Southwest Region University Transportation Center

Real-Time Information for Improved Efficiency of Commercial Vehicle Operations

SWUTC/98/60031-1

Center for Transportation Research
University of Texas at Austin
3208 Red River, Suite 200
Austin, Texas 78705-2650

1. Report No.	2. Government Accession No.	3
SWUTC/98/60031-1		

4. Title and Subtitle	5. Report Date
Real-Time Information for Improved Efficiency of Commercial Vehicle Operations	July 1998
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Amelia Clare Regan, Hani S. Mahmassani and Patrick Jaillet	Research Report 60031-1
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Center for Transportation Research University of Texas at Austin 3208 Red River, Suite 200 Austin, Texas 78705-2650	11. Contract or Grant No. 0079
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Southwest Region University Transportation Center Texas Transportation Institute The Texas A&M University System College Station, Texas 77843-3135	14. Sponsoring Agency Code

15. Supplementary Notes

Supported by a grant from the Office of the Governor of the State of Texas, Energy Office

16. Abstract

Intelligent Transportation Systems (ITS) harness advanced communications and computation technologies in order to make transportation systems more efficient. This work is concerned with the application of ITS to commercial vehicle operations and freight mobility; it identifies and investigates potential uses of real-time information for the efficient management of carrier operations. In truckload and less-than-truckload operations, carriers typically know only a portion of the loads that must be moved more than a few hours before the moves must take place. Therefore, the assignment of an available driver to a load takes place in real-time or shortly after the request is received. The load acceptance decision made by a carrier must also be executed in real-time, and may have a significant impact on the carrier's ability to accept other loads requested in the near future. In this context, vehicle to load assignments as well as the sequence in which loads are to be served may be revisited as demands unfold and traffic network conditions change. Because of the speed with which decisions must be made, the number of possible choices and the fact that the system is changing dynamically and often, unpredictably, locally oriented decision rules offer a promising alternative to approaches seeking global optimality or those which take into account long term or forecast information.

17. Key Words	18. Distribution Statement
Intelligent Transportation Systems, Real-Time Information, Load Acceptance	No Restrictions. This document is available to the public through NTIS: National Technical Information Service 5285 Port Royal Road Springfield, Virginia 22161

19. Security Classif.(of this report)	20. Security Classif.(of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	250	

REAL-TIME INFORMATION FOR IMPROVED EFFICIENCY OF COMMERCIAL VEHICLE OPERATIONS

by

Amelia Clare Regan
Hani S. Mahmassani
and
Patrick Jaillet

SWUTC/98/60031-1

Conducted for the

Southwest Region University Transportation Center
Texas Transportation Institute
Texas A&M University System
College Station, Texas 77843-3135

Prepared by the

Center for Transportation Research
The University of Texas at Austin
3208 Red River, Suite 200
Austin, Texas 78705-2650

July 1998

ACKNOWLEDGMENT

This publication was developed as part of the University Transportation Centers Program which is funded 50% in oil overcharge funds from the Stripper Well settlements as provided by the Texas State Energy Conservation Office and approved by the U.S. Department of Energy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

EXECUTIVE SUMMARY

Commercial vehicle operations consume a vast quantity of economic and environmental resources. Profit margins in the trucking industry are thin, typically less than six percent and often as low as one percent of company revenues. Improvements in operating efficiency lead directly to service improvements and increases in carrier profitability and may result in reduced prices for shippers and, ultimately, in reduced costs to consumers. Transportation cost represent as much as twenty percent of consumer purchases; even a small reduction in these costs can result in significant savings. In addition to improving the operational efficiency and hence profitability and customer responsiveness of operations, improving the energy efficiency of commercial vehicle operations can impact overall energy consumption.

The U.S. Department of Transportation estimated that the cost of domestic, intercity freight transportation in 1995 was almost 220 billion dollars, and the combined cost of intercity and local trucking operations was 348 billion or approximately 5 percent of the 1995 Gross Domestic Product (GDP). In addition, it is estimated that motor vehicle fuel purchases in the same year accounted for 15 percent of common carrier operating expenses, or about 52.2 billion dollars nationally. Advances in Intelligent Transportation Systems (ITS) technologies for commercial vehicle operations offer opportunities for reducing the overall resource consumption of these operations.

The use of automatic vehicle location, real-time communication technologies, along with on-board and dispatch center GIS and database management systems offer significant opportunities for improving the efficiency of commercial vehicle operations. Taking full advantage of these technologies requires the development of tools specifically tailored to information intensive operations. This study has developed a family of operational tools specifically tailored to such operations. These include computer-based dynamic load acceptance and load assignment methods for truckload trucking operations.

An extensive simulation testbed is described for the performance evaluation of these methods, under various demand pattern assumptions and informational scenarios. Extensive numerical tests suggest that the dynamic assignment and dispatching methods developed in this study will perform well, both with respect to customer service and cost measures.

ABSTRACT

Intelligent Transportation Systems (ITS) harness advanced communications and computation technologies in order to make transportation systems more efficient. This work is concerned with the application of ITS to commercial vehicle operations and freight mobility; it identifies and investigates potential uses of real-time information for the efficient management of carrier operations.

In truckload and less-than-truckload operations, carriers typically know only a portion of the loads that must be moved more than a few hours before the moves must take place. Therefore, the assignment of an available driver to a load takes place in real-time or shortly after the request is received. The load acceptance decision made by a carrier must also be executed in real-time, and may have a significant impact on the carrier's ability to accept other loads requested in the near future. In this context vehicle to load assignments as well as the sequence in which loads are to be served may be revisited as demands unfold and traffic network conditions change. Because of the speed with which decisions must be made, the number of possible choices and the fact that the system is changing dynamically and often, unpredictably, locally oriented decision rules offer a promising alternative to approaches seeking global optimality or those which take into account long term or forecast information.

The main hypotheses examined are, that real-time information on vehicle locations and demands can increase the efficiency of carrier fleet operations with respect to measures of trucking company profitability and responsiveness to customer requests, and, that real-time operational strategies perform well, compared to those requiring less real-time information, under certain conditions with respect to fleet size, level of demand and service deadlines. Operational strategies which take advantage of real-time information and, which include methods to perform load acceptance, assignment and re-assignment are examined both analytically, and in simulation framework developed to test these and related strategies under a variety of operating assumptions. Quantitative estimates of the benefits of real-time information for vehicle assignment and routing decisions for trucking operations are developed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

PROBLEM STATEMENT

This work identifies and investigates potential uses of real-time information for the efficient management of carrier operations. In truckload and less-than-truckload (LTL) operations, carriers typically know only a portion of the loads that must be moved more than a few hours before the loads are to be moved. The assignment of an available driver to a load takes place in real-time or shortly after the request is received. The load acceptance decision made by a carrier must also be executed in real-time, and may have a significant impact on the carrier's ability to accept other loads requested in the near future. This research explores ways to make "good" assignment, and ultimately load acceptance decisions, that lead to overall cost effective operations, but rely on local (current) rather than long term or forecast information.

In this context, vehicle to load assignments as well as the sequence in which loads are to be served may be revisited as demands unfold and traffic network conditions change. Because of the speed with which decisions must be made, the number of possible choices, and the fact that the system is in a constant state of flux, locally oriented decision rules offer a promising alternative to approaches seeking global optimality. These local decision rules are primary to this research and can take several forms: first, the decision time frame may be local, so that only current information is employed; second, the decision region may include a "local" or logical subset of vehicles or of demand locations; in addition, if a current solution is to be modified rather than completely re-generated when new demands are accepted, service is completed, or changes in driver and vehicle availability and traffic network occur, then the decision region may be limited to assignments which differ in small ways from the current assignments. In summary, such decision rules are local with respect to one or more of the following: temporal, spatial, or algorithmic considerations.

MOTIVATION

Commercial vehicle operations consume a vast quantity of economic and environmental resources. Profit margins in the trucking industry are very thin, typically less than five percent and often as low as one percent of company revenues [Association of American Railroads, 1992]. Improvements in operating efficiency lead directly to increases in carrier profitability and may result in reduced prices for shippers and, ultimately, in reduced costs to consumers. Transportation costs represent as much as twenty percent of consumer purchases; even a small

reduction in these costs can result in significant savings [Sampson et al, 1985]. In addition to the benefits of improving the operational efficiency and hence profitability and customer responsiveness of operations, improving the energy efficiency of commercial vehicle operations can impact overall energy consumption. In trucking operations alone, a reduction in overall travel of even a few percentage points would represent a significant savings to both suppliers and consumers of such services. The U.S. Department of Transportation estimated that the cost of domestic, intercity freight transportation in 1991 was 167 billion dollars, and the combined cost of intercity and local trucking operations was 278 billion or 4.9 percent of the 1991 Gross National Product (GNP). In addition, it is estimated that motor vehicle fuel purchases in the same year accounted for 7.9 percent of common carrier operating expenses, or about 8.7 billion dollars nationally [Schmitt and Feinberg, 1994]. Advances in Intelligent Transportation Systems (ITS) technologies for commercial vehicle operations offer opportunities for reducing the overall resource consumption of these operations. Telecommunications and information technologies provide opportunities for using real-time information to enhance the productivity, performance, and energy efficiency of the commercial transportation sector. Achieving the benefits of real-time information requires the development of fleet operating strategies, including vehicle assignment and dispatching rules with more flexibility than those currently in use, along with suitable decision support methodologies.

The area of vehicle routing and scheduling, including dynamic vehicle allocation and load assignment models, has evolved rapidly in the past few years, both in terms of underlying mathematical basis and actual commercial software tools. While these approaches may well be adaptable to operations under real-time information availability, they are at present unable to take full advantage of such information because their underlying formulations do not recognize possible decisions that are only meaningful under real-time information.

There is currently little methodology in the literature intended specifically for truckload or other surface carrier operations under the kind of real-time information possible with emerging technologies. This lack of methodological development applies to analysis of carrier operations to evaluate the effectiveness of real-time information, as well as to actual tools that could be used by carriers to take advantage of such information. However, this work is a part of a rapidly growing area of exploration. The field of logistics, driven by the increase in real-time information availability in transportation and supply chain management systems, has witnessed explosive growth in the past few years. Interest in the development of dynamic models of fleet operations and of fleet management systems which are responsive to changes in demand, traffic network

and other conditions is emerging in many industries and for a wide variety of applications [Desrosiers et al 1995, Powell, Jaillet & Odoni, 1995].

RESEARCH CONTEXT: TRUCKLOAD CARRIER FLEET OPERATIONS

The context for this research is truckload carrier fleet operations in which each assignment involves a vehicle moving a single load from the load origin to the load destination. We expect that the results of this inquiry will lend insight to other fleet management problems. The problem studied involves the management of a fleet of vehicles over a typically wide geographic region over time. It is assumed that the vehicle fleet is under the control of a central authority, the dispatcher. Some carrier companies own the vehicles in their fleet, some employ the services of owner-operator drivers, and others maintain a fleet of both company drivers and owner-operators. We ignore these distinctions, which are not essential for the operational strategies considered here, and assume that a central authority has the responsibility for directing the movements of vehicles in their fleet. Several equipment types may be available: tank vehicles for transporting petroleum and other chemical products, refrigerated units for carrying perishable items, flat bed vehicles for carrying, among others, irregularly shaped items and lumber, and, standard or drop frame trailers. In this analysis it is assumed equipment types are homogeneous or at least substitutable. This is without loss of generality and in keeping with the goal of exploring decision rules that are "locally oriented". We further ignore the additional complexity of single and double trailers and of loads which are carried in shipper owned trailers. The assumption is that a driver and vehicle combination includes a trailer that is loaded and unloaded at customer sites. In addition, this research does not take into account federal regulations concerning, for example, the length of time a driver may be on duty or driving. That level of detail, while essential for the successful development of implementable fleet management systems, is beyond the scope of this research.

Once a vehicle is loaded at a customer site, it cannot provide service to another customer until unloaded, typically at the destination point of the load. An exception to this assumption would be in the case of a load swap where two or more drivers meet at some point en-route and either swap trailers or unload a whole load. Those scenarios are not explicitly included in this investigation. Scenarios in which all vehicles are equipped with a continuous Automatic Vehicle Location (AVL) system are examined. All vehicles are also equipped with continuous two-way communications devices which allow driver to dispatcher communication to take place within a short period of time. These technology equipped operations are compared with operations in which real-time communication and vehicle location and status updates are not possible. Further,

in the technology equipped scenarios it is assumed that the dispatch center(s) have a means to display the current locations of all vehicles, typically with a digital map and a GIS (Geographic Information System) interface.

Demands for service are highly stochastic, with carriers typically knowing less than half of the loads to be moved more than a few hours before they are to be picked up [Powell, 1988]. Other aspects of carrier fleet operations may also have a high degree of variability. Travel times may vary significantly, due to unexpected congestion, road closures, weather or equipment failures. The time spent at loading and unloading points may also vary. This variability is not explicitly included in formulations of the problem presented in this research. Rather than work with stochastic representations of travel and loading and unloading times explicitly (i.e. by estimating a probability distribution for these times), these times are considered known. However, when actual conditions vary from those expected, the real-time dispatching systems we envision are designed to react to the unexpected. Of significant interest to this research is the extent to which different assignment strategies are able to react to changes and the effectiveness of their response as changes occur.

The carrier operations planning process can be viewed in two parts. First there is the need to manage the supply of vehicles and drivers to provide timely service to customers. The supply problem includes the assignment of drivers to known loads, reassignment as changes in demands, driver and vehicle availability and traffic network conditions occur, and the repositioning of idle vehicles in anticipation of future demands. Related to this is the need to effectively manage customer requests for service. The load management problem includes both the decision to accept (or reject) requested loads for service and the active solicitation of loads in regions in which the supply of vehicles to provide service currently exceeds or will soon exceed demands. The context of truckload trucking and other dynamic fleet operations is such that it is not possible to serve all requested demands at all times [Powell, Jaillet & Odoni, 1995]. It may be necessary to turn down requests for service if time window, regional or system wide capacity constraints cannot be met. In general, carriers have the ability to refuse customer demands with little risk of loss of customer good will. This increases the importance of making careful load acceptance and solicitation decisions. When some loads are likely to be refused, the problem of effectively choosing the subset of loads to serve has significant effect on the profitability and efficiency of an operation.

The load acceptance and driver-to-load assignment processes may be tightly or loosely coupled, or tackled separately. In general they are separated; the load acceptance decision must be made when a request arrives while, in principle, loads may be assigned to a driver at any time

after the request is received. Shippers call a carrier requesting that a vehicle be available at a pickup location on a specific day, at a specific time, to carry a load to a specific destination. The carrier must decide quickly whether to accept a request to move the load. Advances in computing technology and scheduling and assignment heuristics offer opportunities for coupling the acceptance and assignment process or at least using information about the estimated cost and feasibility of providing service in the load acceptance decision. Assuming the carrier has accepted the load, a vehicle (and driver) moves the load from its origin to its destination and unloads, at which time the carrier must decide what to do with the vehicle. It may be assigned to another load, repositioned to another region in anticipation of future loads in that region, or held in anticipation of future loads in the destination region. A carrier is paid a revenue proportional to the distance of the length of the haul. Loads may have firm or somewhat flexible pickup deadlines. A carrier may decline a load, but it may not accept a load unless it can meet the agreed upon deadlines.

Load Acquisition and Acceptance

Load acquisition may be active or passive. Carriers may simply wait for shippers to call with requests for service or they may identify regions in which supply exceeds demand or in which excess capacity will be available in the near future and actively seek loads originating in regions with excess capacity. Once a shipper has requested that a load be moved, the carrier must decide whether or not to accept the load. This decision may be based upon feasibility only. That is, a carrier may choose to accept all feasible loads, in which case the carrier must have a way to assure that the fleet can move the load, along with previously accepted loads, within the agreed upon time. Or, the decision may be based on an estimate of the cost of providing service and on expectations about (near) future requests. An estimate of the revenue potential of the load may be used both to make the load acceptance decision, and in some cases to trigger price negotiations. Ideally, the current or near-term state of the system with respect to the availability and location of vehicles should be used to update revenue estimates as changes in the system occur.

The profit realized for each loaded movement is highly dependent upon driver proximity and availability at the time the load is moved. In addition, each movement affects the ability of individual drivers or a fleet of drivers to respond to near-term demands for service. Many companies make an effort to predict where and when excess supplies (vehicles and drivers) will be available; some employ load acquisition strategies which discount empty movements in regions that would otherwise require excessive dead-head movements. However, more

comprehensive pricing and load acquisition strategy which takes the current and predicted near term state of the system into account could lead to increased productivity. This could be accomplished in many ways. For example, demands for services could be forecast over time; based on these forecasts the surpluses and deficits in each region or traffic lane could be estimated. These surpluses and deficits could be used to calculate the estimated return on each load, adjusted with the expectation that loaded vehicles moving in particular lanes will be needed at their destination locations or, once empty, repositioned to other locations. These revenue projections could be used internally to identify regions or traffic lanes that should be targeted for aggressive load acquisition and could also be used to adjust the prices charged shippers. Unfolding demands could be compared with forecast demands and dispatchers and managers made aware of unexpected fluctuations. Powell [1985] has suggested several ways to forecast demands and has employed these forecasts both for solving a rolling horizon stochastic programming formulation of the dynamic vehicle allocation (driver to load and region assignment) problem and to estimate the marginal cost (and hence expected return) for movements. Powell explored marginal cost estimation which relies on solving either a deterministic or stochastic formulation of the vehicle allocation problem and found that although the output from the stochastic (and non-linear) version of the model produced useful results, that these were difficult to extract from the model without re-optimizing under many different scenarios. While this and related approaches may indeed lead to the development of valuable insights for carrier companies, it will likely not lead to an approach that can be employed in real-time.

Issues in Load Assignment and Re-assignment

As discussed in the previous section, when a request for service arrives, carrier fleet managers decide whether or not to serve the load. This decision may be made immediately, or within a very short time after the request for service is received. After acceptance, the load is either immediately assigned to a particular vehicle or it is sent to a pool of accepted but unassigned demands for future assignment. The question of how best to handle the tradeoffs between immediate assignment to a vehicle and assignment to a pool is of significant importance to this research. Issues addressed arise in scheduling and assignment in many different fleet-management contexts as well as other distributed or fixed location service systems. Two scenarios, one in which loads are held in a large common pool of accepted demands until assignment to a particular vehicle close to the time at which service is scheduled to begin, and another, in which most accepted demands are assigned to a particular vehicle's queue offer different advantages. The ability to make confident load acceptance decisions--at least when

utilization rates are high--requires that loads be assigned, at least temporarily, to an individual driver or to a small pool of drivers. The ability to create even very short "routes" (sequences of loads to be served) can lead to significant efficiencies in terms of empty distances traveled.

It is possible to achieve the flexibility of a pooled queue of demands while at the same time achieving the economies of generating short "routes" of assigned loads. In these instances, a priori tours that include all accepted service requests are constructed but may be modified as changes occur. The issue of how to best modify assignments in a real-time environment is central to this research. Real-time assignment strategies which rely on route insertion can be extremely efficient. Feasible insertion points can be identified within seconds and all feasible insertion points can be evaluated for efficiency within a comparable period of time. However, considering the re-ordering of routes at the same time as the addition of new loads is computationally expensive; the identification of (cost effective) load swaps between vehicles can be even more so. These assignment techniques lead, in most cases to more efficient assignments, but, rather than explore all alternatives, either the feasible set, or the most promising subset of options should be identified a priori. Decision rules which are local in a temporal, spatial, or algorithmic sense are amenable to the realities of real-time decision processes. The decision time frame may be local, so that only current information is employed. The decision region may include a "local" or logical subset of vehicles or of demand locations generally chosen by geographic location. If a current solution is to be modified rather than completely re-generated when new demands are accepted, service is completed, or changes driver and vehicle availability and traffic network occur, then the decision region is limited to local decisions in the region of the current solution. Each of these techniques results in the identification of small versions of the problem; in some cases these problem instances may be solved optimally (with respect to the chosen criteria) in real-time.

RESEARCH OBJECTIVES

Primary Objectives

The primary objectives of this research are to:

1. State, formulate and analyze the driver assignment (or dynamic vehicle allocation and routing) problem in a way that explicitly takes real-time information on vehicle locations and demands into account.
2. Develop operations research methodologies to assist with dispatching, load acceptance, and dynamic pricing strategies and to test these methodologies under the assumption of the availability of real-time information on vehicle locations and demands. These

methodologies employ real-time information to enhance system productivity and performance under a variety of operating assumptions.

3. Develop a simulation framework to analyze carrier fleet operations under real-time information and to evaluate the effectiveness of strategies developed.
4. Provide quantitative estimates of the benefits of real-time information for vehicle assignment and routing decisions for trucking operations.

Main Hypotheses

Related to the objectives mentioned above, two main hypotheses are tested, both through analytical and simulation investigation. These are:

1) Real-time information on vehicle locations and demands can increase the efficiency of carrier fleet operations with respect to measures of trucking company profitability and responsiveness to customer requests or desires.
2) Real-time assignment rules perform well, with respect to those requiring less real-time information, under certain conditions with respect to fleet size, level of demand and pickup deadlines.

In investigating these hypotheses, the following questions are examined:

- How can operations take advantage of real-time information on vehicle locations and demands?
- How do assignments triggered by changes in the system compare to assignments triggered by the passing of time or of an accumulation of loads or idle vehicles?
- How do local assignment rules compare to assignments generated with the benefit of perfect hindsight?
- Which local assignment rules appear to perform best and under what conditions do they exhibit relative advantages?

RESEARCH APPROACH

The focus of this research is operational strategies that require varying degrees of real-time information and communication. The performance of these strategies is examined in a simulation framework. Prior to that investigation, analytic models of the dynamic vehicle allocation and routing problem are developed; these explore several load assignment strategies and focus on dynamic dispatching strategies that are outside the norms of typical carrier fleet operations. In particular, the diversion of en-route vehicles is examined. This strategy involves diverting a vehicle en-route to a pickup location to make an immediate pickup of a more time-sensitive load, or of a load that (when sequenced first) will improve the efficiency of the vehicle's travel route.

Strategies allowing en-route diversion are examined in detail, beginning with the operations of a single vehicle. Extending this analysis to vehicle fleets, the increase in the ability of a fleet to respond to time-sensitive demands under real-time information is estimated, again with the help of simplifying assumptions. Following a similar line of reasoning, a model of carrier fleet operations as a distributed queueing system is introduced in order to examine how congestion effects the ability of a fleet to respond quickly to requests for service and, to identify the tradeoffs between pooled demands and those assigned to individual vehicle "queues".

