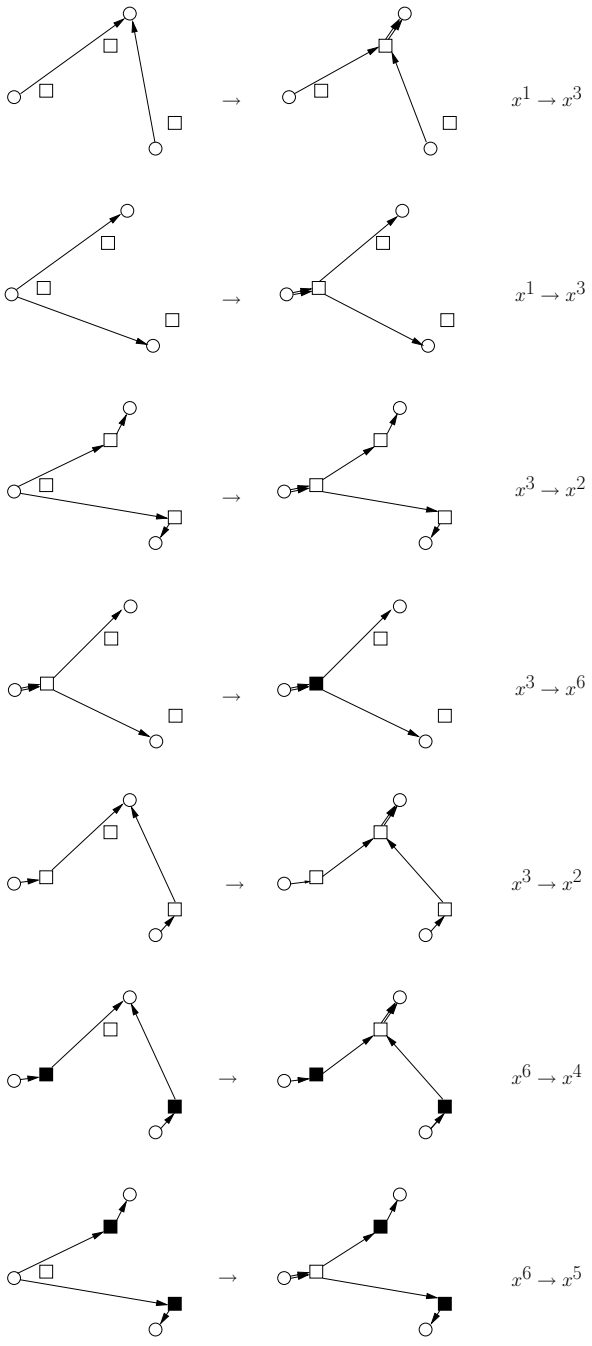


Figure 46: Consolidating shipments and/or trailers from two sub-networks at the overlapping hub can further reduce costs

5.3.2 Consolidating Shipments by Breaking Direct Trailers

Another possible improvement is to break direct loads and consolidate shipments. The trade-off lies in increased sorting costs versus possible decrease in transportation costs. Figure 46 also shows how direct shipments can actually be consolidated at hubs they bypass.

Table 12: Superimposing the shipment routes that are optimal in the sub-networks may yield sub-optimal shipment routes in original network

Shipment route in sub-network	Possible cheaper shipment route(s) in original network
x^1	x^3
x^2	x^4, x^5
x^3	x^2, x^6, x^6
x^4	x^7
x^5	x^7
x^6	x^4, x^5

5.4 Effect of Direct Load Sizes

The routing scheme we have discussed so far does not have any restrictions on the size of the direct load. Most LTL carriers impose restrictions on the size of the direct load. For example, FedEx Ground requires its longhaul trailers to be at least 75% utilized. In this section, we will discuss the influence of the *minimum required* size of the direct load on the network and its operations.

5.4.1 Operating costs

Proposition 1 *The optimal total operating cost decreases with the required direct load factor.*

If we can solve the shipment routing problem for the entire network to optimality, as we allow smaller sized direct loads the total operating cost should decrease. Consider the following constraints which require the average size of direct load to be D_1 and D_2 and also consider $D_1 > D_2$.

$$D_1 t^2_{ijkl} \leq x^2_{ijkl} \leq C t^2_{ijkl}$$

$$D_2 t^2_{ijkl} \leq x^2_{ijkl} \leq C t^2_{ijkl}$$

By combining these two equations we get,

$$D_2 t^2_{ijkl} \leq D_1 t^2_{ijkl} \leq x^2_{ijkl} \leq C t^2_{ijkl}$$

So any routing scheme which is feasible for minimum average direct load size of D_1 is also feasible for the minimum average direct load size of D_2. So by allowing smaller sized direct loads we are

relaxing the shipment routing problem and the total cost should be non-increasing as the minimum allowable average direct load size decreases.

However, even if the shipments are routed optimally in the sub-networks that are generated by the proposed network decomposition scheme, superimposing the solutions from the sub-networks may generate non-optimal shipment routing schemes for the entire network. The decomposition scheme has been specifically designed to centralize and co-ordinate the a direct loading plan between load planners at the hub and terminals. However, it "decentralizes" the direct loading plan at a terminal/hub for the various sub-networks.

Theoretically, a drawback of the proposed decomposition scheme is its sensitivity to direct load sizes, especially, when $D \leq C/2$. When $D \leq C/2$, trailers from the same terminal which are less than half-full may be directly loaded to different hubs (in different sub-networks). In presence of a centralizing scheme at a terminal, such as routing the shipments for the entire network optimally, sending most of these trailers directly may not be justified economically. This is because cost saving maybe achieved either by consolidating directly loaded trailers or by breaking directly loaded trailers and consolidating the shipments (see section 5.3). When $D \leq C/2$, consolidating shipments by breaking direct trailers eliminates excess trailers, which may result in significant cost savings. On the other hand, when $D > C/2$ consolidating shipments by breaking direct trailers does not eliminate excess trailers.

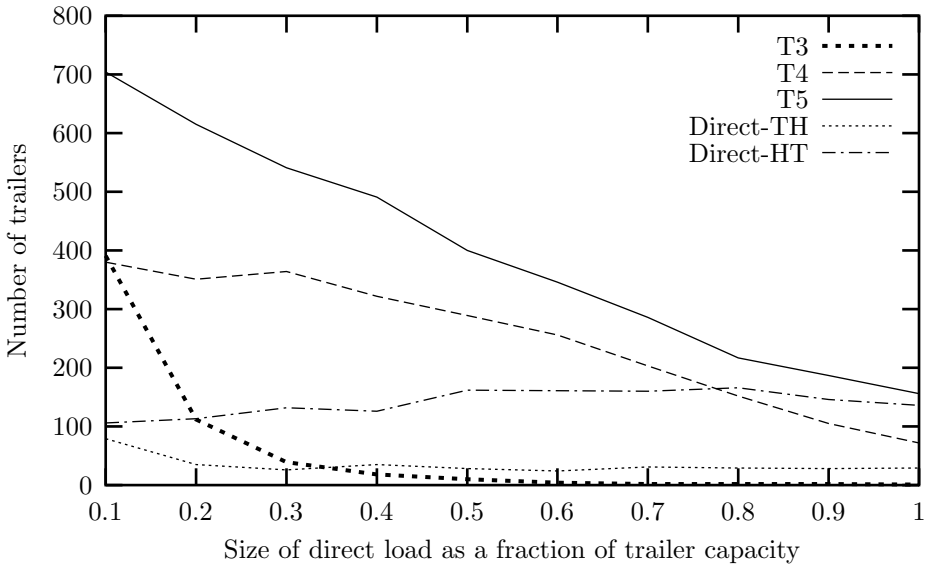


Figure 47: The number of direct trailers used increases as the load factor decreases. Using the proposed decomposition scheme, the number of direct trailers used between terminals assigned to the same hub increases drastically when the direct load factor is lowered from 0.2 to 0.1. This causes the transportation (and total) costs to increase when direct load factor is decreased below 0.2.

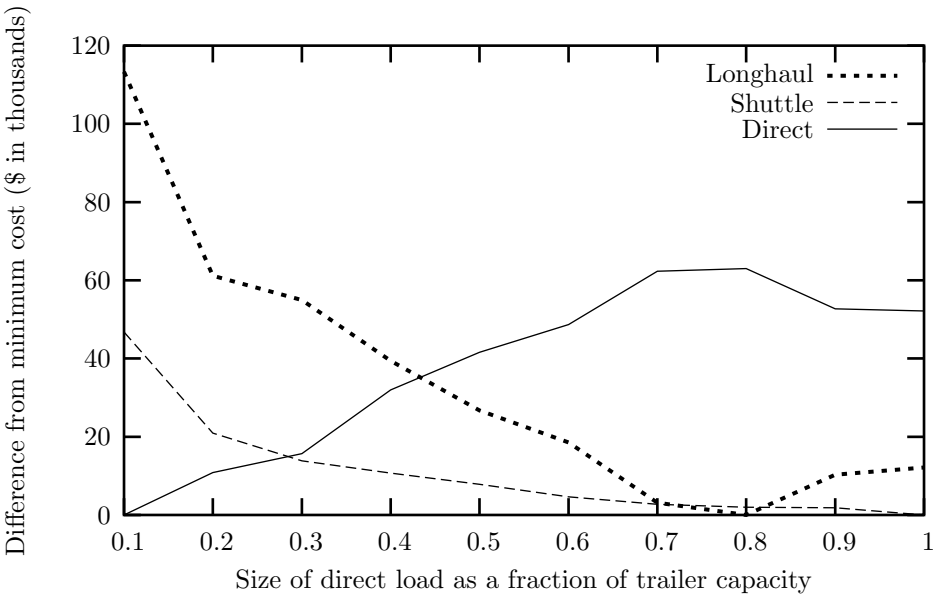


Figure 48: The longhaul and direct transportation costs are inversely related. The shuttle cost increases gradually when the minimum required direct load factor is increased from 0.2 to 1.

Practically, for the FedEx data set our decomposition scheme conforms with proposition 1 for load factors as low as 0.2. As expected, the number of trailers sent directly between terminals assigned to the same hub increases as the direct load factor is reduced. For load factors under 0.2 direct loads between terminals assigned to the same hub exhibit the drawback of the decomposition scheme — decentralization of the same terminal/hub in several sub-networks. The increase in these trailers is significant for direct load sizes under 0.3 (see figure 47).

Figure 48 shows how the individual transportation costs vary with the direct load factor. As the direct load factor is reduced from 1.0 to 0.8 more packages can now be sent directly and the direct load cost increases whereas the longhaul cost decreases. As the direct load factor is further reduced, it may be uneconomical to send a partially full trailer directly and some of these trailers may be sent through two hubs (but may not be sorted at both the hubs). Hence below load factors of 0.7 the direct load transportation cost decreases and the longhaul cost decreases. When direct load factors are further reduced to under 0.2 shuttle costs increase drastically because of directly loaded trailers between terminals assigned to the same hub.