The simulation framework allows for evaluation of the expected performance of assignment and load acceptance strategies under a variety of conditions. Emphasis is placed on examining rather small instances of the problems, under idealized conditions, in an effort to gain insight into the relative merits of flexible assignment strategies and of the benefits to carrier fleet operations of real-time information on vehicle locations, demands and traffic network conditions. "Real-time" operational strategies, consisting of two load acceptance strategies and four assignment strategies are compared to five less information intensive "base case" operational strategies in which the same simple load acceptance policy is followed. In addition, for a single vehicle, the performance of one of the real-time assignment strategies is compared to that of a perfect hindsight solution. Simulation experiments examine the performance of these operational strategies in scenarios in which fleet sizes vary, demand intensities range from one in which five to fifty percent of requests must be turned away to one in which vehicles spend nearly half their time idle, and service requests may or may not have associated pickup deadlines. The long run expected performance of these strategies in terms of several measures of effectiveness is estimated.

RESEARCH SCOPE

A key assumption in this work is that requests for service arrive over time and that assignment decisions are made on a continuous basis as outcomes are observed. No a priori information on the location or timing of future service requests is considered. Two of the "base case" strategies examined, which determine assignments using a classical bipartite assignment approach, allow demands to accumulate in a pool prior to assignment. In none of the strategies examined are demands forecast. This assumption is made with the understanding that not all successful carrier fleet operations provide service to immediate requests, and, that even in an operation providing (immediate) demand responsive service, some fraction of the loads to be moved would be known in advance. The eventual goal of this work is the inclusion of heuristics identified in an overall dispatching system capable of generating both immediate and longer term

operating plans. The research in this report, however, is limited in scope to the evaluation of dynamic dispatching heuristics. This research does not explicitly address the issue of how to best incorporate the assignment of service requests arriving dynamically over time with demands known well in advance. Such investigation, which would examine the performance of local assignment heuristics relative to and in conjunction with "global" optimization systems, in which solutions are generated as changes in the system occur, is a topic of future and continuing research.

Closely related to the issue of scope is the question of how to best evaluate the performance of dispatching strategies. The choice of evaluation criteria is more difficult in a dynamic operation than in a static one. Several benchmark solutions are examined. One, which provides an upper bound on system efficiency involves the assignment of loads to available vehicles in the order in which they arrive, without regard to current vehicle locations. Another, which provides a lower bound on the distance traveled to provided service to a fixed set of loads, requires the development of solutions with the benefit of perfect hindsight. Because cost or feasibility based load acceptance rules confound the issue of comparing real-time solutions to perfect hindsight solutions, a comparison is made between the long run average cost of serving set of randomly generated loads and the corresponding long run average cost when the real-time assignment strategies are applied. The number of loads in the sets are equal to the average number of loads served per week in the real-time assignment strategies. Chapter 3 provides a discussion of criteria used to evaluate the performance of assignment strategies. These are further defined, in the context of evaluation of the relative performance of strategies examined in simulation experiments, in Chapter 5. In some cases, these criteria are linked to those used to make assignment decisions; in others, a lower level proxy for a higher level objective is used in the decision process.

REPORT ORGANIZATION

This report is organized in the following manner. Following this introductory chapter, chapter 2 presents a review of related work in the literature and covers other necessary background information. The second chapter begins with a brief discussion of the growth of the carrier fleet industry and its current state, especially with regard to the nature of competition and demands for increasingly responsive customer service. This is followed by an introduction of the relevant technologies involved in the application of intelligent transportation systems (ITS) advances to commercial vehicle operations (CVO), and discussion of the "state of the art" applications of these technologies in commercial fleet management. CVO applications of ITS technologies are

introduced and briefly reviewed, and the ITS CVO strategic program plan of ITS America presented. Finally, a review of the literature most relevant to this research is discussed and the work placed in context.

Chapter 3 introduces the conceptual and theoretical framework for the analysis of dynamic dispatching strategies for carrier fleet operations under real-time information. The problem is defined in finer detail than in the introductory chapter and assumptions are made explicit. Operational strategies examined in simulation experiments are presented formally, along with mathematical formulations of problems investigated.

Chapter 4 discusses analytic investigations of carrier fleet operations under real-time information, lending insight to the spatial and temporal considerations of carrier fleet operations and dynamic dispatching systems. A strategy of diverting an en-route vehicle to make an immediate pick-up of a more time-sensitive load, or of a load that when sequenced first will improve the efficiency of the vehicle's route is introduced; the increase in the ability of a fleet to respond to time-sensitive demands under real-time information is estimated, with the help of simplifying assumptions. The findings suggest that allowing continuous updates on the location and status of all vehicles in the fleet, coupled with flexible assignment strategies can significantly increase the ability of the fleet of vehicles to respond immediately to new requests. Related to this, carrier fleet operations, a distributed service system, are modeled as an M/G/k queue. A model examining the extent to which congestion affects the ability of a fleet to respond quickly to requests for service is introduced. The model supplies an upper bound on the average wait time for service under varying congestion levels. While the upper bound is quite loose in some cases, it does provide insight into methods for estimating congestion levels that are operationally attractive.

Chapter 5 contains the experimental design followed; a map of the simulation experiments is provided, along with a detailed description of the nine assignment strategies and three load acceptance rules compared and the methods of comparison used. The key factors in the experiments are: the operational strategy (combination of load acceptance rule, and assignment strategy) selected, fleet size, demand intensity and, in the real-time operational strategies, the presence and distribution of deadlines for pickup.

The presentation and analysis of simulation results is the topic of chapter 6. In this chapter results of the simulation experiments comparing the five "base case" and four "real-time" assignment strategies are discussed. The operating environment is by definition dynamic in all of the scenarios examined. Service requests arise over time; the load acceptance decision must be executed when the request for service is received; current load acceptance decisions impact the

ability of the fleet to accept future assignments. The "base case" strategies are intended to represent operations that are less information intensive while the "real-time" strategies rely on continuous driver to dispatcher communication and location and status information for all vehicles in the fleet. Operational strategies are compared over a set of system profitability measures, as well as in terms of their ability to provide satisfactory service to customers. The primary evaluation criteria are: the length of the average empty distance driven between loaded moves, the average and associated variability of wait time for service, and, an estimate of the operating profit generated per driver per week under a set of profit model assumptions. Results indicate that the real-time operational strategies perform well, when compared to the less information intensive cases when evaluated with respect to both profitability and customer service measures under realistic demand scenarios. Under very high demand the base cases fare well. In fact, one of the least information intensive base cases, a purely greedy assignment rule, provides the best performance with respect to most criteria when the system is over capacitated. In more moderate demand environments the real-time strategies perform much better with respect to the criterion of customer wait time for service, and perform well with respect to profitability; the base case with the best performance is a quasi real-time strategy which requires continuous two-way driver to dispatcher communication.

The final chapter summarizes the results of the report, provides conclusions and makes recommendations for continuing research.

CHAPTER 2 BACKGROUND REVIEW

INTRODUCTION

Chapter 2 reviews background material relevant to the modeling of carrier fleet operations under real-time information. Chapter 2 begins with a brief discussion of the growth of the carrier fleet industry and of the current state of the industry, especially with regard to the nature of competition and demands for increasingly responsive customer service. It expands to introduce some of the intelligent transportation systems (ITS) technologies relevant to commercial vehicle operations (CVO).

Chapter 2 also presents a review of the operations research literature that is most relevant to this research. Placing this research in context has posed a significant challenge. Obvious connections exist, for example, to the standard vehicle routing literature and to research on time-constrained vehicle routing. In most cases, however, these now comprehensive bodies of literature focus on problems in which the locations and magnitude of demands are known. In addition, most vehicle routing problems involve the assignment of a fairly large number of customers to a single vehicle route, while the quasi "routes" assigned in this application (truckload trucking operations) tend to be short, containing at most three to five customers, and in some cases containing only one or two. There is a natural affinity too, to the stochastic routing problems that have also been addressed in the literature [for example, Jaillet and Odoni, 1988; Bertsimas, Jaillet & Odoni, 1990] but in general these have been two-stage optimization problems in which the goal is to produce a robust a priori solution to a stochastic problem. Most of that literature has as its focus the generation of solutions that are optimal in the expected sense. The issue of how the system reacts as the demands unfold is closer to the core of this research and is rarely addressed in the literature. Powell [1988], and co-workers, have explored the problem of allocating vehicles to loads and demand regions using various stochastic programming approaches. Indeed, much of their work has examined the general problem we are interested in--the assignment of vehicles to loads in the context of truckload trucking. However, the approach adopted in that extensive body of work is different from the one we are taking. That approach (or, more properly, family of approaches) involves formulating and solving a two-stage or m-stage (with m greater than 2) stochastic program with recourse. In that work, recourse strategies, used to react when realizations of actual demands (and supplies of vehicles to meet demands) differ from the "expected" demands, are a side issue in the analysis. One can argue that our research harkens back to the beginning of the development of algorithms for vehicle routing and scheduling in which the limitations of computers of the day led to the development of greedy

approaches and local improvement heuristics [Fisher, 1995]. In some cases, the time limitations placed on the generation of real-time problems are formidable. Despite relatively fast and efficient computers and significant advances in heuristic and exact algorithms available for static vehicle routing and scheduling problems, standard approaches are likely unable to yield the most efficient solutions to their dynamic counterparts. In highly dynamic systems, developing strategies to react efficiently as customer requests and network conditions unfold may prove to be more important than determining the expected outcome. Furthermore, a strategy which may be optimal in the expected sense may, under certain realizations of the system, perform much worse than a simple strategy for reacting to actual conditions as the system evolves.

THE MOTOR CARRIER INDUSTRY: A BRIEF HISTORICAL PERSPECTIVE AND CURRENT STATE OF THE INDUSTRY IN THE U.S.

This research is concerned with the modeling of carrier fleet operations and the successful implementation of available technologies in these operations. While not directly related to this study, the emergence in recent years of a fiercely competitive and customer service driven environment, one which parallels developments in other service industries, has been driven by the forces of deregulation. To that end, a brief history of the trucking industry in the U.S. is provided to place the current operating environment in perspective.

This section briefly outlines the fascinating history of developments in the carrier fleet industry in the U.S.. We do not presume to mention all of the interesting events and developments that have helped to form the wide and varied enterprise of commercial trucking. The industry has been studied extensively by economists, sociologists, experts in the law and, in recent years, operations researchers and transportation engineers. The history of the many technologies which fueled the development of the industry, trucking culture, and the role of commercial freight operations in the growth of agricultural and manufacturing centers has also been explored.

The Development of the Motor Carrier Industry in the U.S.

The commercial trucking industry began during the first quarter of the twentieth century when the development of passable roads and the use of motorized transportation in logistics operations during the first world war led to the broad acceptance of motorized transportation as a practical alternative to horse drawn carriage. By the late 1920's, the commercial trucking industry, which grew as the interstate highway system developed, was competing directly and successfully against a declining railroad industry for short, medium length and even long-haul

traffic. Significant improvements in the economy, mechanical performance and comfort provided by new vehicles further propelled the rapid growth of the industry [Thomas, 1971].

The Introduction of Regulation

The Interstate Commerce Commission (ICC) was created in 1887 with the goal of protecting shippers from discriminatory pricing by railroads. The politics leading to the creation of the commission are under debate even today (for a comprehensive analysis see [Rothenberg, 1994]), but one widely held view is that a primary motivation was to ensure that small shippers and small communities would be provided service at reasonable rates. This apparent cross-subsidization (a means of using regulation to compensate one set of customers at the expense of others) of small shippers and small communities was not unique to the railroad industry.

During the first two decades of the twentieth century, neither automotive technology nor the system of roads in the country provided commercial trucking with many competitive advantages over rail or even barge operations for large scale movements of freight. However, as conditions and technologies improved, there was increasing pressure from the railroad industry, from the largest trucking companies, and perhaps, from the ICC itself to regulate this relatively new industry. Compounding the competitive issues between modes was a sharp reduction in economic opportunities during the depression. This, coupled with relatively low start up costs, led to a marked increase in the number of entries in to the owner-operator business. These smaller outfits, typically barely surviving economically, threatened the economic well being of both the railroads and large trucking companies at a time when freight was scarce overall due to the depression. In 1935, after nearly ten years of lobbying by the forces mentioned above, the U.S. Congress passed the Motor Carrier Act of 1935 [Rothenberg, 1994; Sampson et al, 1986; Thomas, 1971]. From 1935 until Congress's passage of the Motor Carrier Act of 1980, which greatly reduced trucking industry regulation, the ICC exercised control of trucking entry requirements, rates, mergers and in some cases routes. During that time, the industry was divided into three principal types of carriers: common, contract and unregulated. Common carriers engaged in for-hire transport over fixed or irregular routes and, required to obtain operating certificates from the ICC, were the most heavily regulated. Rates had to be filed thirty days before changes were allowed and were required to be reasonable and not unjustly discriminatory; standards were set for safety, equipment, employee qualifications and allowable work hours. Contract carriers, working under agreements with a small number of shippers (at one point no more than eight) were less carefully monitored than common carriers, although mergers and acquisitions had to be approved by the ICC. Finally, the Motor Carrier Act of 1935 defined a

class of not-for-hire or private carriers unregulated and immune from all regulations but those concerned with safety [Rothenberg, 1994, Thomas, 1971]. Small carriers or owner operators who faced barriers to entry into the market in addition to regulations found it difficult to compete with medium-sized and large carriers even in some niche markets where their costs were lower; the regulation of rates made them unable to compete on price. Railroads and the largest carriers benefited from regulation which artificially raised rates, limited competition and added significant administrative overhead. While regulation led to an increase in rates charged shippers, it also ensured service to those who might have been overlooked or overcharged in the absence of regulation. The largest shippers, the ones that would have been most likely and successful at launching a complaint about inflated rates, were more likely to create their own, unregulated, private fleets [Rothenberg, 1994].

A Deregulated Environment: the Emergence of Fierce Competition

Since the passage of the Motor Carrier Act of 1980 and the subsequent dissolution of the Interstate Commerce Commission in January of 1996, the industry has seen broad and sweeping changes. The effects of deregulation on truckload operations was even more marked than the corresponding effects on the less-than-truckload segment. While the fixed costs of entry into the LTL market are fairly high, requiring geographically dispersed terminals for all but small, local operations, the fixed costs of entering the truckload market include only the application fee for a license, a vehicle lease and the cost of insurance. Hence, deregulation allowed for the entry of thousands of new, generally non-union carriers and had an impact on the very structure of the industry [Corsi, 1993]. Since deregulation many private fleets have been eliminated and their tasks turned over to more competitive common carriers [American Trucking Associations 1987]. A leading industry analyst [Corsi, 1993] believes that far more opportunities exist to convert private (shipper owned) fleets into common carrier operations. In the competitive post-regulation market profit margins are even tighter than before [American Trucking Associations 1976, 1986]. Carriers are increasingly competing on service reliability and on-time performance in addition to cost. One of the principal arguments for the deregulation of the industry was that regulation was preventing carriers from offering shippers new and innovative services. Barriers to entry in the industry preserved the status quo and did not encourage innovation.

The years since deregulation have seen carriers competing more and more on service. Increasingly, carriers are offering logistics services that were neither required in a regulated environment nor possible before the development of reliable two-way communication, electronic data interchange, automatic vehicle location/identification and geographic information systems

technologies. The ability to attract agreements with "core" shippers (shippers that request movements from a particular carrier first), which can help to smooth the fluctuation of requests for service, depends heavily on the carriers' ability to provide reliable service. The availability of technologies, coupled with the popularity of just-in-time manufacturing systems and consumer demand for fresh perishable products and immediate delivery has further increased requirements for reliable and demand-responsive service.

This research is motivated by the increasing importance of providing reliable service that is responsive to shippers' needs for time-sensitive delivery at reasonable cost. ITS technologies and related fleet management decision support tools increase carriers' ability to compete effectively in meeting customer needs.

COMMERCIAL VEHICLE OPERATIONS APPLICATIONS OF INTELLIGENT TRANSPORTATION SYSTEMS TECHNOLOGIES

Intelligent Transportation Systems (ITS) involve the use of advanced communications and computation technologies in order to operate transportation networks more efficiently. The area of commercial vehicle operations or freight mobility has received relatively little attention in the research community when compared to research and implementation of travel and traffic management; however, CVO applications of ITS have achieved some early successes. It may be years before the cost of in-vehicle ITS technologies drops to a level acceptable to significant numbers of individual drivers and before efficient network-wide traffic management systems are in place. In contrast, commercial vehicle fleets have a significant economic incentive to seek to improve operational efficiency wherever possible. Furthermore, while a network traffic controller can suggest routes for individual drivers or classes of drivers, they cannot (in most cases) mandate the routes drivers will take. Members of vehicle fleets, on the other hand, generally accept directions from a central authority and hence react in a predictable fashion. In addition, the largest vehicle fleets involve a much smaller number of vehicles than even the smallest of traffic networks. Some of the issues addressed in commercial applications of ITS technologies correspond to more general work in ITS. Commercial vehicle operators and fleet managers wish to find the most efficient travel paths in the network. Advances in safety, centered on both vehicles and the roadway, are of significant interest to commercial users. Electronic payment services of tolls will save commercial users time en-route. In addition to all of the benefits of ITS experienced by individual (non-commercial) users of the transportation network, fleet managers and commercial drivers alike will benefit from the increased efficiency possible with automated fleet and freight management systems.

Central to this research is the assumption that, in the technology-equipped scenarios, all vehicles are equipped with some kind of continuous Automatic Vehicle Location (AVL) systems, typically GPS or geosyncronous satellite based; that all vehicles are equipped with continuous two-way communications devices; and that driver to dispatcher communication takes place within a short period of time. We compare these technology-equipped operations with ones in which real-time communication and vehicle location and status updates are not possible. Further, in the technology-equipped scenarios, it is assumed that the dispatch center(s) have a means to display the current locations of all vehicles typically with a digital map and a GIS (Geographic Information System) interface. The technologies employed in CVO ITS applications are briefly reviewed in the next section.

Technologies

The primary technological advances affecting commercial vehicle operations have been in automatic vehicle identification, automatic vehicle location and two-way communication systems. Related advances in computer technologies for both on-vehicle and home office use, geographic information systems (GIS) and Electronic Data Interchange (EDI) technologies have also impacted the way carriers operate [Mobility 2000, 1990].

Automatic Vehicle Identification. Transponder and associated reader technologies with both read-write and read-only capability and are widely in use. AVI technologies have many applications for commercial vehicle operations. Electronic toll systems employ time-saving AVI technologies. AVI systems installed at terminal entrance and exit gates allow LTL companies to monitor the movements of their drivers and equipment. Commercial vehicle electronic clearance and weigh-in-motion systems rely on AVI technologies.

Automatic Vehicle Location. Several viable options are available to perform automatic vehicle location. GPS, satellite-based Global Positioning System technology is the market leader and likely to gain ground in the next few years as differential GPS (DGPS) systems, which combine both satellite triangulation and ground based correction signals to enhance accuracy, become more widely available, and, accuracy problems in urban environments are solved. However, for trucking operations, which generally require less than the 1-5 meter accuracy of typical GPS systems, systems which use signals from ground based radio towers (like the Motorola Specialized Mobil Radio (SMR) system) coupled with Loran-C, a ground-based radio navigation system, or, those which use geo-synchronous satellites (for example the QUALCOM Omnitracs system) to perform both communication and tracking are widely in use [Jacobs, 1991,

EnRoute Technology, 1993]. These systems have been geared towards trucking applications and until recently, the 500-1000 meter accuracy they provide has been considered sufficient by most (long distance trucking) users. However, applications which include navigation, either on-board or at the central location, require more accurate position estimates in order to perform street-level calculations. In addition to geo-synchronous satellites and Loran-C, proximity beacon systems use strategically located short-range transmitters to periodically identify the locations of tracked vehicles. Identification may be made when a vehicle passes by a single beacon on the roadside or by triangulation of three or more signals. In addition to these basic systems, dead reckoning, map- matching and map-aiding techniques may be used to improve accuracy and cellular signals may be employed in triangulation schemes [Brown, 1992, Rothblatt, 1992, EnRoute Technology 1993].

Two Way Communication. Feasible options for two-way communications systems are even more numerous than those for vehicle location. Communication links available for this purpose differ in cost and sophistication. VHF, cellular, digital cellular or satellite links are all reasonable alternatives. The link used typically depends upon the desired frequency of communication and the distance between the dispatch center and the vehicles. Some communications systems allow for the transmission of character based messages only while others allow the transmission of both voice and data.

Other Related Technologies. Advances in computers, both for on-board and dispatch center use, are propelling the development of ITS systems for commercial vehicle operations. The state of the art in navigation and dispatching algorithms have continued to be more computationally efficient, and computing power has improved at an even faster rate. In addition, advances in database management and geographic information (GIS) systems have led to the development of sophisticated record-keeping and display systems. Continuing advances in spatial and temporal database management and econometrics methods to analyze data will lead to improved demand forecasting methods. Properly bundled, these technologies should lead to improvements in the efficiency and reliability of carrier fleet and other freight operations.

ITS America CVO Program Plan

The national ITS program plan describes the ITS program in the following way:

> "The Intelligent Vehicles Highway Systems [ITS] program applies advanced and emerging technologies in such fields as information processing, communications, control and electronics to multimodal surface

transportation needs. If these technologies can be effectively stimulated, integrated, and deployed, our society can benefit from more efficient use [of] our infrastructure and energy resources; make more informed choices about modes of travel and route alternatives; achieve improvements in safety, mobility, accessibility, and productivity; and reduce harmful environmental impacts, particularly those emanating form traffic congestion."[ITS America, 1994]

The development of fleet management systems for commercial vehicle operations holds particular promise as an application of ITS technologies which may be deployed on a small scale with little or no governmental involvement and which will have clear immediate economic benefits for users in addition to reducing energy consumption and pollution. The five primary goals of the ITS program are to improve safety, increase efficiency, reduce energy and environmental impact, enhance productivity and enhance mobility of transportation. Fleet management tools which incorporate real time assignment strategies with geographic information systems, automatic vehicle location and communication technologies can help to meet all five of these goals.

CVO User Services. The six CVO user services described in the National Program Plan for Intelligent Transportation Systems are [ITS America, 1994] :

1) Commercial Vehicle Electronic Clearance facilitates domestic and international border clearance, minimizing stops. Transponder equipped vehicles will be able to have their safety status, credentials and weight checked at mainline speeds. Safe, legal vehicles with no outstanding out-of-service citations will be allowed to pass inspection/weigh facilities without stopping. In addition to facilitating the movement of safety and regulation compliant vehicles this service should allow inspectors to concentrate their attention on those vehicles likely to need such attention.
2) Automated Roadside Safety Inspection facilitates roadside inspections by allowing real-time access to the safety and performance record of carriers, vehicles and drivers. Such access helps to determine which vehicles and drivers should be stopped for inspection. In addition to significantly improving the record keeping process and the speed of gathering information, previously identified problems can be monitored and many manual steps in the inspection process can be automated and improved through the use of sensors and diagnostics.

3) On-Board Safety Monitoring is intended to reduce driver and equipment related accidents by facilitating the automated sensing of the safety status of a commercial vehicle, cargo and vehicle at mainline speeds. Critical vehicle components are monitored as is driving time, driver alertness. It is intended that a warning about unsafe conditions would be provided first to the driver and then to the carrier and roadside enforcement officials.

4) Commercial Vehicle Electronic Processes Service provides for the electronic purchasing of credentials and automated mileage and fuel reporting and auditing. This provides the carrier with the capability of electronically purchasing annual and temporary credentials via a computer link. This will replace paperwork and reduce both carrier and state agency processing time.

5) Hazardous Materials Incident Response provides an immediate description of hazardous materials to emergency responders. The service will improve the safety of shipments of hazardous materials by providing enforcement and response teams with timely, accurate information on cargo contents to enable them to react quickly and correctly in emergency situations.

6) Commercial Fleet Management is the least well defined user service discussed in the national ITS program plan. Significant institutional issues remain to be worked out before this user service is put in place in a large scale fashion. This service provides communications between drivers, dispatchers and intermodal service providers. Traffic information will help drivers to avoid congested areas and would improve the reliability and efficiency of carrier operations. It is widely believed that most ITS services that benefit commercial vehicle operations only will be developed by individual companies or private sector consortiums. While it may be in the best interest for the public sector to provide some services directly to commercial vehicles most will be paid for and developed by fleet operators themselves.

The issues addressed in this research fall into the category of Commercial Fleet Management and are intended for use in conjunction with commercial vehicle electronic clearance, automated roadside safety inspection, on-board safety monitoring, automated commercial vehicles administrative processes and hazardous material incident response to form a fully equipped intelligent commercial transportation system. Commercial fleet managers may pick and choose from a wide variety of applications and technologies in order to develop intelligent transportation systems that best meet their needs and resources. In many cases, once an initial investment in technologies is made (on-board computers, two-way communication

systems, automatic vehicle identification or location systems) the cost associated with further increasing functionality is relatively small.

MODELING OF FLEET OPERATIONS

This section presents a review of relevant literature on the modeling of fleet operations. Early work on the modeling of distributions systems is introduced followed by a general discussion of dynamic fleet management. Research on classical vehicle routing and vehicle routing with time windows problems are mentioned for completeness, as is the large body of research (typically involving stochastic programming formulations) on dynamic vehicle allocation problems.

The modeling of distribution systems has received increasing attention since the late 1950's. Eilon, Watson-Gandy and Christofides [1971] introduce the main problems of concern in distribution management and, in addition to introducing important problem formulations and analyses, carefully review early literature on several problems of keen interest to this research: the traveling salesman problem; vehicle scheduling; and expected distances in distribution problems. These three problems, along with the modeling of vehicle loading strategies, form the foundation for early work in the modeling of fleet operations. In some respects, early efforts to model distribution systems for the purpose of deriving fundamental insights and principles are closer to the research presented in this report than some more recent work aimed and finding solutions to specific instances of vehicle routing and dynamic vehicle allocation problems. In spite of the observation that in general, problems addressed early on were restricted to static problems in which customer locations and demands are known, the greedy approaches to load assignment and route generation as well as the explorations of "efficient" local search heuristics conceptually binds our work to these earlier counterparts.

The truck dispatching problem is first defined in a paper by Dantzig and Ramser [1959] in which the goal is the near-optimal routing of a fleet of gasoline delivery trucks between a supply terminal and the service stations served by the terminal. This and other early work focused on static problems, even under a loose definition of dynamic problems [Powell, Jaillet & Odoni, 1995], as those in which one or more parameters is a function of time. The methods described in Eilon, Watson-Gandy and Christofides's book on determining expected distances in distribution problems were an early attempt to take into account the effects of the stochasticity of demand locations. The analyses of en-route diversion in a circular work area discussed in chapter 4 of this document follows a similar line of reasoning. A paper by Knight and Hofer [1968] which describes a manual scheduling approach, introduces the concept of *allocation* and subsequent

routing of vehicles when time window constraints exist and the time spent performing service at each call location is sufficiently long so that only a small number of calls can be assigned to each vehicle at a time. While far from a real-time application, the idea here is to group (allocate to a vehicle) calls which should logically be made by the same vehicle and then to order these in the most cost effective (from a distance traveled point of view) and time-window feasible way. The approach described in the paper is close in nature to the approach followed in this research. A difference is that long assignments (typically 1 to 2 days) in the truckload trucking application cannot be as easily identified for convenient mutual allocation because of the geographic separation of origin and destination locations.

Dynamic Fleet Management

Psaraftis [1988] provides an extensive review of dynamic vehicle routing problems and places these within the broader area of traditional vehicle routing and scheduling. Earlier, Bookbinder and Sethi [1980] present a survey of early work on the problem of selecting, at each instant in time, the optimal flow of commodities to various network sources and sinks so as to minimize the total cost of transportation. They briefly explore applications with stochastic demands, for example, the delivery of home heating oil, as well as travel paths with stochastic time delays. Dejax and Crainic [1987] present a review of models concerned with empty flows and fleet management models, many of them concerned with the dynamic aspect of freight transportation operations. Golden and Assad [1986] present developments in formulations and solution approaches for vehicle routing problems under many different operational assumptions; Powell [1988] presents alternative formulations for the dynamic vehicle allocation problem; Jaillet [1988] and Bertsimas, Jaillet and Odoni [1990] discuss solutions for the probabilistic traveling salesman problem.

Powell, Jaillet and Odoni [1995] present a review of research concerning stochastic and dynamic networks and routing and Gans and van Ryzin [1996a, 1996b] develop methods for analyzing the efficiency of dynamic dispatching operations. Their analysis draws heavily on queueing theory, modeling a depot-centered dispatching operation as a GI/GI/1 queue (in the single vehicle case). The more recent paper analyzes a general model of dynamic vehicle dispatching which seeks to capture the effect of congestion on system efficiency. Dispatching heuristics in which loads are serviced in batches are examined in order to generate an analytically tractable upper bound on the expected work in the system. These heuristics are not applied directly to dynamic dispatching problems but insights gained are used to develop more practical dispatching heuristics.

Two dispatching heuristics based on the analyses are suggested. In the first, a linear program in which system work is minimized is solved. Dual prices from the optimal solution of the LP are used to select the route (column in the LP) to execute. After completion of the selected route, the remaining routes are 'priced out' and the most attractive chosen for immediate execution. A second heuristic employs a similar approach but uses the dual prices from the solution of a different LP, one which gives a lower bound on the system work, to pick the next route to execute. The performance of these policies relative to two simple 'straw' policies are compared and are shown to be favorable. Their analysis introduces a novel approach to modeling dynamic dispatching systems and provides a method of estimating the efficiency of these systems by the total work remaining in the system. Another analysis of carrier fleet operations modeled as a system of queues is found in Tijms [1986]. This analysis uses a model of a distributed service system as an M/G/∞ queue and provides a bound on the delay. These results are used to determine the best allocation of vehicles to a set of fleets under conditions of stochastic demands.

The efficient management of fleets of trucks has been explored as early as 1959, when Dantzig and Ramser [1959] discussed linear programming based formulations for the near-optimal routing of a fleet of delivery trucks. The routing and scheduling of private (company) fleets and of less-than-truckload (LTL) operations has been addressed many times and in many different contexts in the past few years, but the truckload common carrier application has not received as much attention, except for the extensive work of Powell. However, the lessons learned and approaches taken in the related freight and fleet management problems lend significant insight into possible solution approaches for the dynamic vehicle allocation problem under real-time information. In particular, formulations for fleet management problems that treat individual vehicles separately, rather than as part of aggregate network flows, are well suited to modeling operations under real-time information. In order to take advantage of real-time information on vehicle and demand locations and the current state of the traffic network, individual vehicles and demands for service should be uniquely identified, rather than viewed as part of larger regional flows.

Although Powell [1996] introduces a hybrid formulation of the dynamic vehicle allocation problem suitable for the truckload carrier application in which vehicles are modeled individually in an assignment network and flows are predicted in an aggregate manner in a forecast network, most work on the DVA problem to date relies on stochastic (or stochastic-dynamic) programming formulations, which explicitly incorporate the stochastic nature of both future supplies of vehicles and demands for service. An alternative approach is one in which a smaller, simpler problem

(than the stochastic programming formulations) is solved often, and in response to unfolding customer demands. This approach gives little (if any) attention to future uncertainties. It relies instead on the fact that the problem will be re-solved when accepted demands for service, vehicle availability, and in some cases, travel network conditions change. Because demands for service are generally time-constrained, formulations of the real-time problem may share many similarities with the classic Vehicle Routing Problem with Time Windows (VRPTW). The main differences are in the length of the average moves made, a typically smaller number of customers, and in the fact that in most DVA problems it is assumed that a vehicle, whether it be a truck, taxi or rail car, is dedicated to performing work for a single customer at a time, and that not until a job is complete is the vehicle available to take on another assignment. The next section provides a brief review of the extensive literature on the classical vehicle routing problem and the vehicle routing problem with time windows followed by a discussion of and approaches to modeling dynamic vehicle allocation problems.

Classical Vehicle Routing and Vehicle Routing with Time Windows

The vehicle routing problem (VRP) has been studied extensively in the literature. Bodin et al. [1983], Christofides [1985], and Golden and Assad [1988] and most recently Fisher [1995], provide extensive surveys of the different types of vehicle routing problems and solution techniques employed to solve them. These problems have been divided into three categories: 1) routing, 2) scheduling, and, 3) routing and scheduling, and have been classified by various characteristics: fleet size (one or more vehicles), fleet type (heterogeneous or homogeneous), number of depots (one or more), nature of demands (deterministic or stochastic), vehicle capacity restrictions, maximum route times, operations (pickup, delivery, service, mixed), costs and objectives [Bodin et. al, 1983].

Routing problems are primarily concerned with the spatial aspects of the problem while scheduling problems focus on temporal aspects. Problems in which both spatial and temporal aspects are important constitute the routing and scheduling class.

The general vehicle routing problem with time windows (VRPTW) involves the design of a set of minimum cost routes originating and terminating at a central facility (depot) for a fleet of vehicles which services a set of customers with known demands and in which each demand for service must be started and/or completed within a given time interval. For instance, the vehicles may make up a service fleet such as utility, or telephone repair. In such applications, the vehicles are not assumed to be subject to capacity constraints although they may be subject to maximum length-of-day constraints. Another formulation of the problem, one in which the vehicles

themselves are capacitated, corresponds to a package pick-up and delivery problem. The objective in solving this problem is to design a complete tour for each vehicle, starting from and ending at the depot and servicing each customer within its assigned time window at a minimum cost.

The vehicle routing problem with time windows has not been as broadly studied in general, although specific problems have been addressed. Dumas, Solomon and Soumis [1991] address time-constrained routing and scheduling problems; Solomon [1987] provides an overview of heuristic solution approaches for the VRPTW; and, Solomon and Desrosiers [1988] present a survey of approaches and advances made for the VRPTW and also related problems including, among others, the time window constrained traveling salesman problem, the shortest path problem, the minimum spanning tree problem, and the pickup and delivery problem. Christofides, Mingozzi and Toth [1981] and Baker [1983] propose branch and bound and dynamic programming based optimization approaches to solve the single vehicle TSPTW while Kolen, Rinnooy Kan, and Trienekens [1987] introduce a branch and bound method for solving the capacity constrained multi-vehicle problem. Koskosidis, Powell and Solomon [1992] present an optimization based mixed integer formulation that extends to the VRPTW Fisher and Jaikumar's [1981] algorithm for solving a standard VRP.

Bodin et al [1983] classify solution strategies for vehicle routing as: (1) Cluster first-route second, (2) Route first-cluster second, (3) Savings/Insertion, (4) Improvement exchange, (5) Mathematical-programming-based (6) Interactive optimization and (7) Exact approaches. The added complexity of time window constraints mean that many of the methods devised for the standard VRP do not work well for the VRPTW. For example, a cluster first-route second strategy that groups demands by proximity would likely not produce time-feasible subsets for routing. Likewise, a route first-cluster second strategy in which a large route is partitioned into multiple routes would not be applicable since time window constraints would preclude the possibility of building the large route to begin with. However, heuristic methods based on savings insertion and improvement exchange have been used with success on even large problems, and math programming based optimization methods have been successful on some medium sized problems. Exact approaches have been less successful in general because of the size and inherent complexity of most problems. Interactive optimization methods have potential but in large difficult to solve problems would best serve as an addition to another route construction method. In some vehicle routing and VRPTW problems (such as the service fleet problem mentioned) travel and service times may be stochastic. The next section presents issues of

interest in vehicle routing problems in which the system evolves over time. In general, one or more aspects of the system is stochastic.

Dynamic Vehicle Routing

Dynamic vehicle routing typically refers to the dispatching of vehicles to serve multiple demands for service in a real-time manner. Vehicles, for example, may be performing pick-up and delivery services; they may be repair vehicles, delivery vehicles or taxicabs. The chief characteristic of these problems is that, like the dynamic vehicle allocation problem described in chapter 2, a (sometimes large) fraction of the demands become known as the work period goes on. Psaraftis [1988] presents a review of dynamic vehicle routing problems and points out that although published research on vehicle routing was abundant, until that time very little had been published on dynamic vehicle routing problems. Psaraftis identifies the main differences between static and dynamic vehicle routing. In dynamic problems: (1) Time dimension is essential; (2) Problem may be open-ended; (3) Future information may be imprecise or unknown; (4) Near-term events are more important; (5) Information update mechanisms are essential; (6) Resequencing and reassignment decisions may be warranted; (7) Faster computation times are necessary; (8) Indefinite deferment mechanisms are essential (this point applies to problems where time windows and deadlines are loose constraints); (9) Objective function may be different; (10) Time constraints may be different; (11) Flexibility to vary vehicle fleet size is lower; and (12) Queueing considerations may become important. Most of these points apply directly to the dynamic assignment problem for carrier fleet operations. The main difference between what are considered vehicle routing and vehicle assignment problems are the length of time to perform service and the relative number of assignments that each vehicle is responsible for at a given time. In the dynamic vehicle routing problem, multiple customer orders are assigned to each vehicle; in dynamic allocation, the vehicle can serve only one customer at a time and the number of current assignments for each vehicle is one or a very small number. Psaraftis points out the importance of developing techniques to perform local updates on routes and assignments coupled with good initial solutions. These concepts apply directly to the real-time formulations of the dynamic vehicle allocation and routing problem for carrier operations examined in this research.

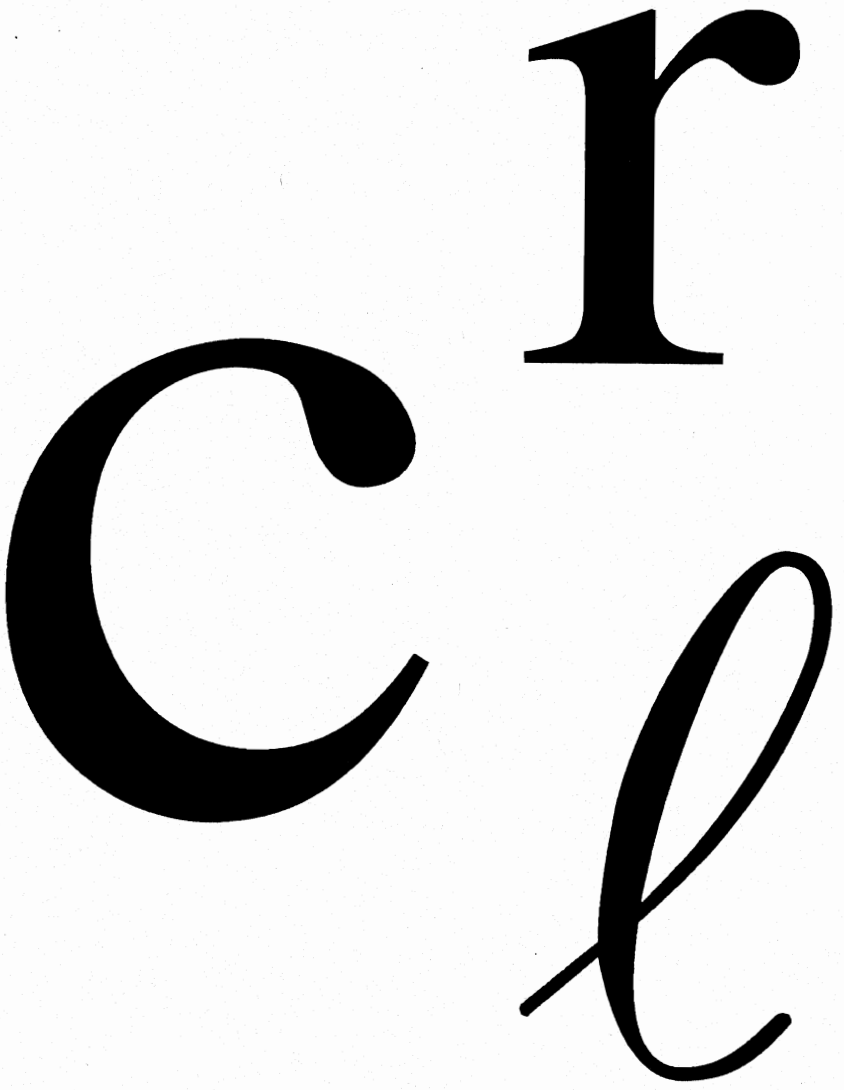
Powell, Jaillet and Odoni [1995] present a survey of dynamic network models. They identify the general issues associated with modeling dynamic problems and list the following key decisions to be made: Deterministic vs. stochastic; Myopic vs. dynamic; Choice of objective function; The planning horizon; and, Spatial and temporal aggregation. These questions are

coupled with the following issues: 1. Developing accurate models under uncertainty; 2. Identifying 'efficient' formulations; 3. Design of efficient solution algorithms; 4. Planning horizons and truncation errors; 5. Errors due to spatial and temporal aggregation; and 6. Evaluating a stochastic and dynamic model. While most of the research presented in this document leans toward deterministic and myopic models which are solved repeatedly over a short horizon, thereby sidestepping some of the complexities of dynamic (and hence typically stochastic) models, the issues identified in this survey offer an insightful method of examining and evaluating real-time routing and assignment systems.

Dynamic Vehicle Allocation

The dynamic vehicle allocation problem has been studied extensively by Dejax and Crainic [1987], Powell [1986, 1987, 1988, 1994], Frantzeskakis and Powell [1990] and Cheung and Powell [1995]. The DVA problem involves managing a large fleet of vehicles over time to maximize profits. This problem arises in many different contexts. Carrier (trucking fleets), railroads, maritime shipping companies, taxi fleets and auto or truck rental companies all must solve vehicle and container allocation problems. The problem is dynamic because the allocation of vehicles at any time affects the state of the system in the future. The problem is typically stochastic, because future demands may not be anticipated with certainty.

Deterministic Assignment. Powell [1988] presents a review of alternative formulations for the DVA. These include deterministic transshipment networks, stochastic/nonlinear networks, Markov decision processes and stochastic programming formulations. Some of these will be discussed in later sections. Powell [1994] discusses state-of-the-art formulations for the DVA applied to truckload trucking. In particular, a hybrid of the deterministic assignment and stochastic-dynamic models which may be promising for real-time implementations is introduced. The following is a simple, static formulation of the DVA problem as an assignment problem; the sum of the cost of moving vehicles to meet known demands, added to the cost of holding vehicles in anticipation of future demands and the penalties incurred for refusing to carry certain loads is minimized subject to the constraints that all drivers must be assigned to a load or held in a region, and that all loads must be assigned to a driver or refused. This formulation and related modified formulations are discussed in chapter 3. It is introduced next:



penalty cost of not assigning any drivers to load ℓ,

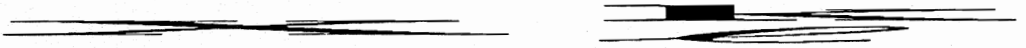
Let L be the number of loads,

K be the number of available drivers (vehicles),

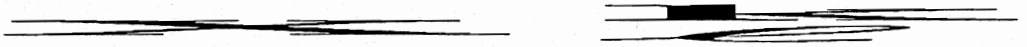
cost for driver k to serve load ℓ,

holding cost for driver k (cost of not assigning driver k to any loads),

Decision variables:

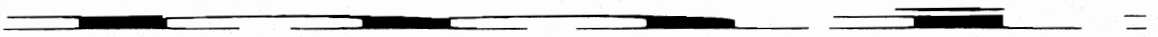


1 if driver k is assigned load ℓ,



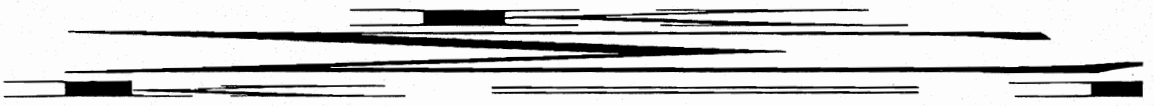
0 otherwise.

Then the problem of assigning drivers to loads may be stated as:

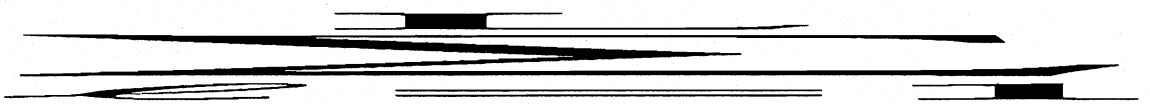


(2.0)

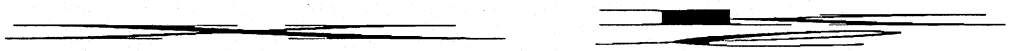
subject to:



for $\ell = 1,2,...,L$ (2.1)



for k=1,2,...,K (2.2)



for ℓ = 1,2,...,L, k=1,2,...,K (2.3)

Constraints (2.1) specify that a load may be assigned at most one driver, (2.2) specify that each driver is assigned at most one new load, while (2.3) ensure the non-negativity of decision variables. This simple assignment model expresses many of the important aspects of the problem and is computationally efficient. It includes the costs of moving empty or holding drivers in an area, as well as the costs of refusing (or not accepting) a load to be moved. The model does not capture the uncertainty of future demands for service, nor deal with the problems associated with solving a 'snapshot' problem. Nor does it include the possibility of moving an empty vehicle to another region to be held in anticipation of future demands. In addition, this formulation does not allow a driver to be assigned a sequence of loads although an appropriate pre-processing step could be introduced in an effort to "chain" loads that should logically be moved in sequence by the same driver. These chains would then be considered by the assignment model to be a single load. This model allows for a fairly high degree of operational realism as it is possible to incorporate many objectives into the cost structure. Waiting loads may be discounted, the cost for a particular driver to move a particular load or class of loads may be discounted, the penalty for not assigning a driver who has been idle for some time can be increased etc.

Deterministic and Stochastic Dynamic Models. The hybrid model for dynamic assignment introduced in Powell [1994] offers a promising alternative. This model has two components. The first is a static assignment model with assigns specific drivers to specific loads. Individual loads and drivers are represented by nodes in the network and arcs representing the assignment of drivers to loads. This model has the advantage of allowing a high degree of operational realism. The second model is a dynamic network which works at an aggregate level and includes forecast

demands in addition to known but not yet serviced loads. Loads which are ready for immediate or near term pickup are represented by an origin node in the assignment network and a destination node in the forecast network. In addition to the (loaded) links connecting origin-destination pairs, empty repositioning links connect drivers with regions in the forecast network. A complete description of the model is found in Powell [1994].