For the proposed decomposition scheme, as the direct load factor decreases the transportation costs increase whereas the sorting costs decrease. There exists a trade-off between these two costs. For the FedEx data set our approach yields the least total cost at a direct load factor of 0.2. Figure 49 shows how these costs vary with the direct load factor.

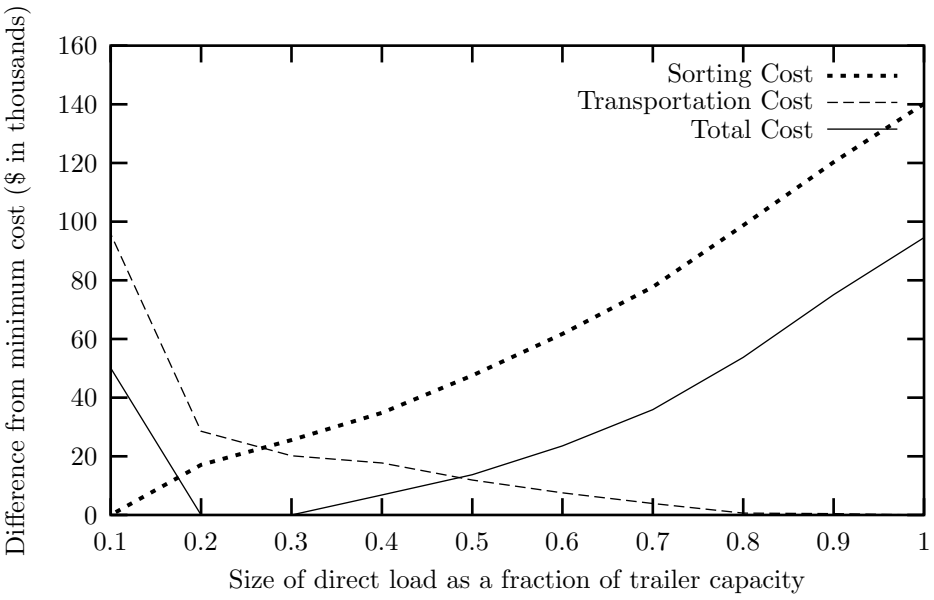


Figure 49: As the minimum required direct load factor is reduced the total cost decreases.

Another interesting observation is the distribution of the distances over which trailers are sent directly as a function of direct load factor. For a direct load factor of 1.0, though the trailers are entirely full more trailers are sent over shorter distances as compared to longer distances. This is possibly because of the inability to fill more trailers directly over longer distances because of freight patterns. Reducing the load factor to 0.6 we see that the number of trailers sent directly increases for shorter as well as longer distances. For direct load factors of 0.4 and under (less than half of trailer capacity) more trailers are sent directly over shorter distances and since it is uneconomical to sent partially full trailers over longer distances the number of direct trailers decreases as the direct load factor is reduced below 0.5. In figure 50 we can see that as the direct load factor decreases the average distance over which trailers are sent directly decreases.

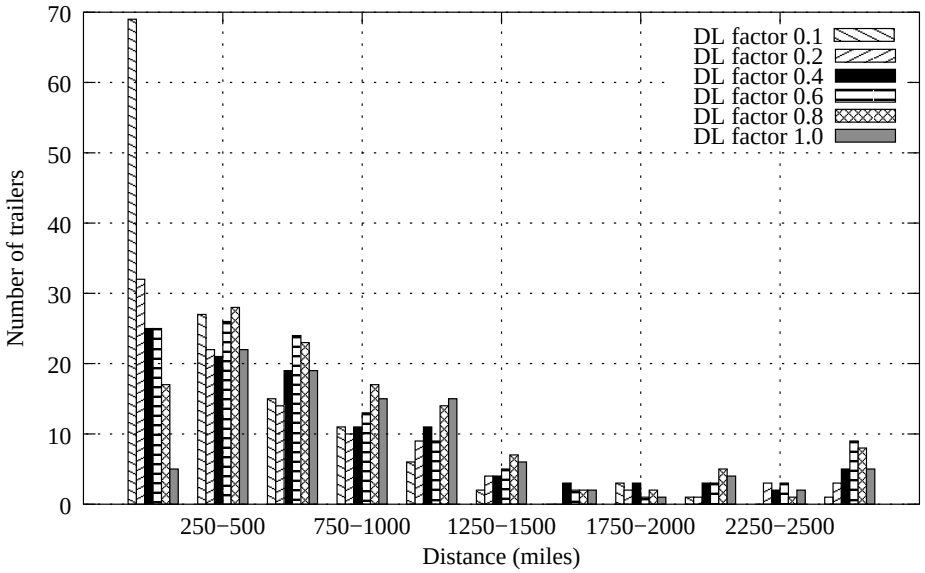


Figure 50: As we decrease the direct load factor the number of trailers sent directly over shorter distances increases whereas those sent over longer distances decreases.

5.4.2 Routing of shipments in the network

Figure 51 shows how the minimum required size of the direct load affects the routing of shipments in the network. Decreasing the direct load factor implies that we are willing to send less utilized trailers directly. As expected, when we decrease the direct load factor, more packages are sent directly.

For the FedEx Ground data set, if no direct loads are allowed then about 89% of the packages will be sent through and sorted at both the hubs. And approximately 11% of the packages will be sorted at the single hub it passes through. Decreasing the direct load factor, we see that the percentage of packages that are sent by the most expensive route decreases. These are routed via less expensive routes which either bypass a hub or avoid sorting at the hub. Table 13 compares the distribution of packages when the direct load factor is 1.0 versus 0.5.

To get an intuition for the effect of direct load factor on shipment routes it may be useful to consider the effect of decreasing direct load factors. For example, when no direct loads are allowed 89% of the packages are sent through both the hubs and sorted at both the hubs. Instead if we allows trucks to be sent directly only if they are entirely full by wisely consolidating packages 8.4% of the shipments are sent directly from terminals to second hub, bypassing the first hub. However, there may be terminals which fail to fill up a trailer to a hub by a small fraction of the trailer, say 10% (100 packages). Due to the restriction on direct load factor, the 900 packages will not be sent directly. Instead, they will be sent through both the hubs and sorted. However, sending these 900

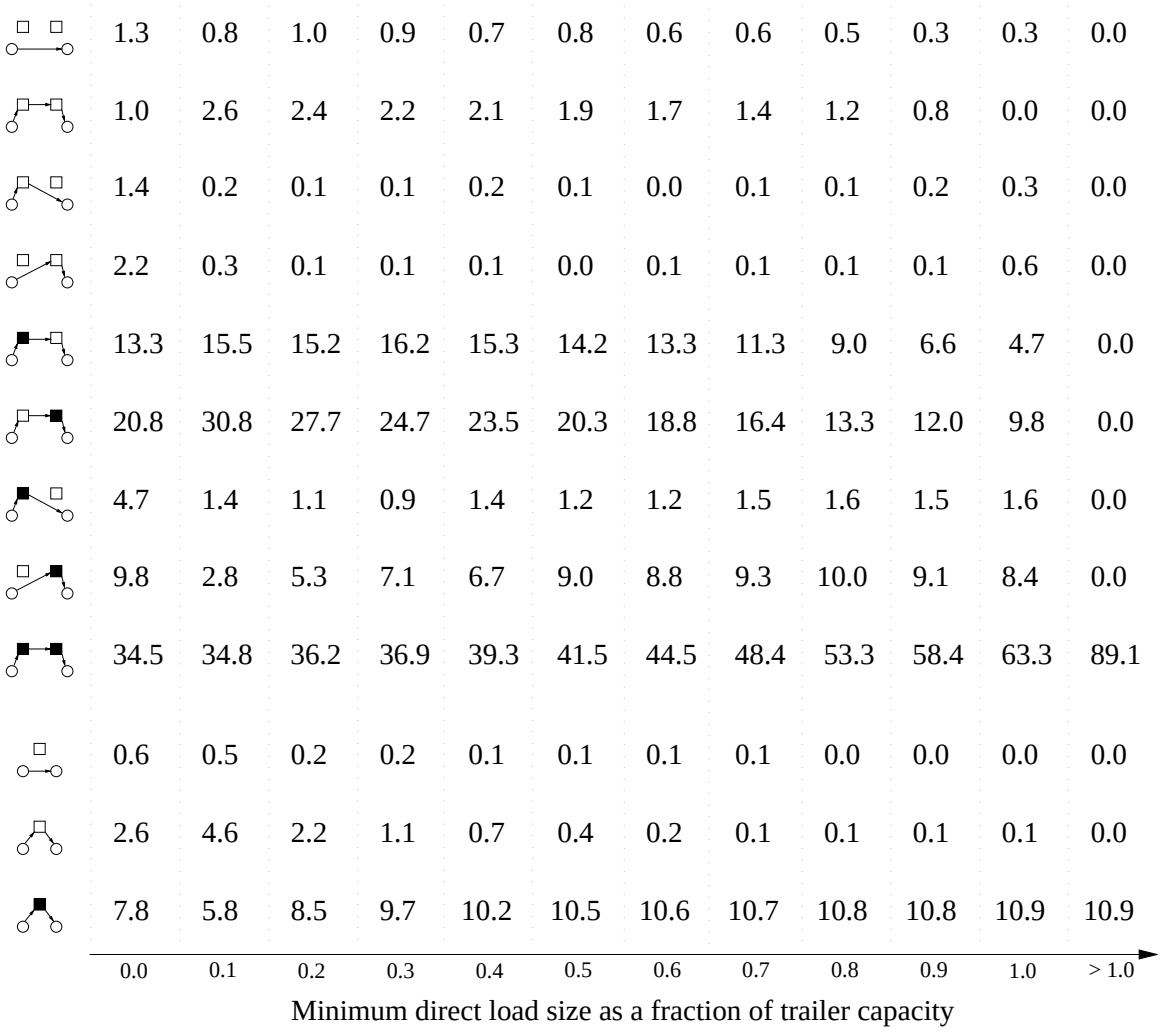


Figure 51: As we allow smaller sized direct loads more shipments are delivered by cheaper routes.A direct load factor of "> 1.0" means that no direct loads are allowed, only default loads.

Table 13: As we decrease the direct load factor more shipments either bypass a hub or avoid sorting at a hub.