SUMMARY

Chapter 2 presents a review of background relevant to the operations and modeling of carrier fleet operations under real-time information. A brief history of the U.S. trucking industry discusses some of the reasons for heightened competitiveness in the past few years and introduces the principal technologies and capabilities of ITS for commercial vehicle operations. The literature review that follows shows how this research, which examines greedy approaches to load assignment as well as "efficient" local search techniques and exact approaches for the routing of vehicles through a small number of assignments in the presence of time window constraints, builds on a large body of work. Early work on the modeling of distributions systems is introduced followed by a discussion of dynamic fleet management in general. Research on classical vehicle routing and vehicle routing with time windows problems are discussed as is research on dynamic vehicle allocation problems. Most work in the published literature on the truckload trucking application has had as its subject the development of large-scale stochastic programs. While that important body of work is mentioned for completeness, the approach followed in this research is fundamentally different in its underlying rationale.

The next chapter introduces the conceptual and theoretical framework for the analysis of dynamic dispatching strategies for carrier fleet operations under real-time information. Modeling assumptions related to the dynamic load acceptance and vehicle to load assignment process are made explicit. Mathematical formulations of the real-time assignment sub-problem of the dynamic vehicle allocation and routing problem, including those implemented in simulation experiments are presented.

CHAPTER 3 CONCEPTUAL FRAMEWORK

INTRODUCTION

Chapter 3 introduces the conceptual and theoretical framework for the analysis of dynamic dispatching strategies for carrier fleet operations under real-time information. The dynamic load acceptance and vehicle to load assignment process for carrier fleet operations, introduced in chapter 1 is defined in finer detail than in the introductory chapter, and simplifying assumptions made for the purpose of modeling these operations are made explicit. In chapter 3, a simple model for estimating the costs associated with carrier fleet operations is introduced, along with the methods used in this study to evaluate the performance of assignment strategies. Also, several mathematical formulations of the real-time assignment sub-problem of the dynamic vehicle allocation and routing problem are introduced under different assumptions regarding the availability of real-time information on vehicle locations, demands and traffic network conditions, and the flexibility of current assignments. Assignment strategies investigated vary in terms of the length of "look ahead" time they include and in the extent to which they allow existing assignments to be modified. Some allow only small modifications while others consider generating completely new assignments when new demands arrive, travel network conditions change or when the status of one of more driver or vehicle changes. This work focuses primarily on flexible assignment strategies which allow modification of existing assignments but, in the interest of computational efficiency do not seek to evaluate all possible assignments.

Simulation plays a central role in the examination of the behavior of various load acceptance, assignment and re-assignment strategies. Potential for the development of analytical models (along the lines of the models discussed in chapter 4) is limited: inherent complexities of the problem must be ignored in the interest of tractable analysis. Analytic investigations rely on identifying simplified problem instances, from which inferences about more general problems can be drawn. In chapters 5 and 6, the analysis begun in chapter 4 is extended in a simulation framework. Typical assignment rules result in service times that are neither deterministic nor identically and independently distributed (IID) random variables, rendering analytic investigation impractical. For this reason, extensive use of simulation is made to investigate the performance of fleet operating strategies, including load acceptance, assignment and re-assignment strategies. The next section provides a motivation for developing decision rules for real-time fleet management that are "locally oriented". This is followed by the context for and definition of the problem examined.

LOCALLY ORIENTED LOAD ACCEPTANCE AND ASSIGNMENT HEURISTICS

This research has as its primary goal identifying decision rules for load acceptance and assignment that are locally oriented but result in long term efficiency. Decision rules can be local with respect to one or more of the following: temporal, spatial, or algorithmic considerations. The decision time frame may be local, so that only current is information is employed; the decision region may include a "local" or logical subset of vehicles or of demand locations, chosen for geographic proximity, substitutability, customer preference or driver domicile; in addition, if a current solution is to be modified rather than completely re-generated when new demands are accepted, service is completed, or changes driver and vehicle availability and traffic network occur, then the decision region is limited to local decisions in the region of the current solution.

Local decision rules are of particular interest in systems in which problems tend to be very large, decisions must be made quickly and information about the future is incomplete or imprecise. Under most conditions, information available to dispatchers is limited both spatially and temporally. Demand forecasts may be constructed to provide limited insight into how the system will evolve, however, ability to accurately predict demand is limited.

Because of relatively limited accurate information available about the future, and potentially extensive information available about current conditions, it is desirable to identify and exploit "locally oriented" real-time dispatching strategies. With perfect hindsight, over time, solutions of global, system-wide assignment problems would naturally out-perform those chosen by local decision processes. Effective solution approaches to solving very large static vehicle routing and assignment problems have been developed over the years. In a dynamic environment, however, demands may not be predictable and traffic network conditions as well as driver/vehicle availability can change, sometimes dramatically. In such an environment "good" local decision rules, coupled with the ability to react quickly as changes in demands, traffic network conditions and driver/vehicle availability occur may offer the most effective way to provide reliable and (customer demand) responsive service.

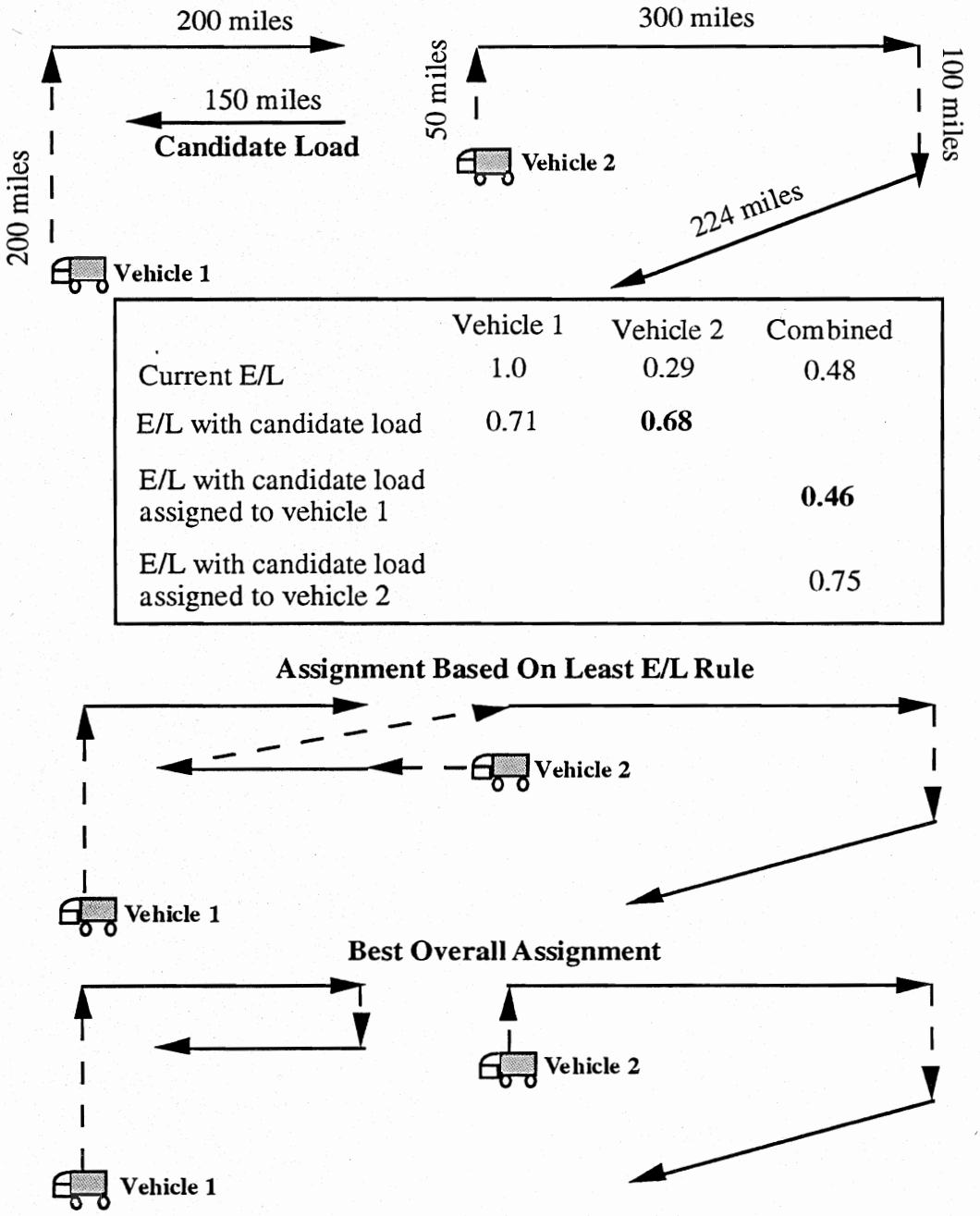
This research examines several operational strategies that make use of local decision rules. Various assignment techniques are examined analytically as well as through simulation. A subset of these involve single rather than multiple load assignment processes. That is, rather than accumulate loads in batches for simultaneous assignment to vehicles, loads are assigned, one at a time, as soon as they arrive. The choice of purely sequential, rather than simultaneous assignment methods is guided by the nature of the problem and an operational framework driven by changes that occur in a continuous manner, rather than at pre-specified decision instances. However, when more than one unassigned load is allowed to wait until an assignment is made,

then simultaneous methods offer opportunities for improvement. Both exclusively sequential assignment rules and applications of classical assignment algorithm are discussed in chapter 5 and results for a simulation based analysis of the performance of these is presented in chapter 6.

Three assignment rules that evaluate the fit of a candidate load with loads already assigned to vehicles are examined. For each vehicle, a measure of the fit of the candidate load is calculated, based on one of three objective function proxies. The first is the ratio of empty to loaded distances that must be traveled in order for the vehicle to serve the candidate load, along with other loads currently assigned. The second is simply the total empty distance that the vehicle must cover in order to serve the candidate load and loads currently assigned. The third is the marginal cost, calculated as the cost of the additional empty move, that the vehicle will incur while serving the candidate load. In each case the "best" (least empty distance, pickup deadline feasible) ordering of loads is evaluated and the load assigned to the vehicle with the best ranking on the measure. Assignment decisions made under this and other local decision rules can lead to inefficiencies with respect to distances traveled to provide service. Figure 3.1 provides an example of this. Two vehicles have loads assigned when a candidate load enters the picture. The new assignment is chosen by adding the load to the "route" including the candidate load and previously assigned loads with the lowest E/L ratio. In the example chosen, this greedy assignment leads to higher overall empty movements than a rule that takes both vehicles into account might provide. Despite the ease with which such examples can be constructed, the long term performance of this simple rule is quite good. Simulation experiments have shown that in the presence of pickup deadlines, and over a horizon that spans a more than just a few loads, this greedy decision rule works fairly well. The three decision rules used to make final load to vehicle assignments in the real-time assignment strategies under investigation are outlined in this chapter; their performance in simulation experiments discussed in chapters 5 and 6.

PROBLEM CONTEXT: TRUCKLOAD CARRIER FLEET OPERATIONS

In this section the context for this research is specified in greater detail than in chapter 1, where it was introduced. The carrier operations planning process is examined, a model of operational costs the primary objectives related to carrier fleet management are introduced, and mathematical formulations for the cost model and measures for these objectives are given in chapter 3.



	Vehicle 1	Vehicle 2	Combined
Current E/L	1.0	0.29	0.48
E/L with candidate load	0.71	**0.68**	
E/L with candidate load assigned to vehicle 1			**0.46**
E/L with candidate load assigned to vehicle 2			0.75

Figure 3.1 Example of poor short term performance of a local "greedy" decision rule

The context for this research is truckload carrier fleet operations in which each assignment involves a vehicle moving one load at a time from the load origin to the load destination. The problem studied involves the management of a fleet of vehicles over a typically wide geographic region over time. An assumption of this research is that the fleet of vehicles is under the control

of a central authority, the dispatcher. The management issues addressed are limited to the acceptance (or rejection) of customer requests for service and the allocation of vehicles to loads and to empty repositioning moves. Some carrier companies own the vehicles in their fleet, some employ the services of owner-operator drivers, and others maintain a fleet of both company drivers and owner-operators. We ignore these distinctions, which are not essential for the operational strategies considered here, and assume that a central authority has the responsibility for directing the movements of vehicles in its fleet. Several equipment types may be available; this analysis assumes however, that equipment types are homogeneous or at least substitutable. We further ignore the additional complexity of single and double trailers and of loads that are carried in shipper-owned trailers. The assumption is that a driver and vehicle combination includes a trailer that is loaded and unloaded at customer sites. In addition, this research does not take into account federal, union or company regulations concerning, for example, the length of time a driver may be on duty or driving. That level of detail, while essential for the successful development of implementable fleet management systems, is beyond the scope of this research. We further ignore the complications that might arise if some vehicles are driven by teams of two drivers rather than individuals.

It is assumed that once a vehicle is loaded at a customer site, it cannot provide service to another customer until unloaded, typically at the destination point of the load. An exception to this assumption would be in the case of a load swap where two or more drivers meet at some point en-route and either swap trailers or unload a whole load. While this research examines en-route diversion and re-assignment of empty vehicles, we do not explicitly examine loaded vehicle re-assignment strategies; this may be examined in future research. Scenarios in which all vehicles are equipped with a continuous Automatic Vehicle Location (AVL) system as well as devices allowing for two-way driver to dispatcher communication are examined. These technology-equipped operations are compared to those in which real-time communication, vehicle location and status updates are not possible. Further, though not essential to this investigation, it is assumed, in the technology-equipped scenarios, that the dispatch center(s) has a means to display the current locations of all vehicles, typically with a digital map and a GIS (Geographic Information System) interface.

Demands for service are highly stochastic. Other aspects of carrier fleet operations may also have a high degree of variability. Travel times may vary significantly, due to unexpected congestion, road closures, weather or equipment failures. The time spent at loading and unloading points may also vary. We do not explicitly include this variability in our formulations of the problem. Rather than work with stochastic representations of travel and loading and

unloading times explicitly (i.e. by estimating a probability distribution for these times) we assume that these values are known. In fact, the current values will be known through relatively frequent updates made possible by real-time information systems deployed as part of the ITS architecture. When actual conditions vary from those expected, the real-time dispatching systems envisioned are designed to react to the unexpected.

The Carrier Operations Planning Process

The carrier operations planning process can be viewed in two parts. First, there is the need to manage the supply of vehicles and drivers to provide timely service to customers. The supply problem includes the assignment of drivers to known loads, reassignment as changes in demands, driver and vehicle availability and traffic network conditions occur, and the repositioning of idle vehicles in anticipation of future demands. Related to this is a second problem - the need to effectively manage customer requests for service. The load management problem includes both the decision to accept (or reject) requested loads for service and the active solicitation of loads in regions in which the supply of vehicles to provide service currently exceeds, or is anticipated to exceed demand. The context of truckload trucking and other dynamic fleet operations is such that it is not possible to serve all requested demands at all times [Powell, Jaillet & Odoni, 1995]. It may be necessary to turn down requests for service if time window, regional or system wide capacity constraints cannot be met. Careful load acceptance and solicitation decisions are therefore very important. When some loads are likely to be refused, the problem of effectively choosing the subset of loads to serve has a significant effect on the profitability and efficiency of an operation.

The load acceptance and driver to load assignment processes may be tightly or loosely coupled, or tackled separately. In principle, they are separated, because the load acceptance decision must be made when a request arrives, whereas loads may be assigned to a driver at any time after the request is received. Shippers call a carrier requesting that a vehicle be available at a pickup location on a specific day, by, or within a specific time, to carry a load to a specific destination. The carrier must decide quickly whether to accept the request. Advances in computing technology and scheduling and assignment heuristics make it possible to couple the acceptance and assignment processes or to use information about the estimated cost and feasibility of providing service in the load acceptance decision. Assuming the carrier has accepted the load, a vehicle (and driver) moves the load from its origin to its destination and unloads, by which time the carrier must decide what to do next with the vehicle. Under some operational strategies, a vehicle may be assigned a sequence of tasks to follow. If no appropriate

assignments are available, the idle vehicle may be repositioned to another region in anticipation of future loads in that region, or held in anticipation of future loads in the destination region. Loads may have firm or somewhat flexible pickup deadlines. A carrier may decline a load, but it may not accept a load unless it can meet the agreed upon deadlines. A carrier is paid a revenue proportional to the length of the haul and may also receive a small fixed revenue for each load (not proportional to distance).

This research examines several operational strategies, defined as the combination of a load acceptance strategy, and an assignment strategy, which may include re-assignment as well. In some, the load acceptance and assignment processes are separate; in others, they are coupled.

The next section presents a simple model of the cost structure for carrier fleet operations. This model is intended to be explanatory and to introduce the costs used in the evaluation of the performance of operational strategies in this analysis.

Operating Costs

The cost structure for carrier fleet operations can be fairly elaborate: in this research, we make simplifying assumptions that retain the essence of the impacts of operational changes on costs. Costs have been separated into higher level fixed costs, fixed costs associated with each vehicle/driver combination and operational costs that vary with the distance driven.

Any discussion of costs, revenues and estimates of operating efficiency must address the issue of the time period for evaluation. An evaluation of the performance of a fleet of vehicles over several years would include in its analysis the longer term costs of maintaining a central control facility while a study of an assignment strategy over a one month period would focus on short term costs including driver wages, fuel, etc. that are directly affected by the performance of that strategy. If the evaluation period is rather short, the effect of the end (or beginning) of the time horizon on performance indices measured over the period of evaluation may be dramatic. An operating strategy might perform quite well over a day or a week, but could have disastrous long term performance if, for example, it sent many of the drivers on long loaded moves to regions where there was little hope of picking up a new load upon delivery.

Several time horizons are chosen for this study. In some cases the long-run expected performance of a set of dispatching rules is examined in a simulation framework, requiring an artificially long time horizon of as much as several years to guarantee that steady state results can be examined. In most cases, a shorter time horizon, generally twenty-six weeks long is sufficient to provide steady state performance estimates. In a typical truckload operation, drivers (or teams of drivers) are on the road for three to six weeks without returning to their home

locations for an extended break. The longer, twenty-six week horizon is chosen to insure that system startup effects are mitigated.

This study ignores fixed costs not associated with individual driver/vehicle combinations. The assumption here is that day to day operations will not effect the size of the dispatch, marketing and management teams and that new facilities will not be erected or old ones sold or abandoned in the short term. Were there a compelling reason to include the fixed costs of operation in similar analyses, these could be apportioned to vehicles and added to that component of vehicle cost not related to distance traveled.

Driver compensation schemes vary across companies, from a simple payment proportional to loaded distances traveled to compensation which includes payment for distances driven, bonuses for safe driving, compensation for layovers, time spent loading and unloading, multiple stops, and premiums for short distance loads. The compensation for loaded and empty traveled distances may be equal, or may be very different. This study makes the assumption that drivers are compensated for both time and distance traveled and that loaded and empty travel are compensated at the same rate. This study does not differentiate between single drivers and teams of drivers. Further, we assume that all drivers are compensated at the same rate. We make the assumption that drivers are paid a fixed amount for each day worked, and on top of that are paid for distance traveled. Idle time is accounted for by the fixed daily charge for each driver.

Costs associated with the vehicle are fixed costs per day and operating costs proportional to distance driven. As mentioned in the previous paragraph, in an analysis of costs in the short run, fixed costs would likely not include a proportion of the cost of maintaining a management team and depot location. In a long term analysis the fixed costs per vehicle could be increased to reflect this cost. These costs are discussed in finer detail later in this chapter.

A simple model of profit over a fixed time horizon given by:

$$\text{Profit} = \sum_{\text{over all loads served}} \begin{bmatrix} \text{Revenue} \\ -\ (\text{empty travel cost}) \\ -\ (\text{loaded travel cost}) \\ -\ (\text{handling cost}) \\ -\ (\text{daily vehicle charges}) \\ -\ (\text{daily driver charges}) \end{bmatrix}$$

The management of a carrier fleet company has several (often competing) objectives. While carriers wish to maximize profits, they also want to meet shipper expectations. For some companies, environmental concerns are important. Reducing costly and irritating service delays may also lead to a reduction in the contribution to congestion (and hence pollution) made by a vehicle fleet. In addition, keeping drivers satisfied is an increasingly important objective. Driver turnover in the truckload portion of the industry is high. Training and monitoring new drivers is costly as is finding qualified replacements for departing drivers. These objectives and the measures used to evaluate them in this analysis are discussed in the next section and outlined in table 3.1.

Objectives

The primary objectives related to carrier fleet operations are to maximize carrier profitability and service quality. Two secondary objectives are to minimize environmental impact and to achieve driver satisfaction. Table 3.1 lists these objectives and some related measures of effectiveness. In chapter 3 a profit model is introduced and the measures are defined mathematically. The measures introduced here are used to evaluate the set of operational strategies outlined in chapter 5 and examined in chapter 6. As mentioned in the preceding section, some of these are competing or even conflicting objectives. For example, the objective of maximizing the ability of a fleet to respond to time-sensitive demands might be achieved by rejecting all loads that do not require immediate service, in order to keep a sufficient subset of the fleet available to accept time-sensitive loads. Such an operational strategy would be in conflict with the objective of maximizing revenue earned.

The next section provides a definition of the problem examined in this research, for which the various measures of performance (objective attainment) are subsequently defined.

PROBLEM DEFINITION

Mentioned in chapter 3, the carrier operations planning process contains two related problems: the problem of assigning loads to drivers and the problem of accepting or rejecting requests for service. Both are discussed in this section along with mathematical definitions for the measures listed in table 3.1. A general model of the carrier operations planning problem is presented and a mathematical representation of the cost model used in this research is presented. The cost model, is introduced and presented in chapter 3.

TABLE 3.1 OBJECTIVES AND MEASURES

Issues	Objectives	Measures
Profitability	• maximize profit (revenue - cost) • maximize revenue • minimize empty distances traveled to provide service • maximize ability to serve loads and high revenue loads	1) operating profit (revenue - cost) over fixed time horizon 2) revenue earned over fixed time horizon 3) ratio of empty to loaded distances traveled 4) fraction of requested high revenue and overall loads accepted
Service Quality	• minimize shipper wait time • minimize missed pickup deadlines • maximize ability of the fleet to respond to requests for immediate or near term pickup	5) average wait time for service and associated variability 6) fraction of pickup deadlines missed 7) fraction of requested time sensitive loads and overall loads accepted
Environmental Impact	• minimize fuel consumption • minimize contribution to congestion	8) ratio of empty to loaded distances traveled
Driver Satisfaction	• achieve reasonable driver compensation • achieve fairness in load assignment	9) fraction of time spent moving loaded 10) variability of time spent moving empty and loaded, across a fleet

The General Model

Service requests arrive at a central location over time. The carrier must decide immediately whether to provide service to a requested load. If accepted the load must be assigned to a driver for service. Loads may have associated time windows for pickup and/or delivery.