	Direct Load Size		
Shipment route	1.0	0.5	0.1
Double stop, double sort	63.3%	41.5%	34.8%
Double stop, single sort	14.5%	34.5%	46.3%
Double stop, no sort	0.0%	1.9%	2.6%
Single stop, single sort	10.0%	10.2%	4.2%
Single stop, no sort	0.9%	0.1%	0.5%
No stop, no sort	0.3%	0.8%	0.8%
Total	89.0%	89.0%	89.0%

packages directly to avoid sorting costs may justify the increase in average transportation cost per package.

In figures 52-56 we can visualize how the direct loads are generated when the direct load factor is changed.

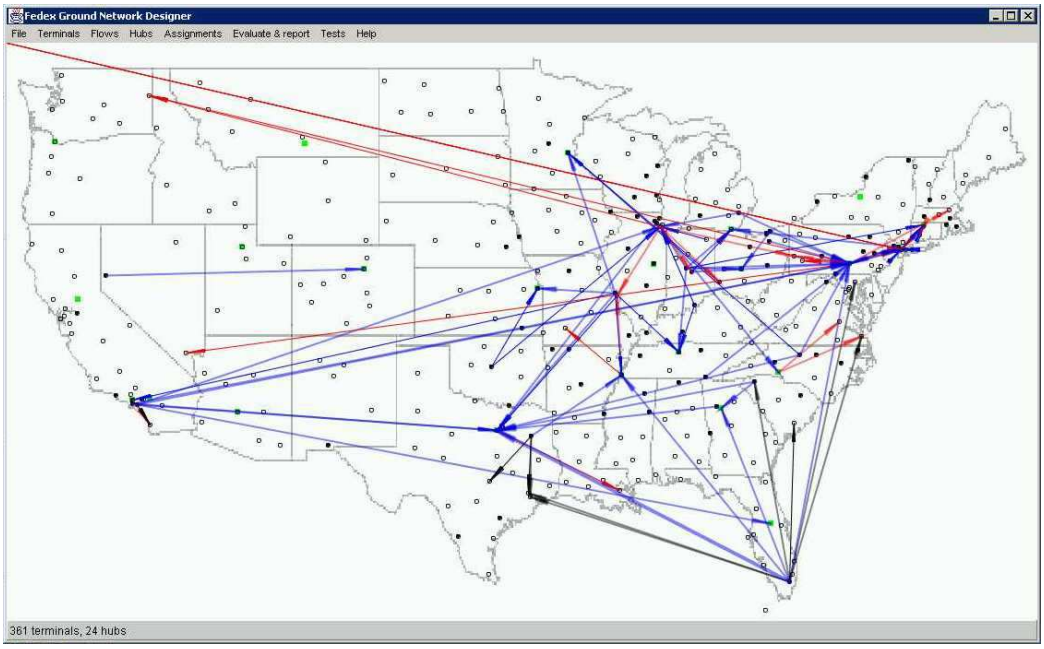


Figure 52: Load plan generated for direct load factor of 0.2.

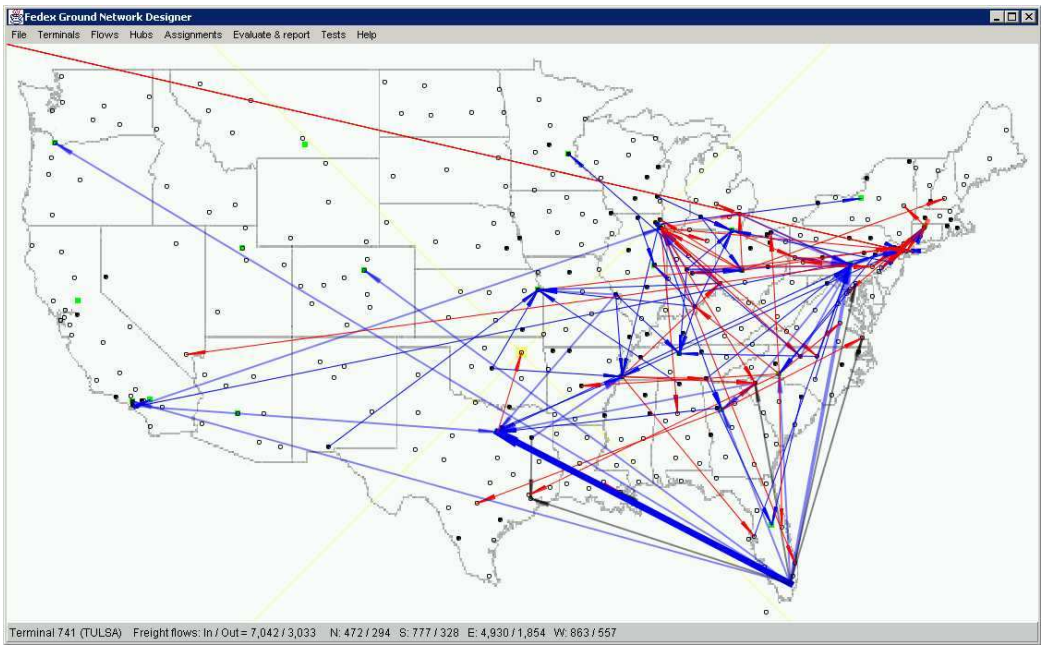


Figure 53: Load plan generated for direct load factor of 0.4. As expected, most directly loaded trucks are pulling two trailers. However, there may be instances of freight patterns where even sending a single trailer directly may be economical as shown by the dark arrow from Miami, FL to Fort Worth, TX.

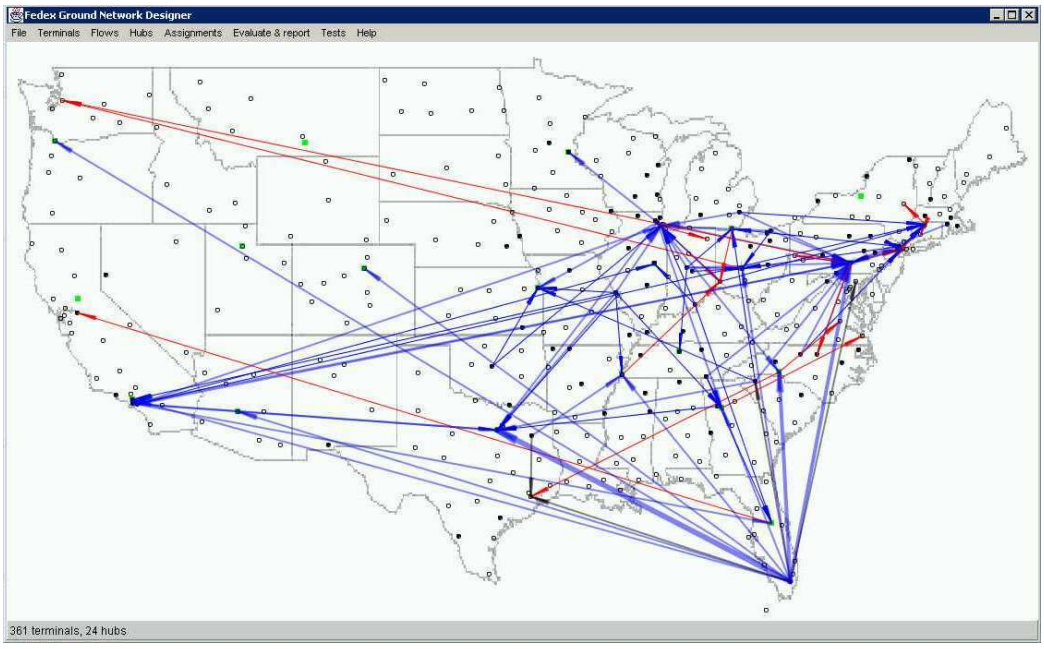


Figure 54: Load plan generated for direct load factor of 0.6.

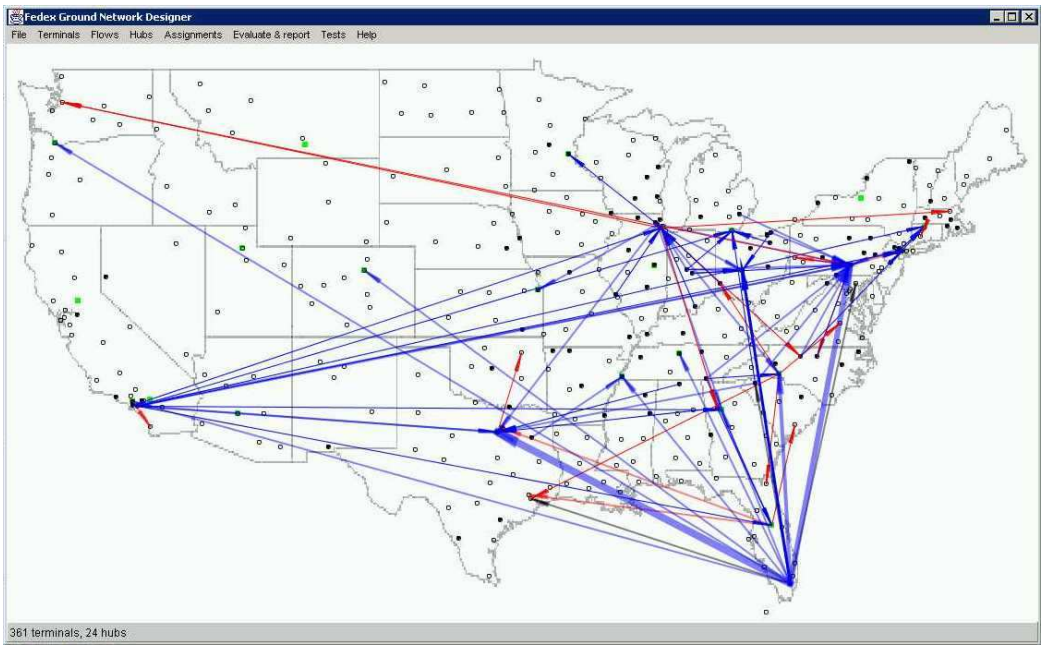


Figure 55: Load plan generated for direct load factor of 0.8.

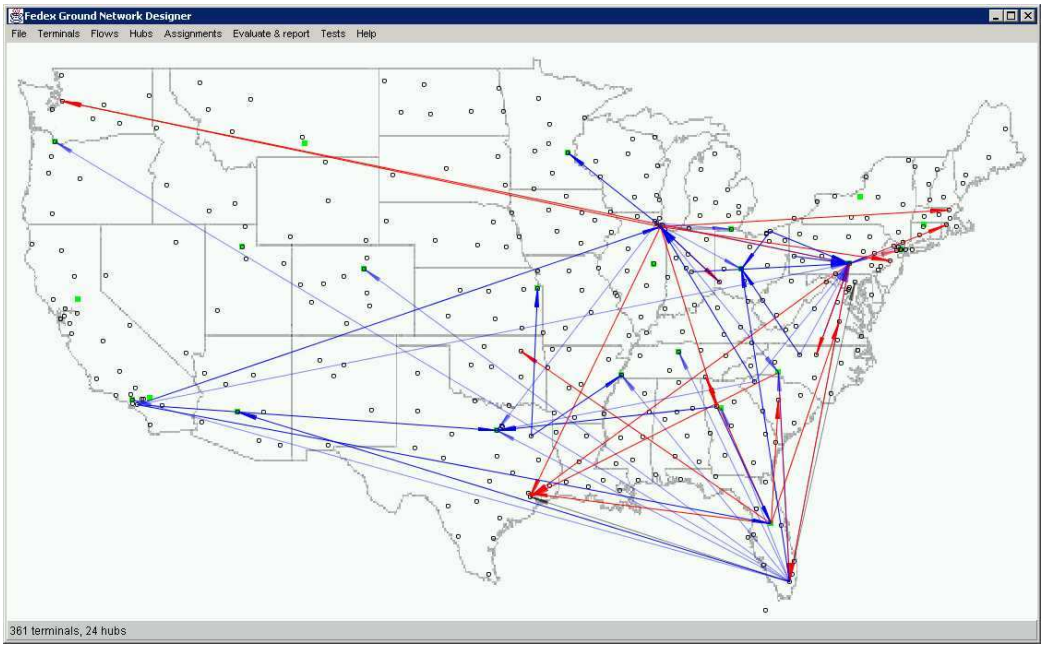


Figure 56: Load plan generated for direct load factor of 1.0.

5.4.3 Average trailer utilizations

Current industry practice is to generate load plans that maximize trailer utilizations. Most LTL carriers evaluate their load plans based on the number of packages put on the trailer. FedEx Ground requires its direct trailers to be at least 75% utilized. In section 5.4.1 we observed that it can be cheaper to send some trailers that are only 30% utilized than to required 75% utilization for all trailers. In fact, in our solution we found that the average realized utilization of trailers is much higher than the imposed minimum required direct load factor.

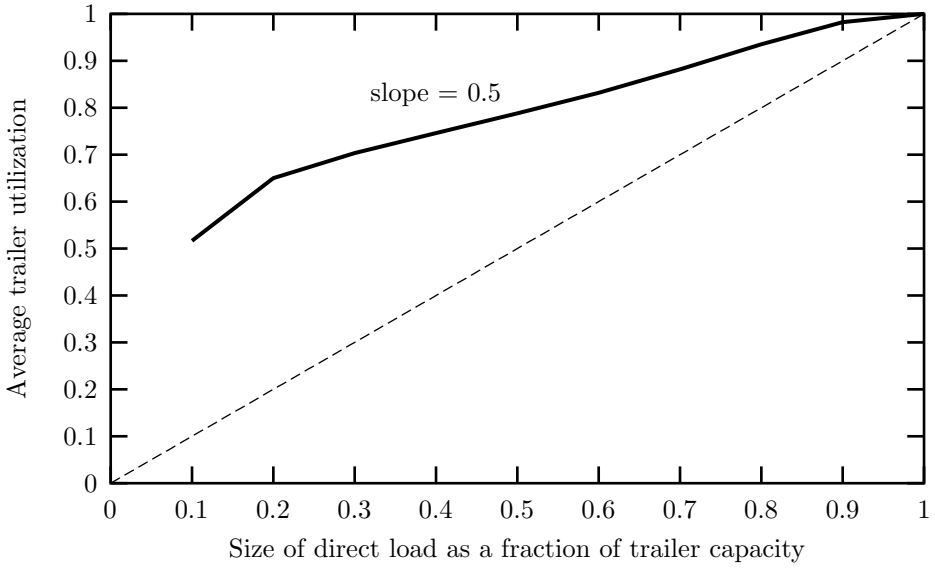


Figure 57: Though the LTL carrier may be willing to send trailers which are not almost full, the average utilization of the trailers sent directly is much higher.

Figure 57 shows how the actual average trailer utilization varies with the direct load factor. For every percentage increase in the direct load factor the average trailer utilization increases by 0.05%.

The average trailer utilization in the network is calculated as

$$\text{Average trailer utilization} = \frac{\text{total number of packages loaded directly}}{\text{total number of trucks routed directly}}$$

There is an inverse relation between the number of packages (and the number of trucks) sent directly and the direct load factor. Consider a percentage decrease in direct load factor. As the direct load factor has decreased more packages are diverted from the traditional hub-and-spoke route to one of the direct loaded routes. For sake of argument let us assume that the number of trailers (and trucks) remain unchanged. For the average utilization to remain the same, on an average 10 additional packages have to be put on every existing trailer. It may not be possible to extract 10 additional packages to be directly loaded on every lane which explains why the slope in figure 57 is less than 1.

Whether the value of 0.5 for the slope is a characteristic of the network still needs to be ascertained. One way to ascertain this is to see the how this slope value changes when the freight flows are perturbed slightly and forms part of future research.

5.4.4 Service level

In our discussion so far, service level was not a consideration. However, it would be worthwhile answering the following question: ' *"What is the service level offered by our loadplan when we design the network'?"*

We define service level as the percentage of packages in the network that are delivered on-time.

The actual transit time of a package depends on the routing from its origin terminal to its destination terminal. Based on the seven possible routes introduced in section 3.3 we have the transit time comprising of the following three components:

Driving time (t_d) : The time it takes to drive the truck carrying the package along the route suggested by the load plan.

Package handling time at hub (t_h) : The time it takes to unload, sort and reload the package at the hub. The trailer that carries this package is opened at the hub and all the packages are sorted.

Trailer matching time at hub (t_m) : The time it takes to match trailers at a hub. The trailer that carries this package is not opened at the hub and the packages contained in it are not sorted.

We assume the driving speed (s) for the trucks to be constant whereas, the time to handle a package and match the trailer to be normally distributed.

$$t_h \sim N(\mu_h, \sigma_h^2)$$
$$t_m \sim N(\mu_m, \sigma_m^2)$$

Let $t_{t,p}$ denote the total transit time of a package p from its origin to its destination. Then,

$$t_{t,p} = t_d + t_m + t_h$$

We can compute the mean ($\mu_{t,p}$) and variance ($\sigma_{t,p}^2$) for the total transit time for each of the flow routes discussed in section 3.3. Refer figure 22 for the x variables.

With a $(1-\alpha)\%$ probability the transit time for the package p on a given route is less than or equal to $\mu_t + z_\alpha \cdot \sigma_t (= t_{\alpha,p})$, where $P(N(0,1) \geq z_\alpha) = \alpha$. If $t_{o,p}$ is the promised delivery time for

Flow path for package p	$E[t_{t,p}] = \mu_{t,p}$	$Var[t_{t,p}] = \sigma^2_{t,p}$
x^1_{ij}	d_{ij}/s	0
x^2_{ijkl}	$(d_{ik} + d_{kl} + d_{lj})/s + 2 \cdot \mu_m$	$2 \cdot \sigma^2_m$
x^3_{ijk}	$(d_{ik} + d_{kj})/s + \mu_m$	σ^2_m
x^4_{ijkl}	$(d_{ik} + d_{kl} + d_{lj})/s + \mu_m + \mu_h$	$\sigma^2_m + \sigma^2_h$
x^5_{ijkl}	$(d_{ik} + d_{kl} + d_{lj})/s + \mu_m + \mu_h$	$\sigma^2_m + \sigma^2_h$
x^6_{ijk}	$(d_{ik} + d_{kj})/s + \mu_h$	σ^2_h
x^7_{ijkl}	$(d_{ik} + d_{kl} + d_{lj})/s + 2 \cdot \mu_m$	$2 \cdot \sigma^2_h$

the package, then with a confidence of $1 - \alpha$ we can determine whether a package was delivered on time or not. Let y_p denote if the package p was delivered on time. Then,

$$y_p = \begin{cases} 0 & \text{if } t_{\alpha,p} > t_{o,p} \\ 1 & \text{otherwise} \end{cases}$$

With a confidence level of $(1 - \alpha)$ we can then calculate the service level of all the packages within a certain set P as,

$$\text{service level} = \frac{\sum_{p \in P} y_p}{|P|}$$

For analysis, we tried to determine the effect of the distance over which the package is sent on the service level. Since the data for the offered delivery times was unavailable we used the following rule of thumb prevalent in the industry: An average of one day in transit for every 500 miles between the origin and destination. [pers. com. Langley, 2003]

	Mean, μ (hours)	Std. Deviation, σ (hours)
Package handling time, t_h	16	2
Trailer matching time, t_m	8	1

Table 14: Values used to compute the service level offered within the network by our load plans.

We have best guessed the values in table 14 based on conversations with industry and academic experts [pers. com. Langley, 2003]. As best as we have tried to pick the appropriate values for these parameters, our analysis is intended to reveal trends in service level rather than provide exact numbers for levels of service provided by the carrier.

As seen in figure 58, for a given minimum required direct load factor the level of service improves over the distance. One possible explanation is that as the distance over which the package travels increases, the effect of variability in the time spent at the hub decreases substantially.

As expected, as the minimum required direct load factor is increased the service level drops. This is because some trucks which could be loaded directly may no longer be allowed due to increased

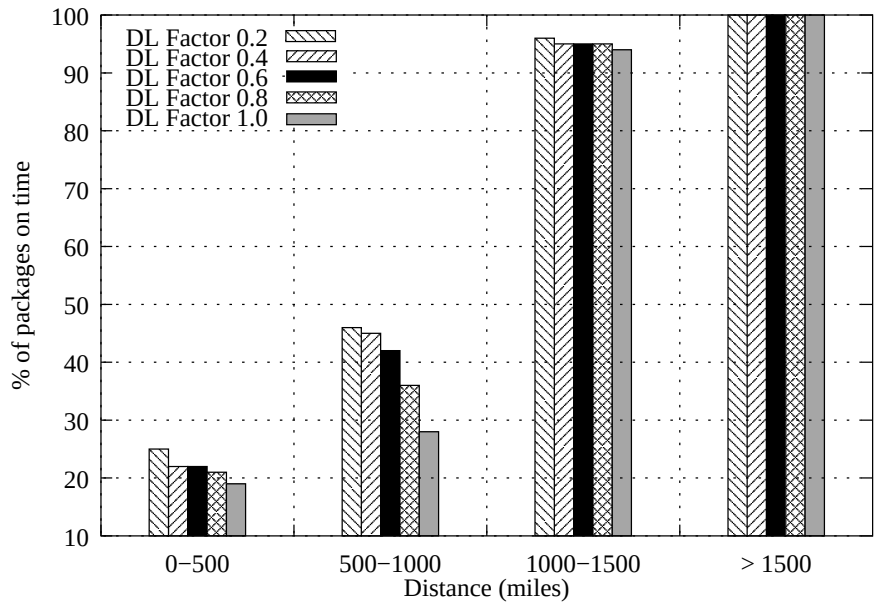


Figure 58: For a given minimum required direct load factor, the service level increases with the distance. Also, for a given range of distances over which packages are sent, as the minimum required direct load factor increases the service deteriorates.

restrictions on utilizations. Hence, more packages are now routed through addition hub(s) and sorted, increasing the transit time.