Each request for service and accepted load has an associated attribute record which includes, at a minimum, the exact location of the origin and destination points, handling time (loading and unloading), the earliest and the latest pickup times for the load. Load attributes (for both requested and accepted loads) might also include equipment requirements, preferences for a particular driver, etc., but these are not included in this analysis.

Each vehicle/driver combination also has an associated attribute record containing information about the exact current location of the vehicle, the state of the vehicle (moving loaded, moving empty, idle and available, idle and unavailable) the equipment type currently in use by this driver, the driver's domicile location, salary, seniority and other relevant characteristics.

In the general model presented, a carrier may accept more loads than it can feasibly serve. In that case, loads must be served by out of fleet drivers. The carrier accepting the load receives the revenue associated with the load but must pay a fee, proportional to the loaded distance associated with the load, to the out of fleet driver.

A schematic of the carrier fleet operations process in which requests for service are filtered through a load acceptance process and then accepted loads assigned to vehicles for service is shown in figure 3.2.

Revenue, Cost and Profits Under Given Assumptions

The following assumptions are made with respect to the costs and revenues associated with providing service.

1) Service requests are received for a total of N loads to be served during a period of H days.
2) K vehicles are in the fleet and all vehicles are in service for all of the H days.
3) Drivers are compensated for time, measured in days, and for distance traveled.
4) Vehicle costs are assessed for time, measured in days, and for distance traveled.
5) A penalty cost, p, is assessed per unit distance for loads accepted for service but served by out of fleet drivers.

6) Vehicles are in service continuously. Breaks and trips off duty are not accounted in this analysis.

7) Revenue has two components, a fixed component and a component linearly related to loaded distance traveled.

Figure 3.3 shows the relationship between fixed revenue and revenue proportional to distance traveled.

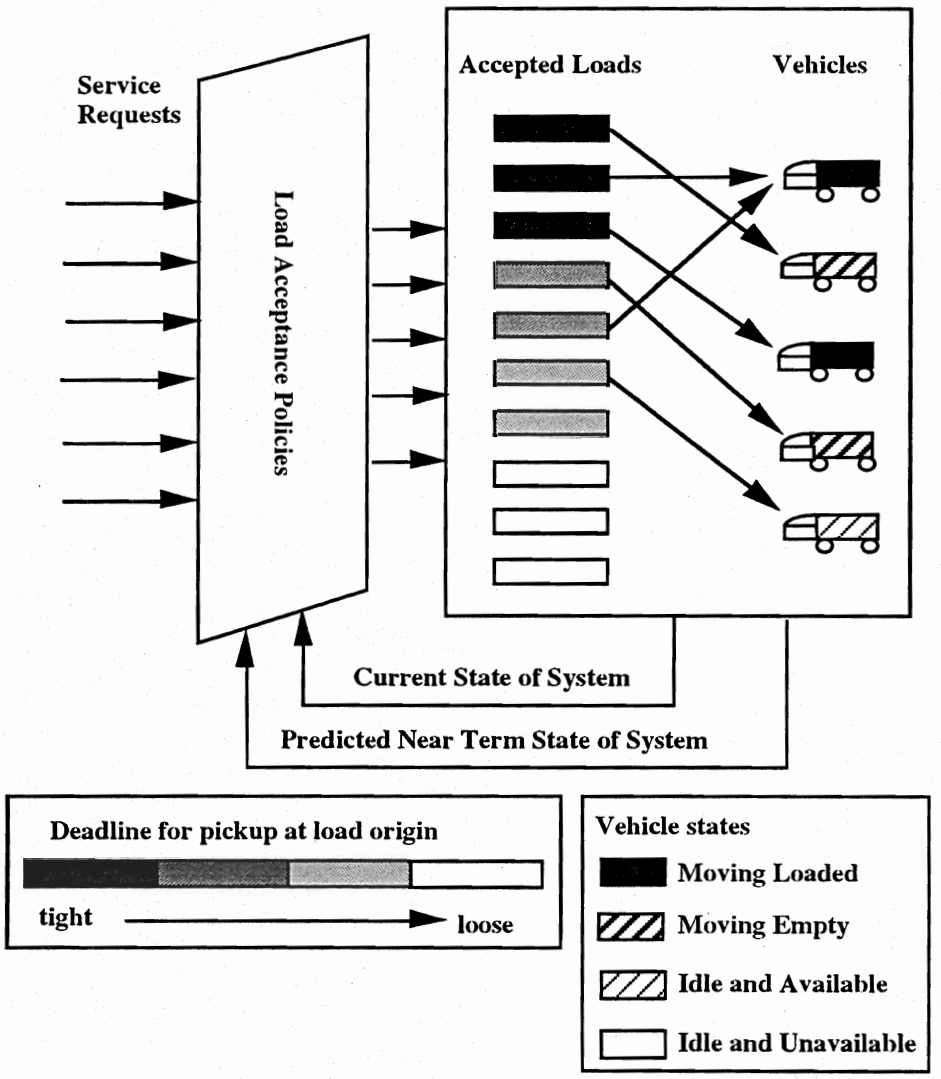


Figure 3.2 Schematic of carrier fleet operations

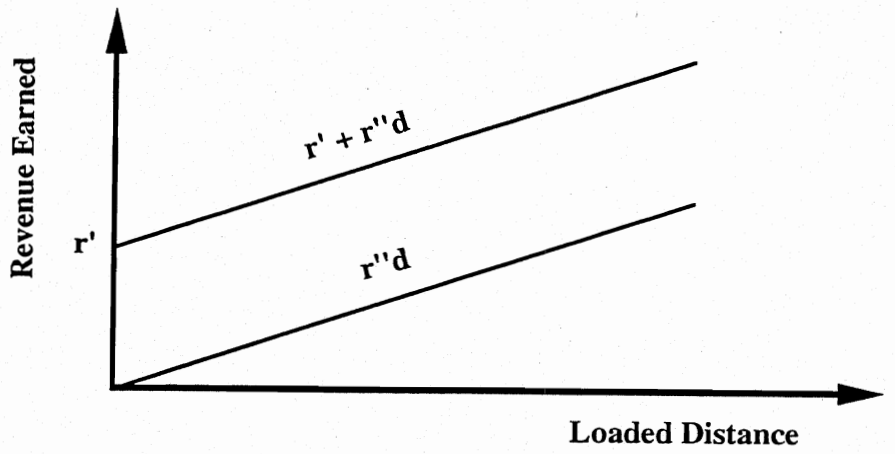


Figure 3.3 Revenue earned for each load carried

The following cost model is developed under these assumptions.

Operating Costs

Driver costs:

w_1 per unit distance for loaded moves,

w_2 per unit distance for empty moves,

w_3 per unit time spent loading and unloading,

w_4 per day

Vehicle Costs

v_1 per unit distance (loaded or empty) driven,

v_2 per day

Penalty for not servicing accepted loads:

p_1 per unit distance (length of load),

p_2 per load but not accepted for service (opportunity cost associated with each load)

Let

$$y_j = \begin{cases} 1, & \text{load j is accepted for service} \\ 0, & \text{otherwise} \end{cases}$$

$$x_{ij}^k = \begin{cases} 1, & \text{load j is served by driver k directly after load i} \\ 0, & \text{otherwise} \end{cases}$$

Letting e_{ij} represent the empty distance between the destination point of load i and the origin location of load j,

d_j the loaded distance associated with load j,

h_j the handling time associated with load j, and,

a_j and b_j the earliest and latest allowable pickup times at the origin location of load j,

the components of cost may be expressed in the following way:

$$(w_1 + v_1)\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{i=1}^{N} d_j x_{ij}^k + \tag{3.1a}$$

(loaded movement cost)

$$(w_2 + v_1)\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{i=1}^{N} e_{ij} x_{ij}^k + \tag{3.1b}$$

(empty movement cost)

$$w_3\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{i=1}^{N} h_j x_{ij}^k + \tag{3.1c}$$

(loading and unloading cost)

$$(w_4 + v_2)HK + \tag{3.1d}$$

(vehicle and driver daily costs)

$$p_1\sum_{j=1}^{N} d_j\left(y_j - \sum_{k=1}^{K}\sum_{i=1}^{N} x_{ij}^k\right) \tag{3.1e}$$

(penalty cost for loads accepted but served by out of fleet drivers) and

$$p_2\sum_{j=1}^{N}(1 - y_j) \tag{3.1f}$$

(penalty cost (opportunity cost) for not accepting a requested load)

Then the profit (loss) may be given by:

$$\sum_{j=1}^{N}(r'_j + d_j r''_j)y_j - \left[\begin{array}{l}(w_1 + v_1)\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{i=1}^{N} d_j x_{ij}^k + (w_2 + v_1)\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{i=1}^{N} e_{ij} x_{ij}^k + \\ w_3\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{i=1}^{N} h_j x_{ij}^k + (w_4 + v_2)HK + \\ p_1\sum_{j=1}^{N} d_j\left(y_j - \sum_{k=1}^{K}\sum_{i=1}^{N} x_{ij}^k\right) + p_2\sum_{j=1}^{N}(1 - y_j)\end{array}\right] \tag{3.2}$$

Under the assumption that the objective is to maximize the profit generated by the system, equation (3.2) defines an overall objective function for the operation.

A set of fairly simple performance measures related to equation (3.2) may be calculated. The revenue over the time horizon (measure 2) in table 3.1, is given by equation 3.3 and the revenue per vehicle per day may be calculated directly by dividing by HK, as in equation 3.4.

$$\sum_{j=1}^{N}(r'_j + d_j r''_j)y_j \tag{3.3}$$

$$\sum_{j=1}^{N}(r'_j + d_j r''_j)y_j \Big/ HK \tag{3.4}$$

The sum of the empty and loaded distances traveled may also be calculated as:

$$\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} d_j x_{ij}^k \tag{3.5}$$

and,

$$\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} e_{ij} x_{ij}^k \tag{3.6}$$

Under certain conditions minimizing the overall empty distances traveled is equivalent to maximizing profit generated in the system.

Dropping the general assumption that some loads may be served by drivers outside the fleet, and rewriting (3.2) to represent the cost, revenue and profits associated with a single load j which was served directly after load i and by the same driver as load i,

$$\text{Cost}_{j|i} = d_j(w_1 + v_1) + e_{ij}(w_2 + v_1) + w_3 h_j \tag{3.7}$$

$$\text{Revenue}_{j|i} = r'_j + d_j r''_j \tag{3.8}$$

$$\text{Profit}_{j|i} = r'_j + d_j r''_j - \left[d_j(w_1 + v_1) + e_{ij}(w_2 + v_1) + w_3 h_j\right] \tag{3.9}$$

Equation (3.7) sums the costs associated with the loaded distance of the move (vehicle costs per unit distance traveled and driver wages per unit loaded distance traveled), the costs associated with the empty distance traveled to make the pickup (vehicle costs per unit distance

traveled and driver wages per unit empty distance traveled), and the hourly rate earned by the driver for loading and unloading multiplied by the handling time for the load. Equation (3.8) sums the revenue associated with the load, both a fixed revenue and a revenue proportional to distance. The profit can be split further into a fixed amount which is not dependent upon the order in which the load is served and the variable amount which is dependent upon the location of the previous load's destination.

$$\text{Profit}_{j|i} = r'_j - w_3 h_j + d_j(r''_j - w_1 - v_1) - e_{ij}(w_2 + v_1) \qquad (3.10)$$

Since $r'_j - w_3 h_j + d_j(r''_j - w_1 - v_1)$ is fixed for each load, that portion of costs which comes under carrier control after a load is accepted is $e_{ij}(w_2 + v_1)$. Since $(w_2 + v_1)$ is assumed fixed, then minimizing e_{ij}, the empty distance traveled in serving load j, is equivalent to maximizing the revenue earned for that load. For a fixed set of loads, minimizing the overall empty distances traveled to provide service is equivalent to maximizing the overall profit. If the operating revenue for a fixed set of loads can be calculated then the corresponding maximum empty travel distance for which the (sub)system will be profitable may also be calculated.

Figure 3.4 illustrates the decrease in operating profit corresponding to an increase in empty distance traveled. While minimizing empty distance traveled also maximizes profit when the set of loads to be served is fixed, such a policy would simply refuse to serve all loads when loads can be rejected. Under these conditions minimizing empty distances does not result in profit maximization. In an operation that is required to reject a certain fraction of loads, or one that is free to reject less profitable loads, the ratio of empty to loaded distances traveled is a better profit indicator. This measure is calculated as in equation (3.11).

$$\frac{\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{i=1}^{N} e_{ij} x_{ij}^{k}}{\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{i=1}^{N} d_j x_{ij}^{k}} \qquad (3.11)$$

The ratio of empty to loaded distances traveled (E/L ratio) is a reasonable measure of efficiency under the assumption that loads are accepted either based on feasibility only or on rejection criteria that, while seeking to eliminate extremely poor loads, accept most feasible requests.

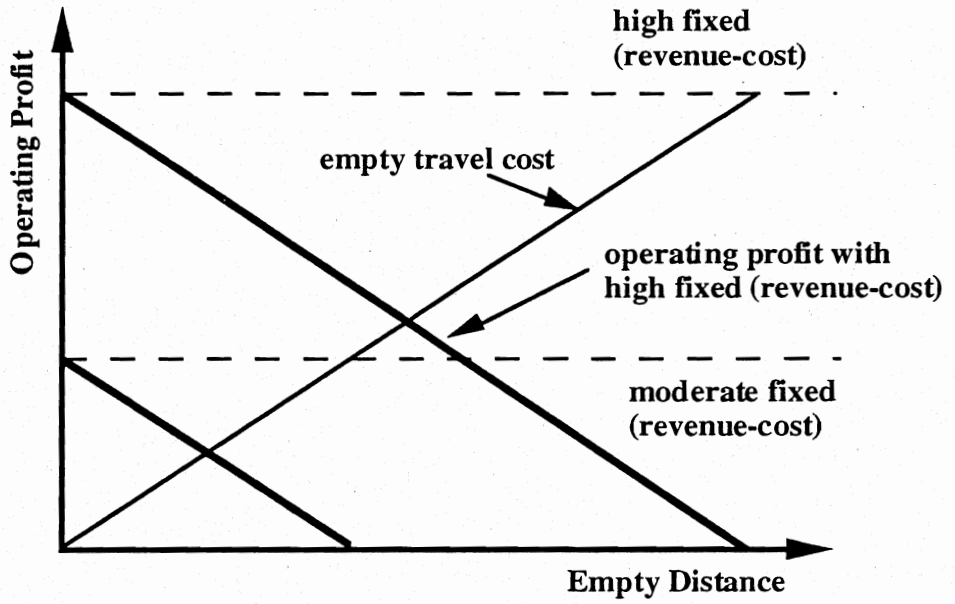


Figure 3.4 Operating profit vs. empty travel distances

A break-even point for (operating) profitability may be estimated in the following way for single load:

$$r'_j + d_j(r''_j - w_1 - v_1) > e_{ij}(w_2 + v_1) + w_3 h_j \qquad (3.12)$$

Making a simplifying assumption, namely that the handling costs are offset by the fixed revenue earned for each load moved we obtain:

$$\frac{(r''_j - w_1 - v_1)}{(w_2 + v_1)} > \frac{e_{ij}}{d_j} \qquad (3.13)$$

A further simplifying assumption that the empty and loaded driver costs (w_1 and w_2) are equal yields:

$$\text{or} \quad \frac{r''_j}{(w_2 + v_1)} > \frac{e_{ij}}{d_j} + 1 \qquad (3.14)$$

Noticing that $\sum_{j=1}^{N} d_j y_j = \sum_{j=1}^{N} d_j$ when all requested loads are accepted, and, generalizing (3.14) to a set of loads and fleet of drivers yields:

$$\sum_{j=1}^{N}(d_j r''_j)-(w_1+v_1)\sum_{j=1}^{N}d_j>(w_1+v_1)\sum_{j=1}^{N}d_j \ . \qquad (3.15)$$

Referring to

$\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{i=1}^{N}e_{ij}x_{ij}^{k}$ as E, $\sum_{j=1}^{N}r_j''$ as R", and, $\sum_{j=1}^{N}d_j$ as D yields:

$$\frac{R''}{(w_1+v_1)}>\frac{E}{D}+1. \qquad (3.16)$$

For example, if $R''/(w_1+v_1)$, the ratio of revenue earned per loaded distance traveled to the cost per distance is equal to 1.5, then for the operation to be profitable the ratio of empty to loaded moves must be no greater than 0.5. Similarly, the ratio of empty and idle time to loaded time may be estimated. Access to historical data about loads moved allows the calculation and comparison of several measures of overall effectiveness.

Furthermore, letting t_j represent the time at which the load was picked up at its origin location, it may be useful to calculate the average time between the earliest allowable pickup time and the time when the load was picked up, the maximum time after the allowable time, and the fraction of pickup deadlines missed. With {a_j, and b_j} representing the allowable window for pickup at the load origin, the average time between the actual pickup and earliest allowable pickup is represented by:

$$\frac{\sum_{j=1}^{N}(t_j-a_j)}{N} \qquad (3.17)$$

And, defining a variable $m_j = 1$ if the pickup deadline is missed, that is ($t_j > b_j$), and 0 otherwise, then

$$\frac{\sum_{j=1}^{N}m_j(b_j-t_j)}{\sum_{j=1}^{N}m_j} \qquad (3.18)$$

represents the average lateness for loads that were pickup late. These equations can be modified easily to exclude loads turned over to non-fleet drivers.

The next section presents the specifications and formulations of the dynamic fleet operational strategies of interest.

DYNAMIC FLEET OPERATIONAL STRATEGIES: SPECIFICATION AND FORMULATIONS

A diagram of a dynamic fleet management system is shown in figure 3.5. In such a system, initial vehicle to load assignments are generated, in some cases taking predictions of (near-term) future requests for service into account. Strategies for reacting to changes as they occur are incorporated in a dynamic assignment sub-system. The next section discusses the load acceptance problem and defines the three load acceptance strategies included in operational strategies (the combination of a load acceptance rule and assignment strategy) examined in the simulation experiments. Chapter 3 also discusses the real-time assignment problem and, where necessary, introduces mathematical formulations for the assignment strategies examined in simulation experiments.

The Load Acceptance Problem

Shippers with loads to be moved call a carrier requesting that a vehicle be available at a pickup location on a specific day, at a specific time, to carry a load to a specific destination. Loads may have firm or somewhat flexible pickup deadlines. A carrier may decline a load, but it may not accept a load unless it can meet the agreed upon deadlines.

Load acceptance decisions must usually be made immediately. A potential benefit of the use of electronic data interchange (EDI) technologies or electronic mail for load request is that the few minutes gained from more efficient communication methods may allow carriers to make better decisions. Carriers typically accept only loads that they believe can be served within the agreed upon time. One way to assure that pickup deadlines can be met is to identify a driver to which the requested load may be feasibly assigned. This process may be fairly simple when utilization rates are low, but when a fleet of vehicles is already working at near capacity this may be computationally difficult and time consuming. From the point of view of the development of computer aided decision tools for carrier fleet operations, several methods could be used to estimate costs and to ensure feasibility. The ability to make good feasibility and cost estimates hinges on the assumption that all accepted loads are at least temporarily assigned to a specific driver (or in some cases a small sub-fleet of drivers) so that these estimates can be based on the near future (expected) locations of drivers and vehicles.

Pool and Queue Limit Based Load Acceptance. When loads do not have explicit pickup deadlines, system and individual vehicle capacity limits (referred to as pool and queue limits,

respectively) are used to limit the average wait time for service. Operational strategies examined in this research in which loads are assigned to individual vehicle queues assume that an individual vehicle may have no more than five loads waiting in its queue, while those in which loads are assigned to a pool of accepted but unassigned loads limit the total number of waiting loads to five times the number of vehicles. While this method provides no guarantee, it can be shown that imposing such limits significantly reduces both the average wait time and associated variability of wait times for service in systems where demand exceeds service capacity (figure 6.3 in chapter 6 illustrates this point).

While keeping average service times within an "attractive" range is a reasonable goal, unless considered explicitly, wait times for service for a fraction of customers may be unacceptable even if mean wait time remains within an acceptable range. When loads do have explicit pickup deadlines or time window constraints, it may be necessary to assign arriving loads to particular vehicles and to sequence the loads assigned to minimize the distance traveled to provide service within given time constraints. The cost and feasibility based load acceptance rules included in operational strategies examined use this approach.

An alternative to such a strategy is introduced here, but not implemented in the operational strategies examined. A priority system can be maintained for the queue of arriving demands and priorities based upon pickup deadlines. Loads can be considered for assignment based solely upon their priority in the queue or, alternatively, on criteria that take both priority and cost to provide service (generally dependent on the geographic locations of the vehicle and the load origins) into account. However, such methods do not guarantee that deadlines will be met.

Pooled vs. Individual Queues. A system that provides a guarantee that pickup deadlines will be respected is one in which loads are assigned to individual vehicles based on explicit examination of feasibility upon arrival. In these circumstances, vehicles maintain individual queues. Instead of multiple vehicles and a pooled queue, such a system has multiple servers and multiple queues (figure 3.6). It is well known that average waiting times in pooled queue systems are in general lower than those in corresponding multiple queue multiple server systems. In the system of interest in this research, however, wait times under the pooled queue scenario are not necessarily less than those in the individual queue cases because once assigned, loads may be efficiently sequenced, resulting in lower service times, and loads may be assigned to individual vehicle queues with both feasibility and efficiency in mind.