For the values we have chosen, for packages sent between terminals under than 500 miles apart, less than 30% of the packages are delivered on time. This is because the rule of one day for every 500 miles of travel is likely to be too stringent. The average distance between a terminal and an assigned hub is 170 miles. So if the packages between terminals less than 500 miles apart are sorted at the common hub, a package sent between an average terminal pair assigned to the same hub would not be delivered on time.

5.4.5 Total number of trucks and trailers

In this research, one of the assumptions we make is that only after all the packages in the current "period" are delivered, new packages will enter the network in the following "period". Our estimate for the total number of trucks and trailers in the network is an underestimate for the actual number of trailers in the network for the following two reasons:

1. We estimate the number of trailers and tractors based on a snapshot view of the LTL network within a period. Suppose it takes n days to travel from location i to location j. We consider that only 1 tractor is used on that leg since new packages are not entering the network until the current packages are delivered. Whereas in practice, since the network experiences a daily

demand for the packages, there maybe n tractors moving freight on that leg.

2. Some trailers maybe used as cross-docks or warehouses and held up even if the freight is not unloaded.

Let,

t_j^{out} : number of trailers sent out from location j

t_j^{in} : number of trailers received by location j

t^{total} : total estimated number of trailers in the LTL network

T_j^{out} : number of tractors sent out from location j

T_j^{in} : number of tractors received by location j

T^{total} : total estimated number of tractors in the LTL network

The total number of trailers in the network is at least equal to the total number of trailers sent out from each of the terminals. Hence, we get

$$t^{total} \geq \sum_{j \in \mathcal{T}} t_j^{out}$$

The equality holds when a hub, which acts as both consolidation and sorting center, does send out more trailers than it receives. If this is not the case, then in addition to $\sum_{j \in \mathcal{T}} t_j^{out}$ trailers we need to account for the difference between the number of trailers that leave the hub and arrive at that hub. Hence, the total number of trailers in the LTL network is,

$$t^{total} = \sum_{j \in \mathcal{T}} t_j^{out} + \sum_{k \in \mathcal{H}} \max\{t_k^{out} - t_k^{in}, 0\}$$

Based on the same argument we have,

$$t^{total} = \sum_{j \in \mathcal{T}} t_j^{in} + \sum_{k \in \mathcal{H}} \max\{t_k^{in} - t_k^{out}, 0\}$$

and

$$T^{total} = \sum_{j \in \mathcal{T}} T_j^{out} + \sum_{k \in \mathcal{H}} \max\{T_k^{out} - T_k^{in}, 0\} = \sum_{j \in \mathcal{T}} T_j^{in} + \sum_{k \in \mathcal{H}} \max\{T_k^{in} - T_k^{out}, 0\}$$

Figure 59 shows that as we the increase the minimum required direct load factor, the requirements for trailers and tractors decreases. Though this is not a representative estimate for the total number of tractor and trailers actually present within the network, this graph provides an insight into the how the direct load factor affects the tractor and trailer needs in the network.

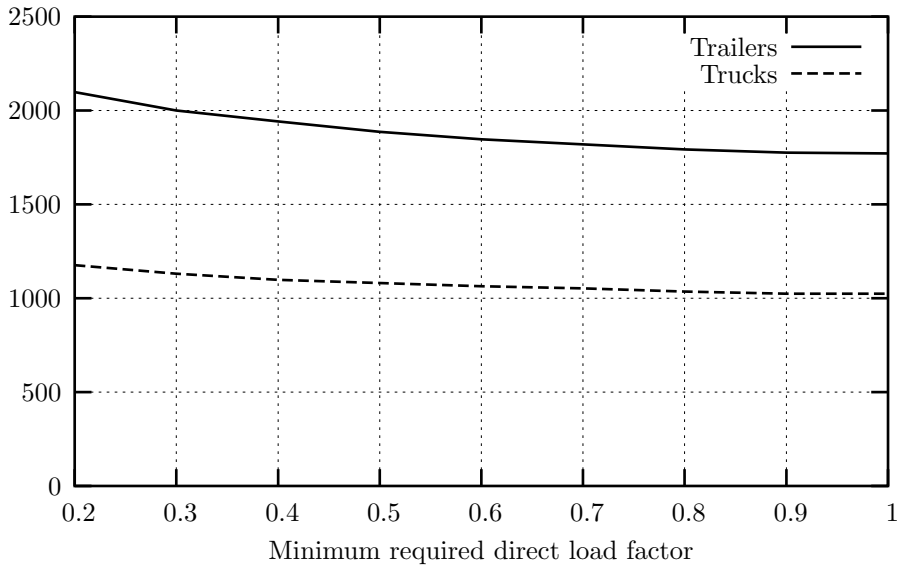


Figure 59: As we increase the required direct load factor the total number of trailers and trucks required in the system decreases.

Moreover, this estimate can be used for an economic analysis to determine whether savings yielded by reducing the required minimum direct load factor dominate the resulting increases in cost of transportation equipment.

CHAPTER 6

ROUTING OF EMPTY TRAILERS

6.1 Problem Description

To hedge themselves against empty backhauls most LTL carriers try to maintain a regional balance (symmetry) of shipments within their network. However in a package delivery network the flow of packages is usually not symmetrical, especially when the carrier cannot be selective about its customers. As a result the number of incoming trailers at a hub (sorting facility) may not be equal to the number of trailers sent from that hub. To balance the fleet in the network, trailers must be redirected from hubs having excess trailers to hubs that are deficit in trailers. To do so most LTL companies have the following options:

1. **Outsourcing the trailers from third party logistics (3PL) providers** One way to reduce the imbalance is to balance the flows. But if that is not possible another way is to externally balance the trailers by renting the "extra" trailers. By doing so the carrier gets rid of the problem of minimizing empty backhaul miles by incurring a cost.

2. **Transporting the trailers back by rail** One of the other ways to reduce empty backhaul costs is to transfer the empty trailers back by rail. Though this is slightly cheaper alternative than renting trailers from 3PL providers one distinct disadvantage is the transit time to transport the trailers.

3. **Minimizing the empty backhaul miles** Finally, some LTL companies use on optimization models to re-balance the empty trailers by their own tractors.

To motivate the idea we first formulate this problem as a transshipment problem, explain its advantages and disadvantages and then improve the formulation to make it more exact.

6.2 Continuous Cost Model

A simple approach is to formulate the problem as a transshipment network flow model. Using twin trailer combinations yields a non-linear cost function. For example, it costs the same to send 4 trailers from hub i to hub j as it cost to send 3 trailers. In either case, we are sending 2 trucks from hub i to hub j. As a first approximation, in the continuous cost model we assume that the cost of

sending trailers from hub h_i to hub h_j is a linear function of the number of trailers sent and the distance. That is, we approximate the tractor miles with trailer miles. A better model would be to consider the cost function as a step function.

This model does not take into account any scheduling within a day, so that it may generate a solution that is technically infeasible. For example, hub h_i may require trailers by noon but they do not arrive until later.

$$\text{Minimize} \sum_{ij} c_{ij} D_{ij} \frac{t_{ij}}{2}$$

Subject to:

$$\sum_{j} (t_{ij} - t_{ji}) = d_i \tag{61}$$

$$t_{ij} \geq 0 \tag{62}$$

$$t_{ij} \quad \text{integer} \tag{63}$$

where,

c_{ij} = cost per mile of sending a truck from hub i to hub j

D_{ij} = distance between hub i and hub j

t_{ij} = number of trailers sent from hub i to hub j

d_i = excess/deficit trailers at hub i

Constraint 61 is the supply and demand constraints (the number of trailers that a hub can send is equal to that it has in excess and a hub receives exactly the number of trailers that it requires). Note that $d_i > 0$ implies that hub i is a freight sink and has excess trailers whereas $d_i < 0$ implies that hub i ships more loaded trailers than it receives and has a trailer deficit for next-day operations. Since a hub receives trailers only from other hubs within the network, $(\sum_i d_i) = 0$. In this case, the network is balanced and we can use equality constraints. Constraints 62 and 63 are non-negativity and integrality constraints. Since the trailer demands are integer and there are no capacity restrictions on the inter-hub routes, relaxing the integrality constraints and solving the problem as a linear program still yields an integer optimal solution [Chvatal, 1992]

6.3 Stepwise Cost Model

The previous model accounted for the trailer miles and not the tractor miles. In this model the marginal cost of a single-trailer tractor pulling another trailer is zero. This is accounted for by the step-wise constant cost function. Eckstein and Sheffi [1987] provides a model to balance tractors

and trailers for the group line-haul movements. He uses Lagrangian branch-and-bound procedure to solve randomly generated instances. We have adapted the model to route empty tractors and trailers for the line-haul movements.

$$\text{Minimize} \sum_{ij} c_{ij} D_{ij} T_{ij}$$

Subject to:

$$\sum_{j} (t_{ji} - t_{ij}) = d_i \tag{64}$$

$$T_{ij} \geq \mathrm{t}_{ij}/2 \tag{65}$$

$$t_{ij} \geq 0 \tag{66}$$

$$t_{ij}, T_{ij} \quad \text{integer} \tag{67}$$

where, T_{ij} = number of tractors sent from hub i to hub j.

This IP model differs from the LP model in constraint 65 which allows for consolidation of trailers (to form trucks) at a hub and constraint 67 which enforces that fractional tractors cannot be sent.

6.4 Results

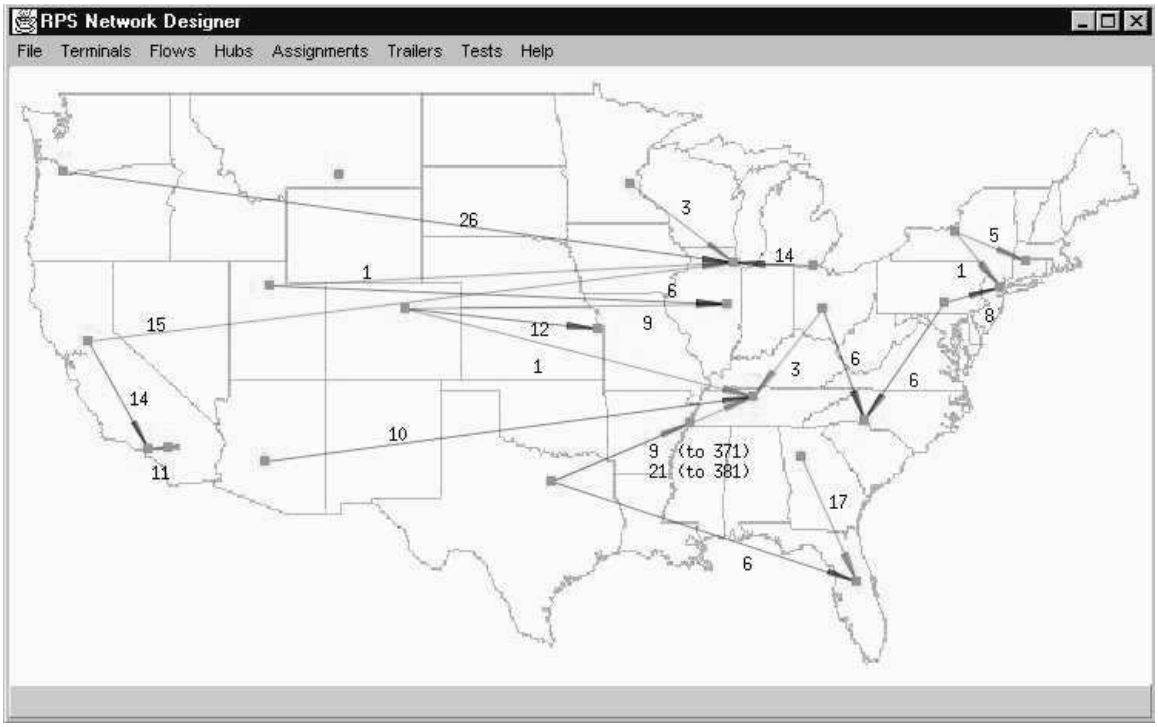
The following table shows the output of the optimization model described above. The output provides the number of trailers sent from a supply node (hub) to a demand node (hub). The empty trailer routing solutions is shown in figure 60.

Example 6.1 *For the continuous cost model (see figure 60(a)) Sacramento, CA and Salt Lake City, UT send 15 trailers (7.5 trucks) and 1 trailer (0.5 trucks) to Chicago, IL. respectively. However, the stepwise constant cost model avoids sending single pups individually to a common destination over longer distances. As seen in figure 60(b) Sacramento sends 14 trailers (7 trucks) to Chicago, IL directly. It redirects the single trailer to Salt Lake City, UT where it piggy-backs with the single trailer en route to Chicago.*

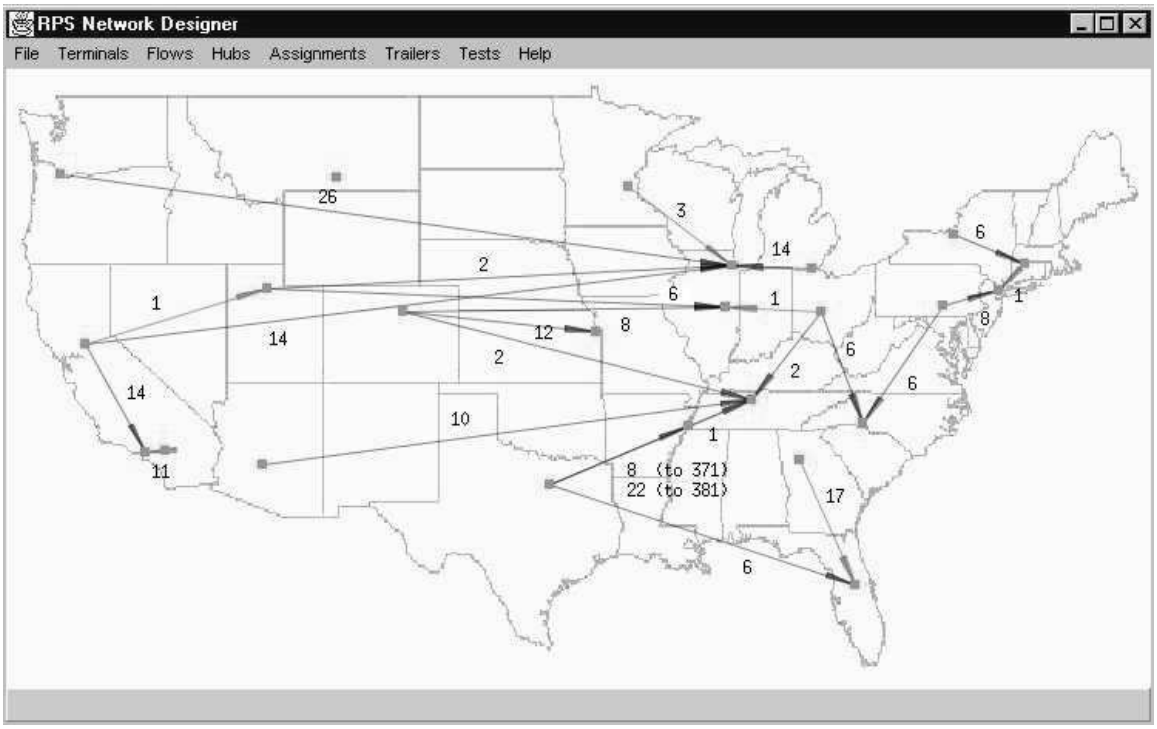
Computational Results

The computations were performed on Sun Ultra 80 Model 2450, 2x450-MHz UltraSPARC-II, 4-MB L2 Cache, 1-GB Memory running on Solaris 8. We used CPLEX v8.1 to route the empty trailers model.

The continuous cost model can be easily solved to optimality. The stepwise cost model generates solution within 1.0% of the optimal value within 35 seconds.



(a) Solution for the continuous cost model



(b) Solution for the stepwise cost model

Figure 60: Inter-hub recirculation of empty trailers

6.5 Consolidating Empty and Loaded Trailers

So far we have considered the recirculation of the empty trailers as an independent subproblem. However, it must be noted that the total tractor movement costs can be further reduced if the empty trailers were allowed to be combined (consolidated) with the loaded trailers.

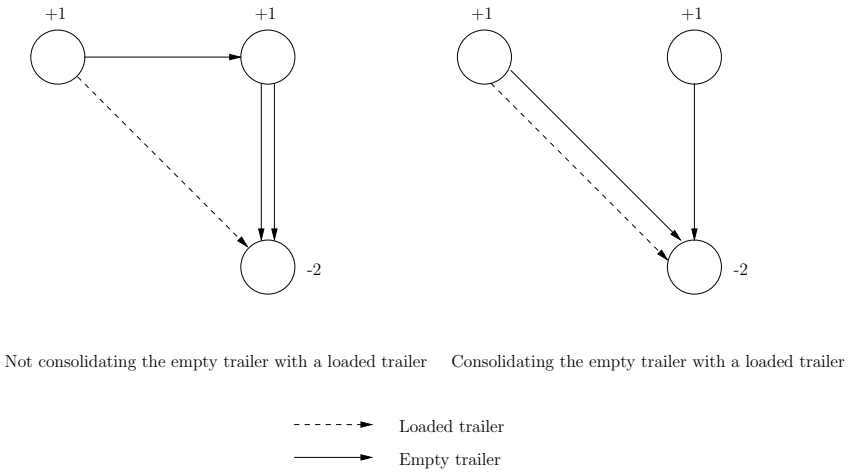


Figure 61: By consolidating the empty and loaded trailers, we can reduce the single-trailer miles and reduce costs.

$$\text{Minimize} \sum_i \sum_j c_{ij} D_{ij} T_{ij}$$

$$\sum_j \{t_{ij} - t_{ji}\} = d_i \tag{68}$$

$$T_{ij} \geq (t_{ij} + l_{ij})/2 \tag{69}$$

$$T_{ij} \quad \text{integer} \tag{70}$$

where, l_{ij} = number of loaded trailers sent from hub i to hub j.

This IP model differs from the previous IP model in constraint 69 which allows for consolidation of empty and loaded trailers. On our data we found that consolidating empty trailers with the loaded trailers resulted in savings of 68% compared to the policy where empty trailers are not allowed to consolidate with loaded trailers.

6.6 Routing Truck Tractors

The industry focus has been primarily on the recirculation of the empty trailers accumulated at a hub to other hubs in need of empty trailers to load the shipments. However, one of the tacit assumptions made in this approach is that a truck tractor pulls two trailers in its long-haul. This might not always be the case.

By adding the following constraint to the empty trailer recirculation constraint set the tractor demands will be satisfied,

$$\sum_j \{T_{ij} - T_{ji}\} \quad = \quad 0 \tag{71}$$

which says that at the end of the day the total number of tractors sent out by each terminal has to be equal to the total number of tractors received.

Adding this constraint will not make the original problem infeasible. It is easy to verify that $t_{ij} = l_{ji}$ and $T_{ij} = \lceil \frac{l_{ji}+l_{ij}}{2} \rceil$ is a feasible solution.

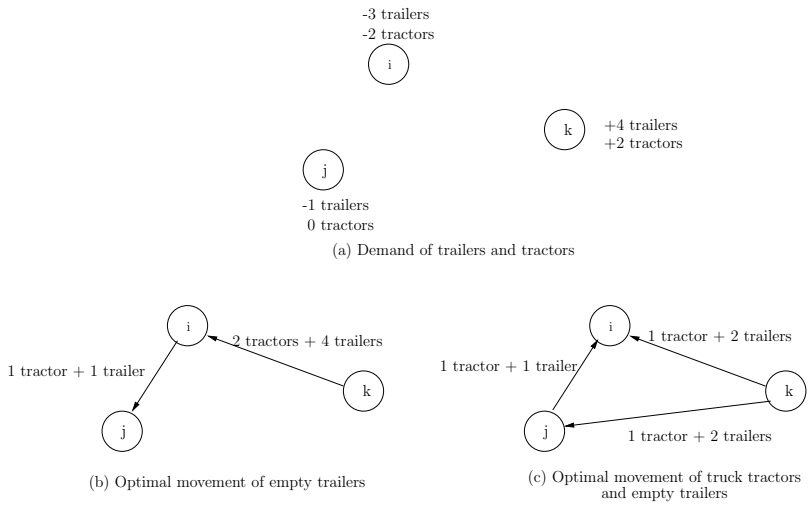


Figure 62: Balancing tractors and trailers

Example 6.2 *To see how this constraint may change the solution consider the example of three hubs in figure 62. Hub i needs 3 trailers and 2 tractors; hub j needs 1 trailer and hub k has an excess of 4 trailers and 2 tractors. As per the trailers recirculation models introduced before the optimal movement of empty trailers is shown in figure 62(b). Though the trailer demands are satisfied at*

each hub, hub i is still in deficit of 1 trailer whereas hub j has an excess trailer. Recirculating the tractors along with the empty trailers yields the solution shown in figure 62(c). In this solution a tractor pulling 2 trailers has been re-routed from hub k to i via hub j instead of routing it from hub k to j via hub i. This re-routing satisfies the tractor demands at all the hubs.