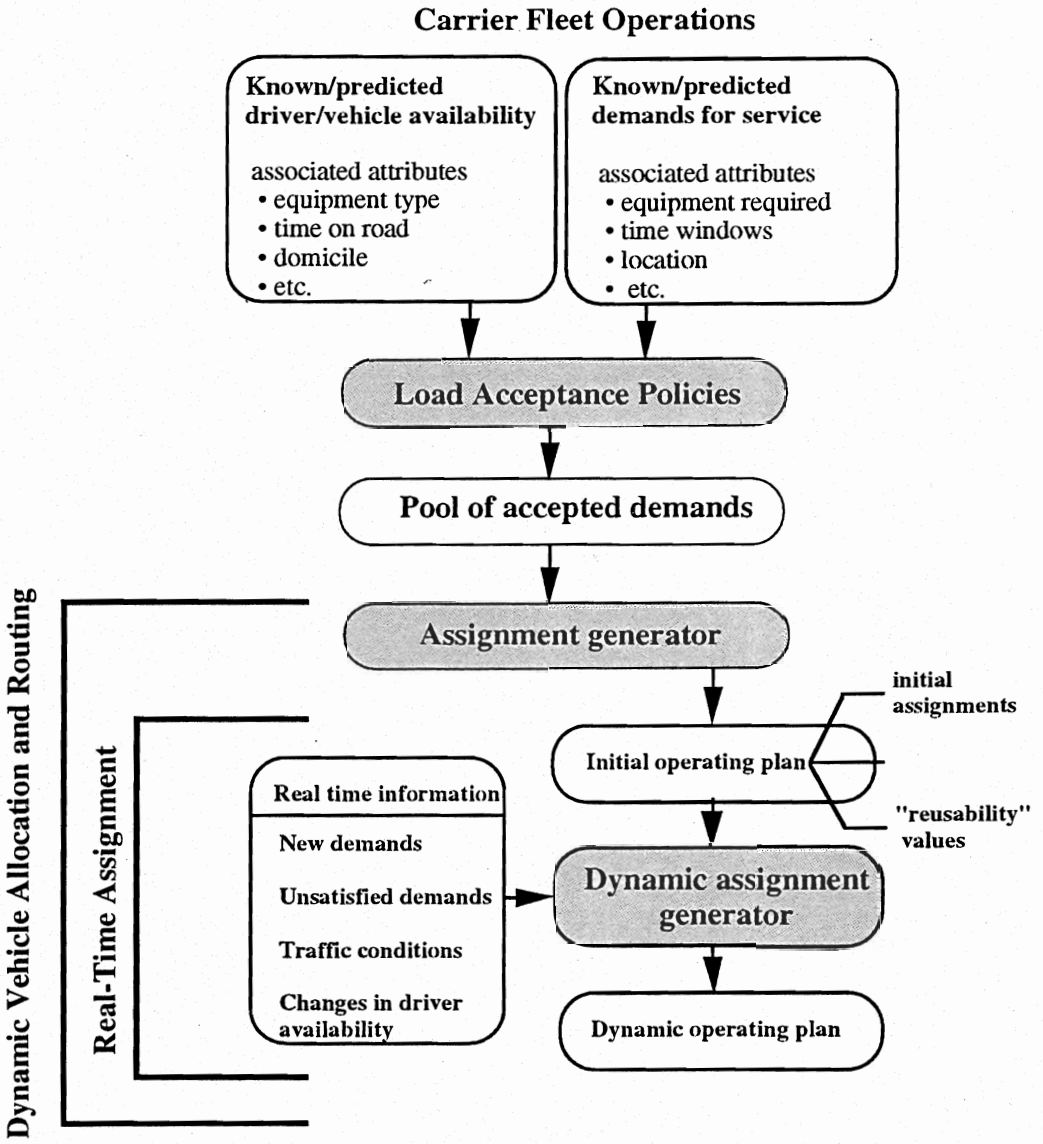


Figure 3.5 Overview of dynamic carrier fleet operations.

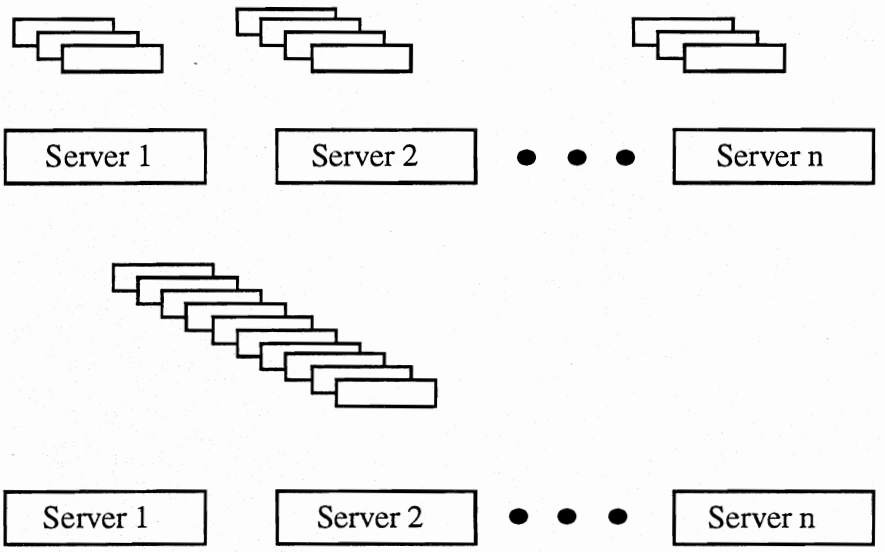


Figure 3.6 Diagram of multi-queue and pooled queue multi-server system.

While assigning loads to individual vehicles immediately upon arrival of the request for service provides a guarantee that pickup deadlines can be met, if deadlines are not binding opportunities more efficient future assignments may be lost.

An important question is how best to handle the tradeoffs between immediate assignment to a vehicle and assignment to a pool. Two scenarios, one in which loads are held in a large common pool of accepted demands until assignment to a particular vehicle shortly before service begins, and another in which most accepted demands are assigned to an individual drivers' queue are investigated. The base cases and the acceptance rules based upon system capacity alone are an example of the former while the real-time scenarios examined, in which load acceptance is based upon deadline feasibility is an example of the latter. A diagram of these is shown in figure 3.7.

Feasibility Based Load Acceptance. One method to ensure that a requested load is feasibly served relies on the identification of feasible insertion points in individual drivers' queues. Each accepted load is assigned to the most cost effective queue. These "virtual" assignments are maintained for all accepted loads, but may be changed later, assuming all service constraints can still be met. A list of feasible insertion points can be quickly generated to make a load acceptance decision.

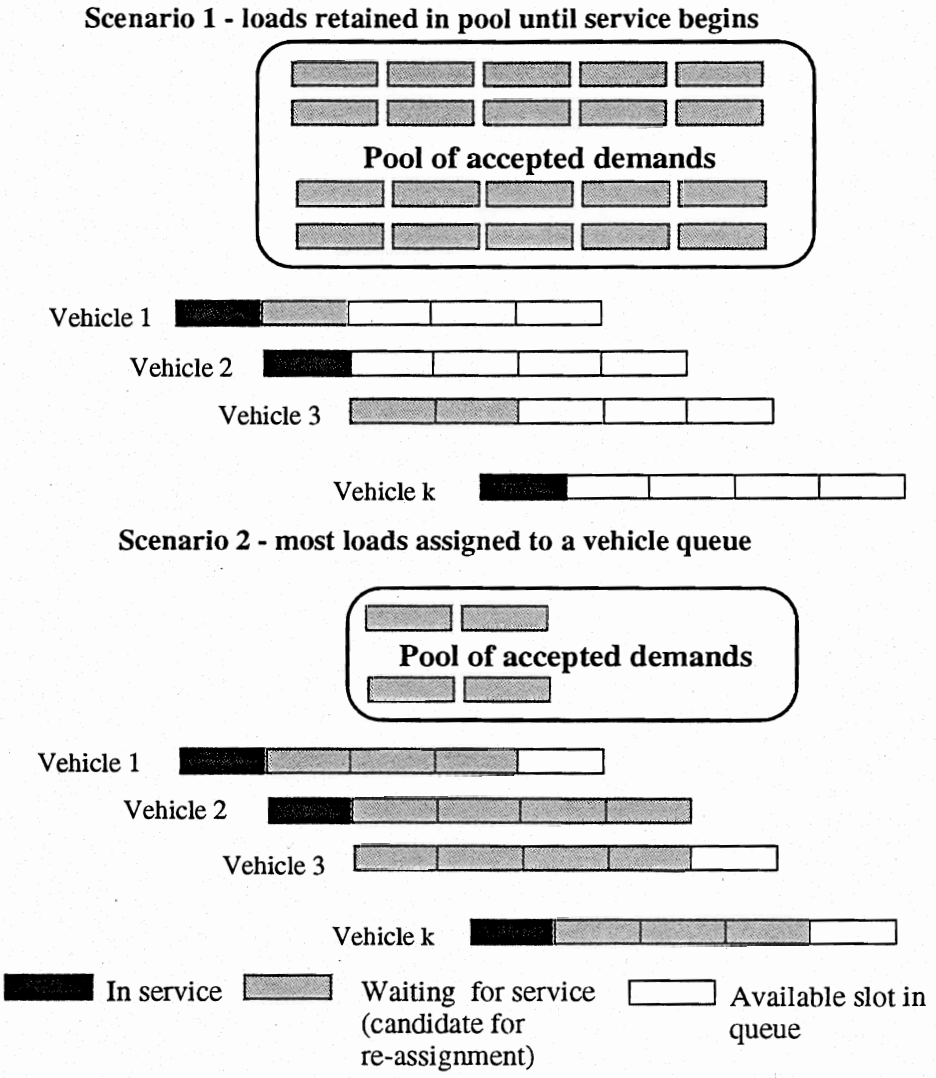


Figure 3.7 Pooled vs. individual vehicle queue assignment strategies

A more computationally intensive approach would solve for the most cost effective insertion point when loads may be re-ordered. This approach is used in the real-time operational strategies implemented in this research and is discussed in greater detail later in this chapter, in the context of driver to load assignment. A traveling salesperson problem with time windows (TSPTW) may be generated and solved for each driver likely to be a feasible choice. The mathematical formulation of this problem follows:

For each driver k in a subfleet of drivers likely to be able to accept a certain assignment, currently assigned loads are indexed by their current service order. Let:

n be the number of loads currently assigned to driver k,

c^k the cost for driver k to serve currently assigned loads (zero if no loads are assigned),

c^k_ℓ = cost for driver k to serve load l along with currently assigned loads,

e^k_{ij} = cost for driver k to serve load j after load i (empty movement cost), with e^k_{i0} = 0 for all locations i = 1,...,n+1 (no cost for the return to the starting location),

a_j = earliest pickup time (arrival time at origin) for load j,

b_j = latest pickup time (arrival time at origin) for load j,

T_{ij} = time needed to travel empty from the destination of load i (or current vehicle location, for i = 0) to the origin of load j, with $T_{i0} = 0$ for all loads i = 1,...,n+1.

τ_i = time needed to complete the loaded portion of load i.

Decision variables:

x^k_{ij} = 1 load j is served directly after load i, by vehicle k

x^k_{ij} = 0 otherwise,

t_j = scheduled time to arrive at the origin of load j.

Then

$$c^k_\ell = \min \sum_{i=0}^{n+1} \sum_{j=0}^{n+1} e^k_{ij} x^k_{ij} \qquad (3.19)$$

s.t.
$$\sum_{i=1}^{n+1} x^k_{ij} = 1 \qquad (3.20)$$

$$\sum_{i=0}^{n+1} x^k_{i\ell} = \sum_{j=0}^{n+1} x^k_{\ell i} \text{ for all loads l = 1 to n+1 and starting location 0} \qquad (3.21)$$

$$t_i + \tau_i + T_{ij} - t_j \le M_{ij}(1 - x^k_{ij}) \text{ for all loads served by driver k} \qquad (3.22)$$

where $M_{ij} = \max(b_i + \tau_i + T_{ij} - a_j, 0)$

and load ℓ is the n+1st load assigned to driver k.

The objective is to find the minimum cost feasible ordering of loads. Constraints (3.20) specify that each load must be served exactly once, while constraints (3.21) ensure flow conservation. Constraints (3.22) specify that all loads must be served within assigned time windows and ensure sub-tour elimination.

As mentioned earlier, a key assumption is that after a load is accepted for service, it is assigned immediately, at least temporarily, to an individual driver. While temporary and permanent assignments need not be identical to the assignment used to determine feasibility in the load acceptance process, the assignment decision must be made soon after acceptance so that new acceptance decisions will not jeopardize the feasibility of already accepted loads. There is no reason that load acceptance decisions, which look for at least one feasible assignment, cannot be made more than one at a time, but the approach outlined in this section requires that at most one new load is assigned to each vehicle at a time. Loads arriving at the same time that should be logically served by the same driver must either be considered sequentially or coupled in a preprocessing step. The preprocessing to couple loads can be performed manually by a dispatcher or may be automated.

Profit Based Load Acceptance. Continuing with the formulation provided above, the profitability of a load that may be feasibly served can be estimated. The cost for driver k to serve load l may be calculated as the ratio of the empty cost to revenue or the marginal cost of serving load l, in addition to already assigned loads ($c_\ell^k - c^k$). Under the assumption that operating cost is closely related to distance traveled, one proxy for the profitability of a load is the ratio of the empty distance attributable to l to the loaded distance associated with the load or, $(c_\ell^k - c^k)/d_\ell$. The load may then be accepted for service if this ratio does not exceed pre-determined or dynamically updated thresholds. For example, an underutilized fleet might be willing to move less profitable loads than one operating at near capacity. Chapter 6 addresses results of simulation experiments which investigate the effects of more or less restrictive load acceptance thresholds.

The next section discusses the real-time assignment problem and specifies strategies examined in the simulation experiments.

The Real-Time Assignment Problem

The real-time assignment sub-problem of the dynamic vehicle allocation and routing problem is concerned with assigning newly arriving loads to specific drivers and with modifying existing assignments as changes in the system occur. Assignment strategies examined in this study vary primarily in the maximum number of loads that may be assigned to a vehicle at any time and, in the extent to which existing assignments may be modified. Some permit drivers to be assigned a small queue of loads, while others restrict the assignment of loads to idle vehicles; some allow only incremental modification of existing assignments while others consider generating

completely new assignments when new demands arrive or the status of one of more driver or vehicle changes.

Flexible operational strategies which allow for the modification of existing assignments but which, in the interest of computational efficiency, do not seek to evaluate all possible assignments are of primary interest in this research. A significant operational benefit of real-time information on vehicle locations and demands, coupled with "seamless" dispatcher to driver communication, is the ability to dynamically assign vehicles to time-sensitive demands, or to recently requested loads that would be more efficiently served immediately. Two modification strategies, en-route diversion and real-time load swapping are investigated. En-route diversion is concerned with reassigning a vehicle en-route to a pickup location to provide immediate service to a more time-sensitive load, or of a load that (when sequenced first) will improve the efficiency of the vehicle's travel route. Tables 3.2 and 3.3, presented after the definition of the strategies provided in the next two sections, contrast some of the characteristics of the operational strategies examined. Base case strategies are intended to represent operations with limited real-time information requirements while the real-time operational strategies require continuous information updates and communication.

Five assignment strategies examined are considered base case strategies. A strategy is defined as a base case strategy if, once an assignment is made, it is carried out with no changes in either the vehicle assignment or the order in which service will be provided by the vehicle. The four real-time assignment strategies require continuous updates on all vehicle locations and demands and service order. As defined, only the real-time operational strategies have the ability to take pickup deadlines (or time windows for service) explicitly into account and only the real-time assignment strategies allow the re-sequencing or re-assignment of currently assigned loads.

Base Cases. Two of the five base case strategies are intended to provide a benchmark for real-time assignment systems. The "first called first served" assignment method should provide an upper bound on reasonable assignment rules while the traveling salesperson tour through points representative of those served in a week should provide a lower bound on the distance traveled to provide service.

First Called First Served (FCFS). This strategy assumes that loads are assigned to available vehicles in the order in which they arrive. Accepted service requests are added to a queue of requests upon arrival; when a vehicle becomes available it is assigned the first load in the queue. If one or more vehicles are idle when the request arrives it is assigned to the vehicle that has been idle longest. The driver must contact the dispatch center upon completion of

service and the dispatch center must be in communication with the driver that has been idle longest.

Nearest Origin Assignment (NO). Accepted loads enter the pool of unassigned loads. It is assumed that upon completion of an assignment, drivers contact the dispatch center for a new assignment. Loads arriving when one or more vehicles are idle are assigned to the nearest idle vehicle. Drivers must contact the dispatch center upon completion of service and the dispatch center must be in communication with all idle drivers.

Classical (Bipartite) Assignment. The classical (or general) assignment problem is as follows (see for example, [Ahuja, Magnanti & Orlin 1993]): Given two equally sized sets N_1 and N_2, a collection of pairs $A \subseteq N_1 \times N_2$, representing possible assignments, and a cost c_{ij} associated with each $(i,j) \in A$, the goal is to pair, at the minimum possible cost, each object in N_1 with exactly one object in N_2. In this application, the sets N_1 and N_2 represent loads and vehicles, respectively. When the number of loads is not equal to the number of vehicles, dummy loads or vehicles are added to the smaller set and assigned infinite costs.

The following is a simple formulation of this assignment problem. This problem is solved at pre-specified times for all available vehicles. Vehicle availability is defined to include vehicles currently idle or those that will become idle at a user-determined time into the next assignment stage. Loads are accumulated in a pool of accepted loads between assignment epochs. Loads not assigned because of an insufficient number of available vehicles are candidates for assignment again at the next assignment epoch. As mentioned in chapter 2, Powell [1994] discusses a similar static formulation of the Dynamic Vehicle Allocation (DVA) problem as a deterministic assignment problem. If the number of loads exceeds the number of available drivers, it may be necessary to limit candidate loads to a set of L loads which include those that have been waiting the longest or which have the nearest desired pickup times. The simulation experiments described in chapter 5 and for which results are discussed in chapter 6 are based on a simpler formulation of this problem which does not include penalties for not assigning drivers and loads. The general formulation is presented here, followed by the formulation examined in the simulation experiments:

Let L be the number of loads considered at this stage,

K = the number of available drivers (vehicles),

c_ℓ^k = cost for driver k to serve load l: relating this to earlier expressions found in section 3.3.2, $c_\ell^k = (w_2 + v_1)e_{0\ell}^k$, where $e_{0\ell}^k$ is the empty distance associated with moving from the current location of vehicle k and the origin of load l, and, w_2, v_1 are the empty driver and vehicle costs charged per unit distance, respectively.

c_ℓ^r = penalty cost of not assigning any drivers to load l,

c_h^k =holding cost for driver k (cost of not assigning driver k to any loads),

Decision variables:

x_ℓ^k = 1 if driver k is assigned load l,

x_ℓ^k = 0 otherwise,

Then the problem of assigning drivers to loads may be stated as:

$$\min \sum_{k=1}^{K} \sum_{\ell=1}^{L} c_\ell^k x_\ell^k + c_h^k \left(1 - \sum_{\ell=1}^{L} x_\ell^k\right) + c_\ell^r \left(1 - \sum_{k=1}^{K} x_\ell^k\right) \qquad (3.23)$$

subject to:

$$\sum_{k=1}^{K} x_\ell^k \le 1 \qquad \text{for } l = 1,2,\ldots,L \qquad (3.24)$$

$$\sum_{\ell=1}^{L} x_\ell^k \le 1 \qquad \text{for } k=1,2,\ldots,K \qquad (3.25)$$

$$x_\ell^k \ge 0 \qquad \text{for } l = 1,2,\ldots,L;\ k=1,2,\ldots,K \qquad (3.26)$$

Constraint (3.24) specifies that a load may be assigned at most one driver, (3.25) specifies that each driver is assigned at most one new load, while (3.26) ensures the non-negativity of decision variables. The cost c_ℓ^k may include a penalty for loads not picked up before their latest pickup time and the cost of not assigning a load

Formulation Examined in Simulation Experiments.

The simpler formulation actually implemented in the simulation experiments is the following:

$$\min \sum_{k=1}^{K} \sum_{\ell=1}^{L} c_{\ell}^{k} x_{\ell}^{k} \qquad (3.27)$$

subject to:

$$\sum_{k=1}^{K} x_{\ell}^{k} \leq 1 \qquad \text{for } l = 1,2,..,L \qquad (3.28)$$

$$\sum_{\ell=1}^{L} x_{\ell}^{k} \leq 1 \qquad \text{for } k=1,2,..,K \qquad (3.29)$$

$$\sum_{k=1}^{K} \sum_{\ell=1}^{L} x_{\ell}^{k} = \min\{L, K\} \qquad \text{for } l = 1,2,..,L;\ k=1,2,..,K \qquad (3.30)$$

$$x_{\ell}^{k} \geq 0 \qquad \text{for } l = 1,2,..,L;\ k=1,2,..,K \qquad (3.31)$$

Rather than penalize loads not picked up and drivers not assigned to loads, constraint (3.30) simply requires that either all available drivers are assigned loads when the number of loads exceeds the number of available drivers, or, that all loads awaiting service are assigned drivers when the number of available drivers exceeds the number of available loads.

Two applications of this assignment method are examined. These applications vary with respect to the assignment trigger and the subset of vehicles considered candidates at each assignment period. The fundamental strategy is as follows: 1) Loads accumulate over time in a pool of accepted loads; 2) At decision points loads are assigned to idle vehicles or in some cases vehicles that will become idle in the near future. Exactly m assignments are made where m = min{ available loads, available vehicles }; the assignment that minimizes the overall distance from the current (or next available) location of the vehicles to the origin locations of the loads is chosen.

Time triggered bipartite assignment (BAT(a)). Assignments are triggered by fixed, evenly spaced assignment points. Fixed assignment points are separated by a length of time aDL where a is a real number in the interval (0, 2) and DL is the average duration of loaded moves. Figure 3.8 provides a time-line highlighting assignment points in this case.

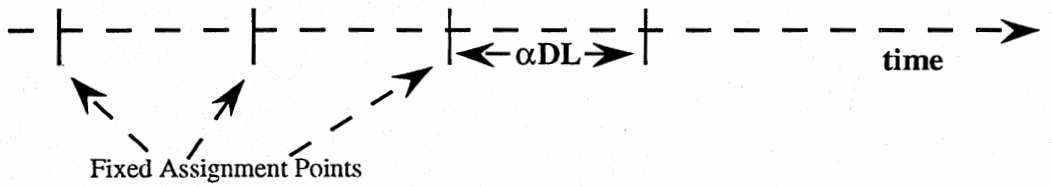


Figure 3.8 Diagram of assignment points

This method may be applied with a look ahead strategy in which, in addition to considering currently idle vehicles for assignment, vehicles that are expected to become idle a specified fraction of the way between the current assignment point and the next assignment point are included in the set of candidate vehicles.

State Triggered Bipartite Assignment (BAS(b)) In state triggered or state-based bipartite assignment, assignments are made when the number of loads awaiting service is equal or greater than a multiplier b times the number of idle vehicles, or, when the number of idle vehicles is equal to the same multiplier times the number of waiting loads: for example, when the number of loads exceeds {0, 1, 2,..} times the number of idle vehicles.

Asymmetric TSP (ATSP) The solution of a Multiple (Asymmetric) Traveling Salesperson Problem provides a lower bound on cost of solutions in the real-time operational strategies examined. Rather than solve the MTSP, an asymmetric Traveling Salesperson Problem is solved for a set of loads and a single vehicle. This case is used to generate a benchmark for single vehicle operation of the real-time operational strategies examined. A problem based on a set of randomly generated loads of approximately the same number of loads served per vehicle per week is solved and the expected distance traveled under a perfect hindsight (or perfect look-ahead) assignment estimated. The objective is simply to minimize the empty distances traveled. This bound is compared to the performance in systems in which loads become known over time.

One formulation of the TSP problem is as follows:

Let n+1 be the number of loads to be ordered.

e_{ij} = cost to serve load j after load i or, where i = 0, is the current location of the vehicle (empty movement cost) and e_{i0} = 0 for all loads i = 1,..,n+1.

Decision variables:

$x_{ij} = 1$ load j is served directly after load i

$x_{ij} = 0$ otherwise,

The objective is to find

$$\min \sum_{i=0}^{n+1} \sum_{j=0}^{n+1} e_{ij} x_{ij} \qquad (3.32)$$

s.t.

$$\sum_{i=0}^{n+1} x_{ij} = 1 \qquad (3.33)$$

$$\sum_{i=0}^{n+1} x_{i\ell} = \sum_{j=0}^{n+1} x_{\ell i} \text{ for all loads } \ell = 1 \text{ to } n+1 \text{ and starting location } 0 \qquad (3.34)$$

The objective is to find the minimum cost ordering of loads. Constraint (3.33) specifies that each load must be served exactly once, while constraints (3.34) ensure flow conservation.

Information Requirements, Advantages and Disadvantages of Base Case Assignment Strategies. The information requirements of the five base case strategies vary. Under FCFS assignment, the driver must communicate with the dispatch center upon completion of service. The dispatcher need not know the current location of the driver to make the next assignment. If no loads are available when the driver completes service, then the dispatch center must be able to contact idle drivers with new assignments in the order in which they became idle.

Under NO assignment, the driver must initiate contact with the dispatch center upon completion of service and the dispatch center must know the driver's location at that time. If no loads are available when the driver completes service, the dispatch center must be able to contact all idle drivers when loads arrive to the system.

Under time-triggered bipartite assignment, the locations of all available drivers (vehicles) must be known before each assignment step so that costs may be estimated. Whenever a vehicle becomes available to move another load its location must be known and it must be in communication with the dispatch center in order to receive directions to pick up a new load. Continuous driver to dispatcher communication is generally not needed, since drivers need only communicate when they are ready for a new assignment or at pre-specified assignment times. Furthermore, if look ahead is allowed the location and status of each vehicle must be known (or, predicted with a high degree of certainty) at each assignment period. Because of its additional information needs, BAT(a) with look ahead could be termed a quasi real-time assignment method.

The ATSP solved to provide a benchmark is a purely static assignment method. All loads are known at the beginning of the assignment period. A tour is formed and the vehicle is dispatched.

Under time-triggered bipartite assignment, when near-idle vehicles are included in the assignment stage, the locations of vehicles must be known at all times. The dispatch center must have continuous communication with idle drivers. State-triggered bipartite assignment is also a quasi real-time assignment method. However, since once assigned, neither the order in which loads are served nor the driver to which they are assigned are open to change, state-based assignment is included in the set of base case assignment strategies.

The primary advantage of the FCFS and NO assignment methods is their simplicity. No real-time information other than the location of an available driver when the driver is ready for the next load is needed. Since a driver will typically contact the dispatch center upon completion of service (or upon arrival at the load destination), this information is readily available. Time-based bipartite assignment, with no look ahead, and with assignment periods separated by a reasonable length of time has light information and communication requirements also. The ATSP is included merely because it is used to generate a lower bound on the average distance traveled to provide service under other strategies.

Another advantage of the classical assignment method is that despite its simplicity a fairly high degree of operational realism may be expressed. Powell [1994] discusses this aspect of the simple static assignment problem for the DVA and describes many ways that driver and dispatcher desires can be incorporated into the costs. Holding costs may be applied to reduce the likelihood that a particular driver is idle, or applied across the board if keeping all drivers working is more important than reducing empty distances traveled. Since drivers are modeled individually, preferences for assigning a driver or subset of drivers to certain loads may be easily incorporated.

The primary drawbacks of all but the ATSP base cases are that only one new load is assigned to a driver at a time and that there is no guarantee or even expectation that pickup deadlines will be respected. In addition, it can be shown that under some conditions (moderate to low demand, for example) the ordering of even a small number of loads can lead to reduced empty distances traveled. Base case assignment methods preclude the generation of such "routes". The first called first served process is clearly not efficient and is presented as a benchmark case. The nearest origin assignment can be shown to be very efficient in the limiting case where the number of loads to be served is very high. However, this assignment method performs less well at typical congestion levels. In addition, in all but the ATSP case the decision

rule takes only the length of the next empty move into account when assigning vehicles, leaving drivers open to possibly long empty moves after completion of the assigned load.

Conditions under which the nearest origin assignment performs fairly well are examined in chapter 4. Simulation experiments designed to analyze the performance the five base case assignment rules are described in chapter 5, and their results are presented in chapter 6.

Assignment Under Real-Time Information. In this section, a formulation for a modified bipartite assignment problem in which service time constraints are met is introduced. The following assumptions are made. A feasible and cost efficient assignment of drivers to loads has been constructed for an initial set of accepted requests for service. If there are more service requests than available vehicles then some vehicles may have an associated pseudo-route of assigned loads. As new service requests are accepted or changes in driver and vehicle availability occur, these new loads are added to *current* driver to load assignments.

The formulation is very similar to the one introduced in chapter 3 except for one important modification of the cost function to include previously assigned loads. Specifically, the cost function term c_ℓ^k is now the cost for driver k to serve load l and currently assigned loads, whereas previously it only applied to load l. The terminology and formulation are otherwise identical to those of chapter 3.

The main difference in the underlying assumptions is that drivers may already have a set of assigned loads. This formulation allows for the insertion of a load into a driver's queue of assigned demands. The cost c_ℓ^k represents the cost for driver k to serve all currently assigned loads *and* the candidate load. Holding costs c_h^k are equal to zero for drivers with other assignments and non-zero for idle drivers.

Assignment costs, c_ℓ^k, must be updated whenever changes in the system occur. These are updated by calculating the cost of inserting candidate loads into each driver schedule, given the current location and status of the vehicle. The assignment costs come from the solution of another sub-problem. For each driver and each candidate load, the least cost (time window feasible) tour must be found. This problem may be formulated as an integer linear program and is an instance of a traveling salesperson problem with time windows (TSPTW) with asymmetric costs (the cost to serve load i after load j is not the same as the cost to serve load j after load i). The basic formulation is modified to take into account the fact that the origin and destination location for loads are different, and for the associated loaded travel time. For a detailed discussion of the general (m-vehicle) VRPTW and TSPTW see for example, Desrosiers et al. [1995].

Formulation Examined in the Simulation Experiments. As in the last section, a simplified adaptation of this formulation is examined in a simulation framework. Only one load is considered for assignment at a time. The TSPTW sub problem is solved by complete enumeration for each vehicle's current assignments and the candidate load. In each case, pickup deadlines are viewed as hard rather than soft constraints. En-route diversion is possible under each of these rules. When en-route diversion is allowed the current queue of assigned loads will include the first load in the queue, assuming the load had not been picked up: otherwise the current queue will include all but the first load in the queue. When en-route diversion is allowed this strategy is referred to as DR^C; when en-route diversion is not allowed it is referred to as D^CR^C. This is the case regardless of the local assignment rule used to make the final assignment.

Rule 1) Least empty to loaded ratio assignment

Let c_ℓ^k represent the empty distance associated with the least empty distance, deadline feasible ordering of loads currently assigned to vehicle k and candidate load l. Let d_ℓ^k represent the corresponding loaded distance. Then, the candidate load l is assigned to the vehicle for which $\left\{ c_\ell^k / d_\ell^k \right\}$ is lowest.

Rule 2) Least additional distance assignment

Let c^k represent the empty distance associated with the queue of order loads currently assigned to vehicle k. Again, c_ℓ^k represents the empty distance associated with the least empty distance, deadline feasible ordering of loads currently assigned to vehicle k and candidate load l. Then, the candidate load l is assigned to the vehicle for which $\left\{ c_\ell^k - c^k \right\}$ is lowest.

Rule 3) Least overall empty distance assignment

The candidate load is assigned to the vehicle for which c_ℓ^k is lowest.

The relative merits of these three decision rules are examined under simulation. Experiments performed are outlined in chapter 5 and results presented in chapter 6.

Information Requirements, Advantages and Disadvantages of Approaches Allowing En-route Diversion But Not Re-assignment of Loads. Assignment strategies allowing en-route diversion require continuous driver to dispatch center communication in addition to real-time vehicle location and status updates.

The current state of the driver (and vehicle) is taken into account; the current location of the driver is taken as node zero and constraints (3.34) are modified so that t_0, the arrival time at node zero, is assumed to be the current time. For en-route (and not divertable) drivers and loaded drivers this time may be taken as the expected time at the destination location of the current load and the corresponding location information may be updated accordingly. Similarly, this formulation permits the diversion of en-route vehicles without increasing the complexity of evaluation of alternatives. Since assignment costs are based on the current location of the vehicle, no distinction need be made between en-route and idle vehicles.

This approach also presents some difficulties. First, it only allows the assignment of one new load to a driver at a time. While this restriction increases the likelihood of finding an optimal solution for the problem quickly, the optimized problem is a local, rather than a global problem. Examples may be readily constructed to illustrate the fact that the most efficient solution would assign more than one new load to a single driver. However, if the initial assignments are made well, it may be that the addition of loads to drivers' schedules in real-time will produce good solutions. In addition, if it is clear that a set of loads ought to be served by the same driver, then these loads could be combined into a single load. Second, although en-route diversion is considered, and the order in which loads are served (within time constraints) is considered flexible, this strategy does not consider the re-assignment of previously assigned loads to other vehicles. Examples in which the most cost effective and efficient solution would remove an assignment from a driver and divert the driver to make another pickup can easily be constructed. Finally, this formulation does not explicitly identify empty repositioning moves for idle drivers.

In the next section, a dynamic assignment strategy which allows re-assignment of loads from one vehicle to another is presented.

Dynamic Assignment and Re-assignment

As in the previous section, the assumption is made that a feasible and cost efficient assignment of drivers to loads has been constructed for an initial set of accepted requests for service. The difference here is that as demands arrive or changes in driver and vehicle availability occur, previously assigned loads are re-evaluated. This re-evaluation is performed in two ways. The most flexible approach views the load assignment problem as a multi-vehicle TSPTW problem like the single vehicle TSPTW outlined earlier. This m-TSPTW varies from the standard definition in that vehicles are not assumed to begin at a single location. The basic problem is as follows:

Let n+1 be the total number of loads to serve.

e_{ij}^k = cost to for driver k to serve load j after load i, with e_{0j}^k is the cost of moving from the current location of driver k to the origin location of load j, and $e_{i0}^k = 0$, for all loads i = 1,..n (the cost of returning to the starting location of driver k).

a_j = earliest pickup time (arrival time at origin) for load j,

b_j = latest pickup time (arrival time at origin) for load j,

T_{ij} = time needed to travel empty from the destination of load i (or current vehicle location (i = 0)) to the origin of load j.

τ_i = time needed to complete the loaded portion of load i.

This time may also include loading and unloading time associated with load i.

Decision variables:

$x_{ij}^k = 1$ load j is served directly after load i, by vehicle k

$x_{ij}^k = 0$ otherwise,

t_j = scheduled time to arrive at the origin of load j.

Then

$$c_\ell^k = \min \sum_{i=0}^{n+1} \sum_{j=0}^{n+1} e_{ij}^k x_{ij}^k \qquad (3.42)$$

s.t.
$$\sum_{k=1}^{K} \sum_{i=0}^{n+1} x_{ij}^k = 1 \qquad (3.43)$$

$$\sum_{i=0}^{n+1} x_{i\ell}^k = \sum_{j=0}^{n+1} x_{\ell i}^k \text{ for all loads l = 1 to n and starting location 0} \qquad (3.44)$$

$$t_i + \tau_i + T_{ij} - t_j \le M_{ij}(1 - x_{ij}^k) \text{ for all loads served by driver k} \qquad (3.45)$$

where $M_{ij} = \max(b_i + \tau_i + T_{ij} - a_j, 0)$

The objective is to find the minimum cost feasible assignment of loads. Constraints (3.43) specify that each load must be assigned to exactly one vehicle's route, while constraints (3.44) ensure flow conservation. Constraints (3.45) specify that all loads must be served within assigned time windows and ensure sub-tour elimination.

Information Requirements, Advantages and Disadvantages of Approaches Allowing Both En-route Diversion and Re-Assignment of Loads. This approach requires continuous driver to dispatch center communication in addition to real-time vehicle location and status updates. It allows significant changes in assignments to be made as new demands arrive and traffic network conditions change. It requires no more information than the approach outlined in the previous section. The current state of vehicles may be easily taken into account by the addition of K dummy loads, one for each driver. For idle or en-route (but empty) drivers the pickup time at this dummy node is taken to be zero (as is the loaded time for the load) and for loaded drivers this time is taken to be the expected time at the destination for the load currently in service.

Although there has been a significant effort among researchers to develop techniques to solve VRPTW and m-TSPTW problems efficiently in the past few years, the VRPTW is well known to be *NP*-complete. Even small instances of this problem are difficult and time-consuming to solve. This approach does not explicitly consider the repositioning of idle vehicle, this must be tackled in a separate step A more realistic approach would involve a load swapping heuristic which would intelligently chose loads to include in the decision process, or at the very least would identify a small subset of vehicles and loads to be candidates for load-swapping.

Alternative Approach Examined in Simulation Experiments. A less computationally complex re-assignment method identifies loads with flexible pickup deadlines and returns them to the pool of demands for re-assignment. A difficulty is that if more than one load is returned to the pool, then new assignments could be infeasible, if the decision process is not reversible. A method in which loads are returned to the pool and immediately re-assigned to the current best vehicle is shown, in Chapter 6, to be surprisingly effective in improving assignments. The success of this purely local, and not particularly intelligent re-assignment method in which feasibility is maintained at all times suggests that more sophisticated methods should be examined and that route improvement techniques, well known in the vehicle routing literature, offer significant promise in this application.

In simulation experiments this strategy is referred to as DR, when en-route diversion is allowed, in addition to re-assignment, and as $D^{C}R$ when en-route diversion is not allowed.

TABLE 3.2 CHARACTERISTICS OF BASE CASE OPERATIONAL STRATEGIES

Characteristics of Base Case Strategies		
Assignment Method / **Characteristics**	**BAT(α)** **BAT(α - with look ahead)**	**FCFS** **NO** **BAS(β)**
Information/ Communication Requirements	Location of idle vehicles at assignment points. Ability to communicate with idle vehicles at assignment points. Under *look ahead*-location of all vehicles at assignment points, ability to communicate with vehicles as they become idle.	Location of idle vehicles, ability to communicate with idle vehicles at all times.
Assignments triggered by	Fixed assignment periods, predicted future availability	State of system, driver availability
Load acceptance methods	Based on estimate of system capacity	
Length of planning horizon	Length of longest load currently in service (typically less than one day)	
Management of accepted requests	Demand pool	Demand pool, or ordered queue
Resequencing possible?	no	no
En-route diversion possible?	no	no
Re-assignment of loads possible?	no	no

TABLE 3.3 CHARACTERISTICS OF REAL-TIME OPERATIONAL STRATEGIES

Characteristics of Real-Time Assignment Strategies		
Assignment Method / Characteristics	D^CR^C, DR^C	D^CR, DR
Information/ Communication Requirements	Continuous status and location information for whole fleet and continuous driver to dispatcher communication	
Assignments triggered by	Start of period (i.e. day), load acceptance, changes in driver or vehicle availability	
Load acceptance methods	Feasibility guaranteed by (pseudo)route construction, estimated profitability of providing service based on currently accepted loads	
Length of planning horizon	Time until all currently accepted loads are served (typically one to two days)	
Management of accepted requests	Immediate assignment to individual vehicles	Immediate assignment to individual vehicles, subject to change
Resequencing possible?	yes	yes
En-route diversion possible?	yes	yes
Re-assignment of loads possible?	no	yes

SUMMARY

This chapter has introduced the conceptual and theoretical framework for the analysis of dynamic dispatching strategies for carrier fleet operations under real-time information. The problem has been defined and the assumptions made explicit. A simple cost model and the motivation for its specification is described, as are the evaluation criteria for the performance of a real or simulated dispatching system. Mathematical formulations for the three load acceptance and nine assignment strategies, components of the operational strategies examined in the simulation experiments of chapters 5 and 6 are provided. The next chapter provides analytical examinations of en-route diversion and an estimate of the increase in the availability of a fleet of vehicles to respond to time-sensitive demands under real-time information. In addition, a model of carrier fleet operations as an M/G/k queue is presented.

CHAPTER 4 ANALYSIS OF CARRIER FLEET OPERATIONS UNDER REAL-TIME INFORMATION

A central focus of this research is to identify and test ways in which operations might change in order to take advantage of real-time information. In this chapter, idealized instances of the driver to load (and load to driver) assignment problem are examined. These provide insight into dynamic dispatching strategies that are outside the current practices of typical carrier fleet operations. The common thread in the analyses of this chapter is the examination of how, in a system in which customer requests arrive stochastically over time and space, to provide service within a reasonable length of time to all customers while at the same time maintaining the flexibility to respond to requests that require immediate service. Chapter 4 discusses a strategy of diverting a vehicle en-route to make an immediate pick-up of a more time-sensitive load, or of a load that (when sequenced first) will improve the efficiency of the vehicle's travel route is introduced. Strategies allowing en-route diversion are examined, beginning with the operations of a single vehicle. Moving from a single vehicle to fleets of various sizes, the increase in the ability of a fleet to respond to time-sensitive demands under real time information is estimated, again with the help of simplifying assumptions. Related to this is the issue of how congestion affects the ability of a fleet to respond quickly to requests for service.

Queueing models provide a natural approach to analyzing service systems under congestion. Of considerable importance in dynamic fleet management is how congestion and the spatial and temporal variability of demands should be managed (Psaraftis [1988], Powell, Odoni and Jaillet [1995]). Like many spatially distributed service systems, heavily utilized systems offer sometimes significant economies with respect to distances traveled to provide service. However, customers must still be served within a "reasonable" length of time. A queueing model of the system may be used to estimate target congestion/utilization levels for the system. These target values can be used in the load acceptance decision process and in the assignment of loads to individual vehicles or sub-fleets. A system with Poisson arrivals and a general service distribution is examined. An approach for generating an upper bound on the efficiency (measured as the average wait time for service to begin) for a fleet of vehicles is developed and the conditions under which this bound is relatively tight are discussed.

The analysis begins with an M/G/1 (single server) system in which service times are approximately[1] independently and identically distributed (IID) variables. The relationship between

[1] Service times are only approximately IID because consecutive customers share a geographic location - the destination point of the last load served is the starting point for the empty move of the next load served.

such a system and systems in which there exist systematic dependencies between service times and the number and geographic locations of customers waiting for service is examined. It is shown that while analytic approximations for M/G/1 systems with independent travel times do not provide accurate estimates of the performance of the kinds of fleet assignment strategies of interest here, that they do provide bounds on the efficiency of such systems. This relationship is demonstrated directly for a single vehicle and generalized for the multi vehicle, M/G/k system. The discussion in this chapter is limited to assignment rules in which vehicles are assigned at most one load at a time and in which pickup deadlines, if they exist, are not honored explicitly. Strategies in which vehicles receive multiple load assignments and those addressing time windows and pickup deadlines are discussed in chapters 5 and 6. Simulation results related to the analysis introduced are examined in chapter 6. Limited simulation results are presented as well, as that analysis does not fit neatly into the experimental framework outlined in chapter 5. Limited simulation findings are necessary for the discussion of assignment strategies that lack the (approximate) IID assumption for the service times.

INTRODUCTION TO EN-ROUTE DIVERSION

Because of the length of some empty moves made to pick up loads, it is possible that new information on demands to be serviced may arrive while a driver is en-route to a pick-up. Assuming time windows for movements are flexible, this new demand information may be used to order demands in such a way as to reduce empty miles driven. Quasi continuous dispatcher to driver communication makes possible the en-route diversion of a driver moving to a pick-up location to an alternative load, thereby inducing a re-sequencing or re-assignment of the original load. Such diversion strategies are not generally feasible under current operations because dispatcher-driver communication takes place at discrete instances only, typically at a load pick-up or delivery point.

The relative improvement possible under this strategy depends on the relative locations of the alternative pick-up and delivery points. Under some distributional assumptions about the locations of these points, we are interested in the probability that diverting the driver to a new demand while en-route to a previously assigned pick-up will be beneficial. Results and related insights from the investigation of single vehicle diversion strategies are discussed in the next section.

Section 4.3.2 includes a discussion of the extent of the lack of independence of consecutive service times and the impact of the approximation on the analysis.

Diversion Probabilities Under Simple Assumptions: A Single Vehicle

Beginning with the most basic case, while a driver is en-route to a load origin, information about another load (and in this initial case, only one other load) to be moved becomes available. The following questions are addressed: what is the probability, given various diversion decision rules, that the driver will be diverted to serve the new load first? What is the probability that following such diversion decision rules will result in a reduction of overall distance traveled? And, what is the associated expected reduction in distance traveled?

To clarify, consider in Figure 4.1, a vehicle that begins at the center, c, of a circle and moves toward the origin of a loaded movement between points x_1 and x_2, where these points are uniformly and randomly generated over the area of the circle. Given a diversion point (the point at which another load to be moved becomes available) some fraction of the distance from the center of the circle and origin x_1, the probability that the distance between the diversion point to a new origin x_3 will be less than the distance from the diversion point to origin x_1 is first derived. Let a, $0 \le a \le 1$ denote the fraction of the distance from the center to x_1 traveled to reach the diversion point. The probability that the distance from the diversion point to the new origin is less than to the old origin is given by

$(1\text{-}a)^2/2$, as shown hereafter.

Let B(c,r) denote the circle of center c and radius r, and d(x,y) the Euclidean distance between points x and y. Consider two random points in B(c,r), say x_1 and x_3. For $0 \le a \le 1$, let $W_1(a)$ be the point on the segment (c, x_1) such that $d(c,W_1(a)) = a(d(c, x_1))$. Define the following two random variables $Y_1 = d(W_1(a), x_1)$ and $Y_2 = d(W_1(a), x_3)$, where Y_1 and Y_2 represent the distances from the potential diversion point to the current and potential load origins.

Let Z be the radial distance of $W_1(a)$ so $Z = d(c,W_1(a))$ and $f_Z(\bullet)$ be its probability density function. With $Z/a = d(c, X_1)$, $f_Z(\bullet) = 2z/a^2$.

$$P(Y_2 < Y_1) =$$

$$\int P(Y_2 < Y_1 | Z = z) f_z(z) dz = \int P(x_3 \in B(W_1(\alpha), z/\alpha - z)) f_z(z) dz \qquad (4.1)$$

Since $W_1(a)$ is a random point in B(c,a),

$$P(Y_2 < Y_1) = \int_0^{\alpha} ((1-\alpha)z/\alpha)^2 (2z/\alpha^2) dz = (1-\alpha)^2/2. \qquad (4.2)$$

Under a myopic strategy of diverting to the new demand origin, x_3, if it is closer to the diversion point than origin x_1, $(1-\alpha)^2/2$ represents the fraction of loads for which diversion is selected. This probability ($P(Y_2 < Y_1)$) is shown graphically as a function of the diversion point location parameter, a, in Figure 4.2. However, under this myopic strategy, even if we evaluate

the diversion decision at point a = 0, the resulting average savings (determined through simulation of the system described) in terms of reduced distance traveled while serving the two loads is less than one percent, and, diverting at points further downstream actually results in a slight increase in traveled distance, on average.