For the FedEx data set we observed that when we enforced balancing of tractors the cost increased insignificantly (by less than 0.075%). Table below lists the tractor/trailer routing costs for each of the models.

Tractors not balanced:		
empty and loaded trailers routed independently	:	1,247,740
combination of empty and loaded trailers allowed	:	1,181,590
Tractors balanced:		
empty and loaded trailers routed independently	:	1,248,670
combination of empty and loaded trailers allowed	:	1,181,800

6.7 Direct Loaded Trailers

In our model we have balanced the empty trailers only on the inter-hub lanes. The difference in the inbound and outbound trucks from (to) a hub to (from) a terminal are balanced individually on each shuttle lane. This is reasonable when we do not allow direct loads. However, when terminal-to-hub and hub-to-terminal direct loads are allowed the model may be extended to include terminals as either one of demand, supply or transshipment points. For the FedEx data set this increases the size of the network almost by a factor of 7 which may suggest computational intractability for the MIP models. Future research may be directed towards incorporating the knowledge of load plan (direct-load trailers) to minimize the empty trailer miles.

CHAPTER 7

HUB LOCATION

Our focus so far has been to optimize the network operating costs for a given configuration of terminals and hubs. As mentioned earlier, selection of terminals is demand-driven and beyond the scope of this research. The performance of any LTL freight system is inherently limited by the design of the freight network (for example, the locations and capacities of hubs). Thus, LTL operations may be optimized yet not be the best possible because they are constrained by the network design.

The hub location problem involves locating an appropriate number of hubs in the network to minimize the total costs. The hub location problem is also commonly modeled as a p-median problem. In the p-median problem we have to select p hubs from a given set of hub locations and the objective is to minimize the total costs.

7.1 Literature Review

O'Kelly [1987] presents a quadratic integer program for the p-hub median problem. Campbell [1990a] analyzes freight routing schemes for routing freight shipments via hubs — nearest terminal, minimum distance and minimum transportation costs. Campbell [1990b] develops a continuous approximation model of a freight carrier serving a fixed region with an increasing density of demand. The papers suggests that myopic strategy with limited capability to relocate is nearly optimal unless the terminal relocation costs are high. Klincewicz [1991] compares one-hub and two-hub exchanges (for the p-hub median problem) and the clustering techniques. Klincewicz [1992] use tabu search and greedy randomized adaptive search procedure to examine local optima and try to find better solutions. Kuby and Gray [1989] include stop-overs and feeders, equivalent to relay-points and spider-legs, for designing the network. Campbell [1994a] provides a concise survey of the work in the field of hub location as regards the objectives, proposed heuristics, their effectiveness and the size of the network. O'Kelly and Miller [1994] review research analytical papers and give brief empirical examples of eight different hub-and-spoke protocols. Skorin-Kapov and Skorin-Kapov [1994] provide a tabu search heuristic for the p-hub median problem. Campbell [1994b] presents integer programming formulations for the p-hub median problem, the uncapacitated hub location problem, p-hub center problem and hub covering problem. Formulations considering flow thresholds on spokes are also

considered. O'Kelly, Skorin-Kapov, and Skorin-Kapov [1995] use the knowledge from a known heuristic solution to strengthen the lower bounds for the hub location problem. Campbell [1996] defines a p-hub median problem and presents integer programming formulations. Two heuristics are proposed to find a solution for the single allocation p-hub median problem. Aykin [1996] proposes exact and heuristic solution procedures for the design of hub-and-spoke based distribution system. The decision involves whether a terminal may function as a hub. Skorin-Kapov, Skorin-Kapov, and O'kelly [1996] develop new mixed 0/1 linear formulations with tight programming relaxations. They found that in most cases the solutions to the LP relaxations were integer and in cases with fractional solutions, integrality was obtained by adding a partial set of integrality constraints. Jaillet, Song, and Yu [1996] propose flow based models for designing capacitated networks and routing policies. Mathematical programming based heuristic schemes are used. O'Kelly and Bryan [1996] analyze the sensitivity of the hub location with respect to the inter-hub discount factor. They also determine the optimal number of hubs as the fixed costs and interhub discount factors change.

7.2 Single Additional Hub Location using Enumeration

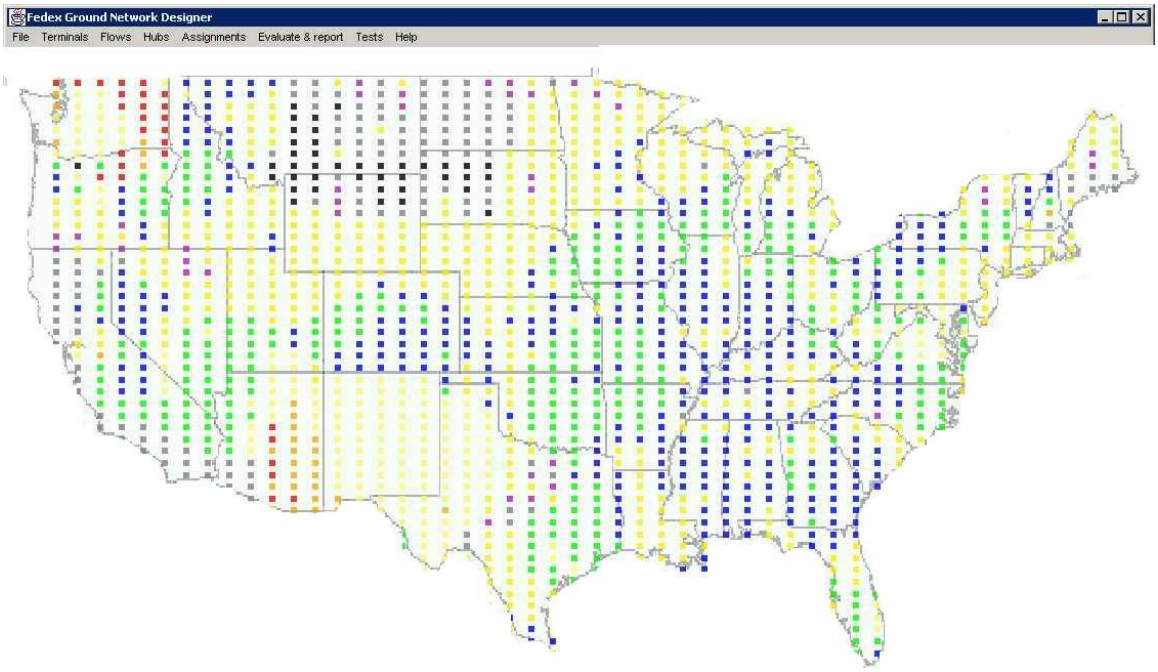


Figure 63: Iso-cost contour for an additional hub added to the FedEx Ground network

Hub location is an extremely difficult problem. Instead of focusing on finding optimal set of hub locations we directed our efforts to answering the following question: *Given a set of existing hub*

locations where should an additional hub be located?

We superimposed the map of USA with grids and sequentially located a hub at each grid point and generated the set of assignments using the heuristic mentioned in chapter 2 and estimated the total costs. The iso-cost contour plot of the costs associated with an additional hub at each grid point is shown in figure 63. Any grid point which is associated with a decrease in cost is denoted by a black hub. No change in costs is denoted by a grey hub. All grid points which are associated with an increase in cost are divided into seven groups colored from violet to red (in the same order as that of a rainbow) — violet denoting the least increase in cost and red denoting the most increase in cost.

Some hub locations are associated with no increase in costs because they do not affect the existing set of assignments. Most hub locations are associated with increase in costs either because they are the only feasible (and more expensive) hubs for certain terminals or because of the approximation of fractional trucks on the longhaul segments.

7.3 Remarks

Typical network design methodology involves an iterative procedure where after the hubs are perturbed the network is re-configured and re-evaluated. We have laid down the foundation which involves quick re-configuration of the hub-and-spoke network and accurate direct load planning.

However, the idea of perturbing the hub locations is not easy to formalize and is a research question that still needs to be addressed. However, more insight than that provided by this research needs to be gained before we can devise heuristic approaches to tackle hub location. This may also aid further research issues regarding adding additional hub(s) or deleting one (or more) from the existing set.

CHAPTER 8

CONCLUSIONS

Our key conclusions fall into four categories.

Network design: LTL load planners seem to do a good job of assigning terminals to hubs, but it is possible to automate this. The ability to redesign quickly is important in the presence of changes such as the recent change in "hours of service" [pers. com. Trussel, 2003]. Automating the hub-and-spoke network design may also provide valuable insights to redesign the network when it handles freight flows beyond the realm of any load planner's experience.

Dynamic Load Planning: Since the load planning problem has a huge computational overhead, traditionally the approach within the industry has been to provide a load plan for a certain period, which may extend a few weeks or to accommodate a freight seasonality. This load plan may only serve as a guideline to the terminal and hub load planners who then have to use their experience and expertise to build direct loads based on the actual demands.

Using the design technique we suggest in this research, for a network of the size of FedEx Ground we can generate a load plan within approximately 90 minutes (1.5 hours).[1]

Step in network design	Approximate time (minutes)
Generating the hub-and-spoke network	5
Network decomposition and generating sub-network MIPs for direct load optimization	60
Solving the MIP sub-problems	15
Generating solution for the entire network	5
Total Time	85

Currently, load planners at terminals/hubs look for opportunities to build direct loads, but they look from local perspective and so can miss globally superior solutions. Our methodology allows the rapid generation of globally-economical direct loads and so offer a practical way

[1]This estimate is if we generate the sub-network MIPs using a single processor but use 8 processors to solve the MIP sub-problems for minimum required direct load factors of 0.2 or higher.

of reducing system-wide costs. Furthermore, the significantly reduced computation time also makes possible the analysis of what-if scenarios.

Building direct loads: It is common wisdom within the industry that there are more opportunities to build terminal-to-hub direct loads rather than hub-to-terminal direct loads. Though these rules may be handy in absence of decision-support tools, they depend on freight patterns and may not always yield good load-plans. In fact, for the FedEx Ground data set about twice as many terminal-to-hub direct loads were sent as hub-to-terminal direct loads. This may be because the shippers may have been mostly freight producers, such as a manufacturing facility, sending product to customers, so that there were relatively large amounts of freight going to destination hubs, to deliver to many terminals served by that hub.

FedEx Ground is a package carrier and hence there is less diversity in the freight it handles as compared to other general LTL carriers. There may be other considerations that cannot be captured by programming. The hub being a consolidation center has greater mix of freight that may be compatible to be loaded onto a single trailer increasing trailer utilizations [pers. com. Rowe, 2003]. In spite of greater cost-saving opportunities to build direct loads at terminals rather than at hubs, hub-to-terminal direct loads may still be preferred.

Competing with Regional LTL Carriers: It is common practice in the industry to maximize trailer utilizations to reduce transportation costs. But the total operating cost also depends on the sorting costs. By maximizing trailer utilizations over *longer* distances the average savings per package dominate the sorting costs per package thereby reducing the overall operating costs. Over shorter distances, higher load factors may reduce transportation costs but any savings in transportation costs may be offset by sorting costs.

Higher load factor severely deteriorates service level over shorter distances. Consider a shipment that has to be sent within a region over a *short* distance. Requiring higher load factors may route the shipment from the origin to the destination via a hub. This hub may be located significantly off the direct route and outside the region, drastically increasing the transit time. Regional carriers satisfy customer expectations by providing lower transit times.

Our research suggests that sending some lightly-loaded trailers not only is cheaper but also greatly improves regional service – which is exactly where the national LTL carriers are facing stiff competition from regional LTL carriers.

We have shown how to coordinate the fine details of the movement of freight across networks that span a continent. Our methodology continues the by-now-familiar extension of automated decision-making that has been made possible by the revolution in computation and communications.

Bibliography

Mehmet O. Akyilmaz. An algorithmic Framework for Routing LTL Shipments. *Journal of Operational Research Society*, 45(5):529–538, 1994.

T. Aykin. On modeling scale economies in hub-and-spoke network designs. manuscript, 1996.

M. O. Ball, T. L. Magnanti, C. L. Monma, and G. L. Nemhauser, editors. *Network Routing*, volume 8 of *Handbooks in Operations Research and Management Science*. Elsevier, Amsterdam, 1995.

Robert E. Bixby, William Cook, Alan Cook, and Eva K. Lee. Computational Experience with Parallel Mixed Integer Programing in a Distributed Environment. *Annals of Operations Research*, 90:19–43, 1999.

J. W. Braklow, W. W. Graham, S. M. Hassler, K. E. Peck, and W. B. Powell. Interactive optimization improves service and performance for Yellow Freight System. *Interfaces*, 22(1):147–172, 1992.

J. F. Campbell. A Survey of Network Hub Location. *Studies in Locational Analysis*, 6:31–49, 1994a.

J. F. Campbell. Integer Programing Formulations of Discrete Hub Location Problems. *European Journal of Operations Journal*, 72:387–405, 1994b.

J. F. Campbell. Hub Location and the p-hub Location Problem. *Operations Research*, 44:923–935, 1996.

James F. Campbell. Freight Consolidation and Routing with Transportation Economies of Scale. *Transportation Research – B*, 24B:345–361, 1990a.

James F. Campbell. Locating transportation terminals to serve an expanding demand. *Transportation Research – B*, 24B:173–192, 1990b.

Alan Chalmers and Jonathan Tidmus. *Practical Parallel Processing*. International Thomson Press, London, WCIV 7AA, 1996. second edition.

V. Chvatal. *Linear Programing*. W. H. Freeman and Co., New York, 1992. second edition.

Jack J. Dongarra, Iain S. Duff, Danny C. Sorensen, and Henk A. van der Vorst. *Solving Linear Systems on Vector and Sahred Memory Computers*. Society of Industrial and Applied Mathematics, 3600 University City Science Center, Philadelphia, PA 19104-2688, 1993. second edition.

Jonathan Eckstein and Yossef Sheffi. Optimization of Group Line-Haul Operations for Motor Carriers Using Twin Trailers. *Transportation Research Record*, 1120:12–23, 1987.

C. J. Emerson, C. J. Grimm, and T. M. Corsi. The Advantage of Size in the U.S Trucking Industry: An Application of the Survivor Technique. *Journal of the Transportation Research Forum*, 112032 (2):369–378, 1992.

Geoffrey C. Fox, Roy D. Williams, and Paul C. Messina. *Parallel Computing Works*. Morgan Kaufmann Publishers, San Fransisco, CA 94104, 1994. second edition.

Richard L. Francis, Leon F. Jr. McGinnis, and John A. White. *Facility Layout and Location: An Analytical Approach*. Prentice Hall, Upper Saddle River, NJ 07458, 1992. second edition.

B. Gendron and T. G Crainic. Parallel Branch and Bound Algorithms: Survey and Synthesis. *Operations Research*, 42:1042–66, 1994.

Patrick Jaillet, Gao Song, and Gang Yu. Airline network design and hub location problems. *Location Science*, 4(3):195–212, 1996.

J. G. Klincewicz. Heuristics for the p-hub location problem. *European Journal of Operations Journal*, 53:25–37, 1991.

J. G. Klincewicz. Avoiding local optima in the p-hub location problem using Tabu search and GRASP. *Annals of Opns. Res.*, 40:121–132, 1992.

M. J. Kuby and R. G. Gray. The Hub Network Design Problem with Stopovers and Feeders: The Case of Federal Express. *Transportation Research–A*, 27A(1):12–29, 1989.

Bruce W. Lamar and Yossef Sheffi. An implicit Enumeration Method for LTL Network Design. *Transportation Research Record*, 1120:1–11, 1987.

C. John Langley. Conversation with author. Atlanta GA., 14 November 2003.

Eva K. Lee. A Linear-Programming Based Parallel Cutting Plane Algorithm for Mixed Integer Programming Problem. *Proceeding for the Third Scandinavian Workshop on Linear-Programming*, pages 22–31, 1999.

Eva K. Lee. Branch-and-Bound Methods. In *Handbook of Applied Optimization*, chapter 3. Oxford University Press, 2001. Invited book chapter. Honorable mention by the Associate of American Publishers' (AAP) Outstanding Professional and Scholarly Titles of 2002 in Computer Science.

Eva K. Lee. Generating Cutting Planes for Mixed Integer Programming Problems in a Parallel Computing Environment. *INFORMS Journal on Computing*, 16:1–28, 2004.

J. M. Y. Leung, T. L. Magnanti, and V. Singhal. Routing in Point-to-Point Delivery Systems: Formulations and Solution Heuristics. *Transportation Science*, 24(4):245–260, 1990.

Cheng-Chang Lin. The Freight Routing Problem of Time-Definite Freight Delivery Common Carriers. *Transportation Research – B*, 35:525–547, 2001.

Starr B. McMullen and Hiroshi Tanaka. An Econometric Analysis of Differences Between Motor Carriers: Implications for Market Structure. *Quarterly Journal of Business and Economics*, 34 (4):16–29, 1995.

Sev McMurtry. Conversation with author. Pittsburgh, PA., 27 June 2000.

M. E O'Kelly. A quadratic integer program for the location of interacting hub facilities. *European Journal of Operations Journal*, 32:393–404, 1987.

M. E. O'Kelly and D. Bryan. Hub network design with single and multiple allocation: a computational study. *Location Science*, 4(3):125–138, 1996.

M. E. O'Kelly and H. J. Miller. The Hub Network Design Problem. *Journal of Transport Geography*, 2:31–40, 1994.

M. E. O'Kelly, D. Skorin-Kapov, and J. Skorin-Kapov. Lower bounds for the hub location problem. *Management Science*, 41(4):713–721, 1995.

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

Warren B. Powell and Ioasnnis A. Koskosidis. Shipment routing algorithms with tree constraints. *Transportation Science*, 26(3):230–245, 1992.

Warren B. Powell and Y. Sheffi. Design and Implementation of an Interactive Optimization System for Network Design in the Motor Carrier Industry. *Operations Research*, 37:12–29, 1989.

James Rowe. Conversation with author. Atlanta GA., 19 August 2003.

J. Roy and L. Delorme. NETPLAN: a network optimization model for tactical planning in the less-then-truckload motor-carrier industry. *INFOR*, 27(1):22–35, 1989.

D. Skorin-Kapov and J. Skorin-Kapov. On Tabu search for the location of interacting hub facilities. *European Journal of Operations Journal*, 73:502–509, 1994.

D. Skorin-Kapov, J. Skorin-Kapov, and M. E. O'kelly. Tight Linear Programming Relaxations of Uncapacitated p-hub Median Problems. *European Journal of Operations Journal*, 94:582–593, 1996.

Marc Snir, Steve W. Otto, Steven Huss-Lederman, David W. Walker, and Jack Dongarra. *MPI: The Complete Reference*. The MIT Press, Massachusetts Institute of Technology, Cambridge, MA 02142, 1996.

Ross M. Starr and Maxwell B. Stinchcombe. Efficient Transportation Planning and Natural Monopoly in the Airline Industry: An Economic Analysis of Hub-Spoke and Related System. Discussion Paper 92–25, University of California, San Diego, Department of Economics, June 1992.

Janet M. Thomas and Scott J. Callan. Constant Returns to Scale in the Post-Deregulatory Period; The Case of Specialized Motor Carriers. *Logistics and Transportation Review*, 25(3):271–288, 1989.

Teresa Trussel. Conversation with author. Atlanta GA., 12 September 2002.

Teresa Trussel. Conversation with author. Atlanta GA., 5 November 2003.

Vijay Vazirani. *Approximation Algorithms*. Springer, 2001.

Lawrence A. Wolsey. *Integer Programming*. John Wiley & Sons Inc., 605 Third Avenue, New York, N.Y. 10158-0012, 1998. second edition.

D. D. Wyckoff. *Organizational Formality and Performance in the Motor-Carrier Industry*. Lexington Books, D.D. Heath and Company, Lexington, MA, 1974.

VITA

Devang Bhalchandra Dave was born on February 14, 1975 in Mumbai, formerly Bombay, India. He graduated with Bachelor of Engineering with Honors, in Mechanical Engineering, from Sardar Patel College of Engineering, affiliated to the University of Mumbai, in 1996. During his year of work experience as a maintenance engineer at National Organic Chemical Industries Limited (NOCIL), a petrochemical complex in India, he got exposed to and interested in Industrial Engineering and came to Georgia Tech, Atlanta to pursue a Master of Science in Industrial Engineering. He started work as a Graduate Research Assistant under the guidance of Dr. John J. Bartholdi, III during his Masters program. Liking what he was doing and doing what he was liking he continued to stay at Georgia Tech where he recieved his Doctor of Philosophy on 2004. Besides research, Devang aspires to be an Ironman sometime in his lifetime.