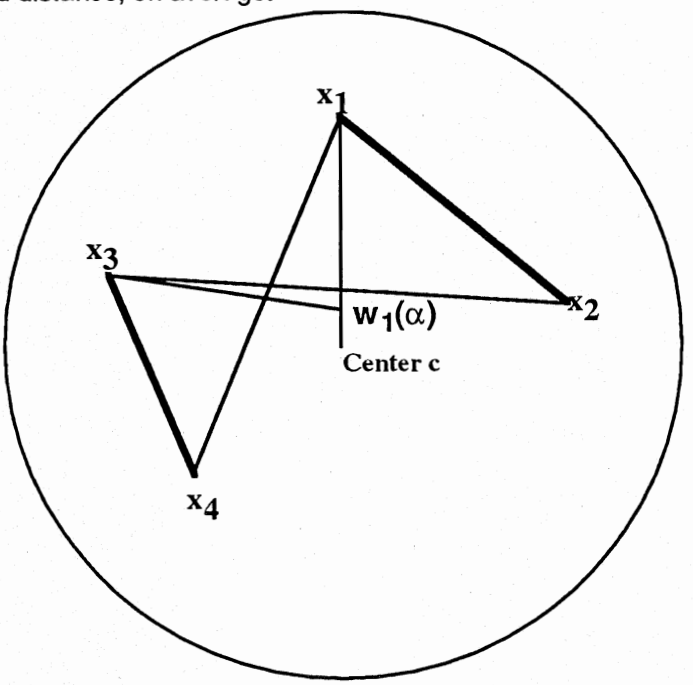


Figure 4.1 Diversion example

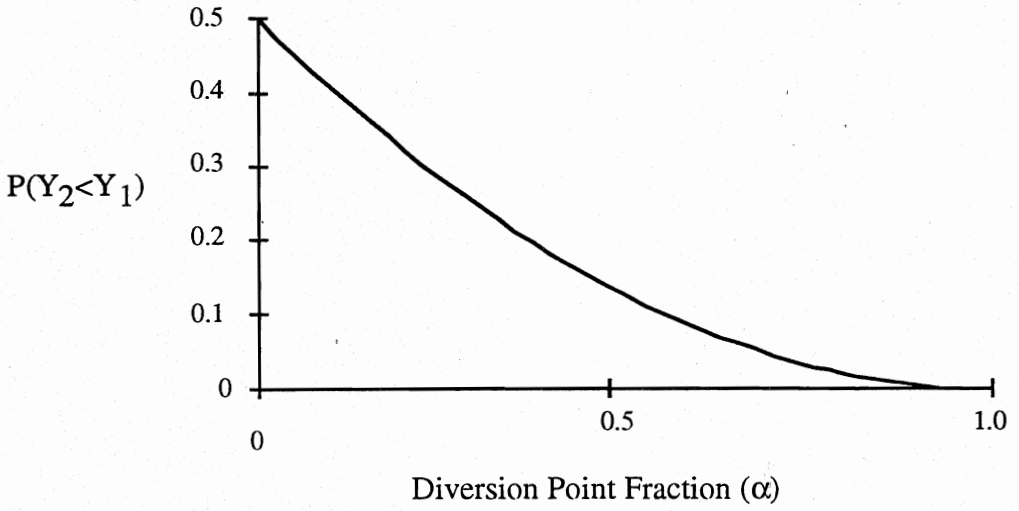


Figure 4.2 Probability of diversion under myopic strategy: $P(Y_2 < Y_1) = (1\text{-}a)^2/2$

A more plausible diversion strategy would also consider the relative distances between the destination point of the first movement and the origin point of the next load. In Figure 4.1 these are given by $d(x_2,x_3)$ and $d(x_4,x_1)$. In this case diversion is chosen if,

$$d(p,x_3) + d(x_4, x_1) < d(p,x_1) + d(x_2, x_3). \qquad (4.3)$$

While it is possible, using analytic methods, to derive insights into the behavior of certain strategies (diversion strategies for example), the fact that each move is dependent upon the moves that precede it make such derivations impractical under all but the simplest assumptions. For this reason, performance measures, including the probability that employing diversion will be beneficial, are evaluated through simulation of such diversion strategies over service horizons of varying lengths, under different arrival stream distributions, and under load acceptance rules that either require all loads to be serviced, or allow less profitable loads to be rejected. The scenarios examined are not intended to replicate actual operating conditions, but to provide a simplified representation that allows derivation of basic insights into the potential benefits of real-time information and the factors that affect these benefits. In addition, this examination is intended to assist in the identification and design of strategies that merit evaluation under more realistic operating conditions. An examination of the effects of allowing en-route diversion are discussed in Chapter 6.

ABILITY OF FLEET TO RESPOND TO TIME-SENSITIVE DEMANDS

A primary operational benefits of real-time information on vehicle locations and demands, coupled with "seamless" dispatcher to driver communication, is the ability to dynamically assign vehicles to time-sensitive demands, or to recently requested loads that would be more efficiently served immediately. The diversion strategy is explored for vehicle fleets in several contexts.

Figure 4.3 depicts a time history of the operational states of each vehicle in a given fleet. Four states are possible: moving loaded, moving empty, idle and available, and idle and unavailable.

In figure 4.3, assume that a new load has been accepted at the time marked by the first of the two vertical lines. With en-route diversion allowed, three of the vehicles shown would be considered candidates for immediate dispatch to the load. Without en-route diversion, only vehicle 3 in the diagram would be a candidate, since the other vehicles are moving toward pick-up points or moving loaded. It can be observed from this figure that the possibility of re-assigning vehicles during empty travel states can offer a significant increase in operational flexibility by increasing the availability of vehicles for real-time diversion or re-assignment at any time.

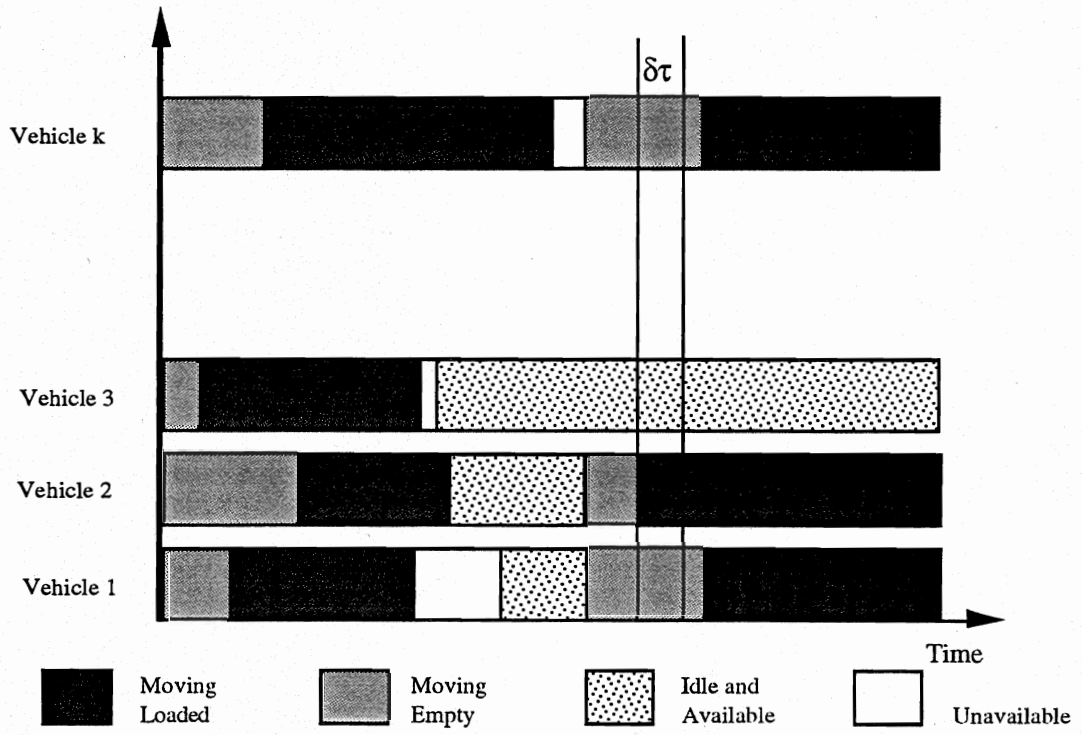
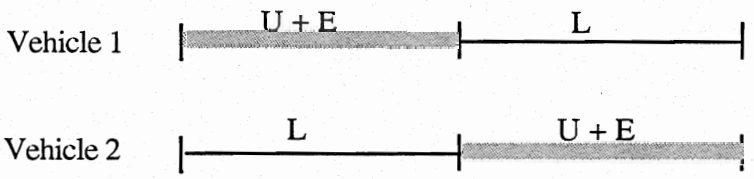


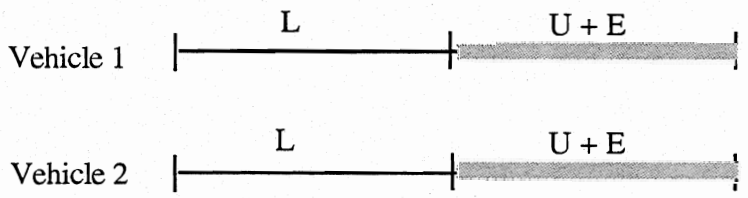
Figure 4.3 A record of vehicle states

To illustrate this point, we consider the idealized situation of a steady state in which the duration of each empty or loaded move and the idle time between loaded and empty moves are constant, independent of the load and of the vehicle. Let E denote the duration of each empty move, L the duration of each loaded move, and U the time spent idle (unassigned) between loaded and empty moves; in addition, let $e = (U+E)/L$ and $u = U/L$. In this illustration, the distinction between idle and available and idle and unavailable is not made, with no loss of generality.

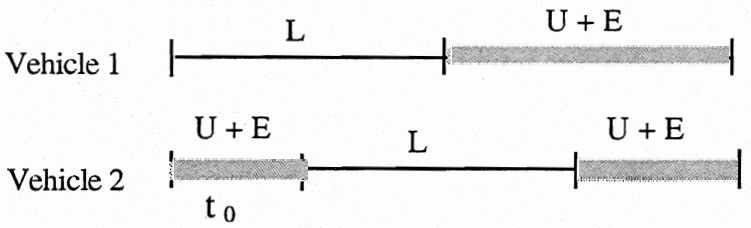
Assume, for illustration purposes, a constant $e = 1$ so that each vehicle is moving empty or is idle 50% of the time. With two vehicles, each with $e = 1$, the fraction of time when at least one vehicle is available to divert varies from 1.0 in the case where vehicle 1 is always loaded during the period when vehicle 2 is idle or empty (case a in figure 4.4) to 0.50 in the case where the loaded and empty moves of vehicles 1 and 2 overlap completely (case b in Figure 4.4). The more general case in which an offset, t_0, separates the start of the loaded move for vehicle 2 from the start of the loaded move for vehicle 1 is shown as case c in Figure 4.4.



a. No Overlap of U + E States



b. Complete Overlap of U + E States



c. General Case: Partial Overlap of U + E States

Figure 4.4 Overlap of vehicle states

Let t_0 = t_0/L represent the "relative offset"; for a given t_0, the conditional probability that at least one vehicle is available to divert is given by:

Proba(at least one vehicle available to divert | t_0) = $0.5 + t_0 0.5$ (4.4)

If we further assume that the times at which vehicles begin the loaded portion of their respective assignments are independent (i.e. no synchronization), and that t_0 is uniformly distributed between 0 and 1 (with pdf $ft_0(x) = 1.0$ for $0 \le x \le 1$), then for the two vehicle example shown in Figure 4.4 (case c), the conditional probability is given by:

$$\text{Proba(at least one vehicle able to divert)} = \int_0^1 [0.5 + (x)0.5]\, f\tau_0(x)dx$$

$$= \int_0^1 [0.5 + (x)0.5]\, dx = 0.75 \qquad (4.5)$$

Beyond the special cases of figure 4.4, with two vehicles and e = 1, vehicle availability for immediate dispatching can be quantified for a fleet of n vehicles and general e, but still under the assumptions of (1) steady state with constant duration of L and U+E, and (2) independence of the states associated with a particular vehicle (empty, loaded, or idle) of the states of other vehicles. The increase in vehicle availability under real-time information can be estimated by calculating the respective probabilities that at least one vehicle is a candidate for immediate assignment to a load under the diversion strategy possible with real-time information, and, without real-time information. Under the former, the probability that a given vehicle is available for immediate dispatch to a random call is given by the fraction of time that the vehicle is idle or moving empty, or e/(1+e). Without real-time information, a vehicle is available for immediate dispatch only when idle, so the corresponding probability is u/(1+e). For a fleet of n vehicles, the probability that at least one vehicle is available for immediate response is equal to [1- Proba(no vehicles available)]. Under the above assumptions and for the diversion strategy under real-time information this probability is given by

$$1-\left(1-\frac{e}{1+e}\right)^n \tag{4.6}$$

Without real-time information, the corresponding probability is given by:

$$1-\left(1-\frac{u}{1+e}\right)^n \tag{4.7}$$

The number of vehicles available at a given time is a binomially distributed random variable with expectation equal to n(e/(1+e)) in the case with real-time information and n(u/(1+e)) in the case without.

Naturally, values for u and e will vary significantly, depending upon the demand stream and assignment strategies employed. The probability that at least one vehicle is available to divert are given in Table 4.1 for two sets of values:

(1) u = 0.4 and e = 1.0, which corresponds to a relatively low demand environment with high idle time, and (2) u = 0.182 and e = 0.818, corresponding to relatively higher demand and more efficient operation.

TABLE 4.1 PROBABILITY OF VEHICLE AVAILABILITY WITH DIVERSION ALLOWED AND WITHOUT

u = 0.4 e = 1.0 Less Efficient Operation

n	no diversion $1-\left(1-\frac{u}{1+e}\right)^n$	diversion $1-\left(1-\frac{e}{1+e}\right)^n$	% increase
3	0.488	0.875	79.3
6	0.738	0.984	33.3
9	0.863	0.998	15.6
12	0.931	~1.00	7.4
15	0.965	~1.00	3.6

u = 0.182 e = 0.818 More Efficient Operation

n	no diversion $1-\left(1-\frac{u}{1+e}\right)^n$	diversion $1-\left(1-\frac{e}{1+e}\right)^n$	% increase
3	0.271	0.834	207.7
6	0.469	0.972	107.3
9	0.613	0.999	63.0
12	0.718	~1.00	39.3
15	0.794	~1.00	16.4
•	•	•	•
•	•	•	•
•	•	•	•
75	~1.00	~1.00	0.0

As shown in the table, even without en-route diversion, the probability that at least one vehicle is available to serve a load immediately increases rapidly with the size of the fleet. However this probability, and the expected number of available vehicles, is always higher under the diversion strategy with real-time information and increases more rapidly with fleet size. The increase in availability with the diversion strategy may be as high as 200% with very few vehicles (n = 3), but drops off rapidly to more realistic levels as n increases. This increase in the number of available vehicles is important because of the associated reduction in the expected distance from the origin location of the new load to the nearest available vehicle. The ability to divert increases the likelihood that the load can be served immediately and efficiently (from a travel distance point of view). This is illustrated in Figure 4.5, where in one case five of the vehicles are available to pick up a new demand and in the other only two are candidates for immediate assignment.

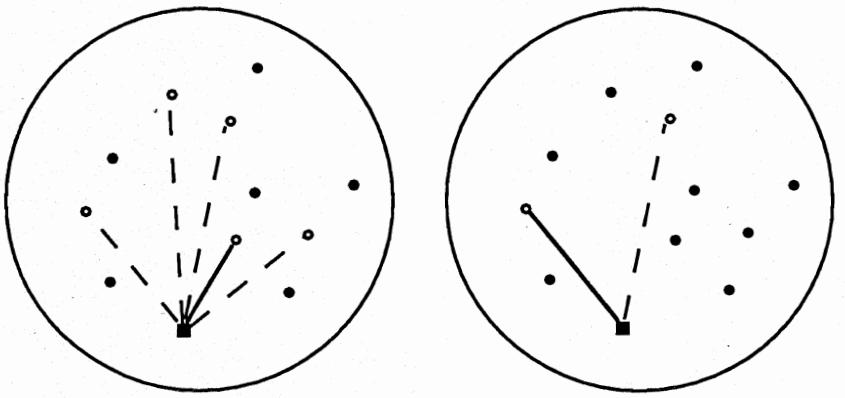


Figure 4.5 Distance from a new demand to available vehicles

In general, when m vehicles are uniformly distributed over unit a circle, the expected value of the distance from a randomly generated point and the nearest available vehicle can be approximated by (see for example, Larson and Odoni [1981])

$$0.2\sqrt{\frac{2\pi^2}{m}}. \qquad (4.8)$$

This value is shown as a function of m, the number of points in figure 4.6.

Of course, an available vehicle will not always be diverted to the new load, especially if it is en-route to pick up a load. A decision rule must be devised to determine whether to divert or not. Continuing our analysis of highly idealized situations, we first consider the myopic, greedy decision rule under which a vehicle will divert from its current (next) load to service a new load if the origin location of the new load is closer to the vehicle than the origin location of its current next load. Under this rule, a diversion will take place when the origin location of the new load falls within a circle with a radius equal to the distance remaining to the current load origin. In this case, the new load is said to be within the reach of the vehicle. In Figure 4.7, the vehicle en-route to load l_2 (with travel trajectory marked by thin arrow) would be a candidate from the point of view of proximity *and* time availability to divert to load l_3 while the vehicle moving empty toward the origin of load l_1 would be a candidate from the point of view of time availability only.

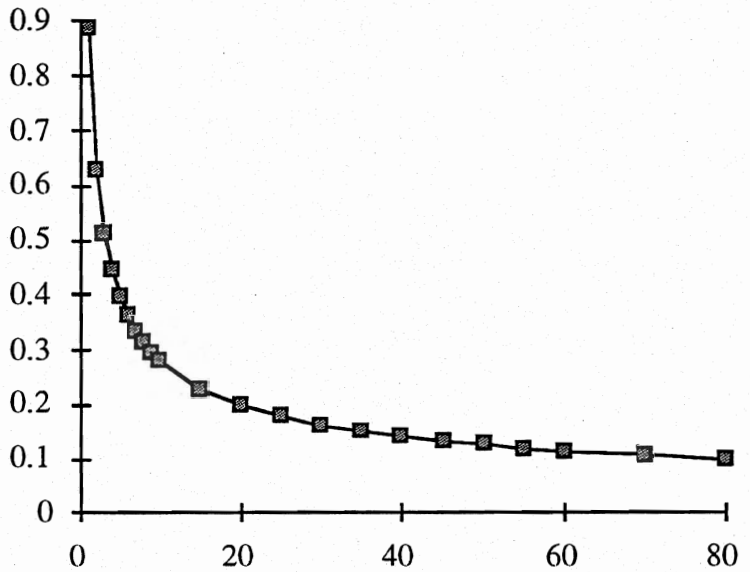


Figure 4.6 Expected distance between a randomly generated point and the closest of m randomly generated points in a circle.

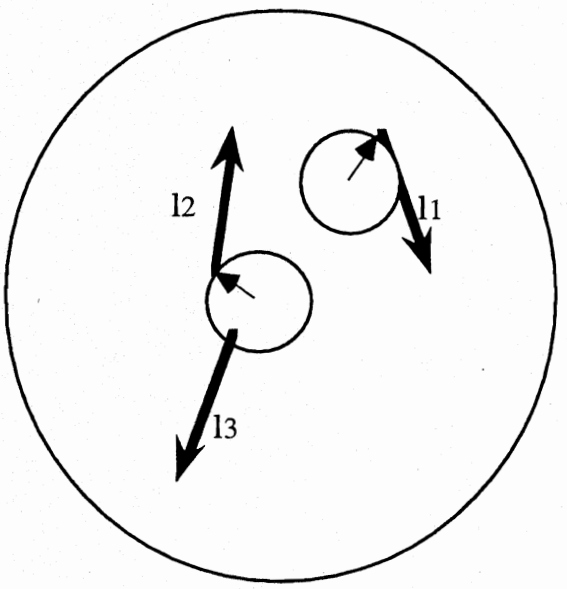


Figure 4.7 Circles of diversion

We extend the analysis to include the probability that at least one vehicle is available, with respect to both time and distance, to serve a newly requested load. Under our assumption that the vehicle states (and the state transition times) are independent, the expected distance from each vehicle to its load origin is uniformly distributed over the length of the empty distance traveled. Letting b represent the (steady state) length of an empty move, we assume that, at the time a new load is accepted, the distance remaining from a current vehicle location to the origin of its next load (for those vehicles moving empty) is uniformly distributed over $[0,\beta]$. Restricting ourselves to the interior of the work area (in order to avoid boundary effects and to keep independence between vehicles) it is not difficult to see that the probability that a new demand (in the interior of the work area) will be within the reach of a given vehicle is

$$\frac{\pi\int_0^{\beta} r^2 dr}{\pi R^2} = \frac{\frac{\beta^3}{3}}{R^2} = \frac{\beta^3}{3} \qquad (4.9)$$

(since R, the radius of the work area is 1). Denoting this probability P_p (for proximity), and assuming independence between vehicles, the probability that there is at least one vehicle within reach of this new load is then $1-\left(1-P_p\right)^n$ and the number of vehicles within reach of the new load may be estimated as a binomial random variable B(n, P_p). We showed earlier that the

probability that an individual vehicle is available, from the point of view of time, to respond to a new load is given by e/(1+e) in the case where we allow diversion. Since the probability that an individual vehicle is available in the case where diversion is not allowed is u/(1+e), the additional fraction of the fleet available is (e-u)/(1+e). Keeping in mind that all idle vehicles are considered close enough to respond to all new loads, but that en-route vehicles are within reach of a new load only if the new load is closer to the vehicle than the next load to which the vehicle was assigned, and, letting $P_a = e/(1+e)$ and $P_a' = u/(1+e)$ for (time) availability, then the expected number of vehicles available and close enough to respond to a new load, is given by

$$E[\text{vehicles available to respond}] = n[P_p(P_a - P'_a) + P'_a] \qquad (4.10)$$

While the expected additional vehicles available to respond when diversion is allowed is given by

$$E[\text{additional vehicles available to respond}] = nP_p(P_a - P'_a). \qquad (4.11)$$

Of course, the assignment of a demand to a particular vehicle, and its sequencing depend on the load's characteristics (e.g. time sensitivity) and its ability to fit in well with a set of loads. Analysis of the probability and associated benefits (measured in terms of reduction of overall distance traveled to serve a set of loads) for a single vehicle indicate that both the probability and associated benefits increase when a (short) sequence is used to estimate the benefit (or cost) of a diversion decision (Regan, Mahmassani & Jaillet [1995]). With a single vehicle, diversion from a previously assigned load requires a corresponding return to serve the load in the (near) future. However a fleet of vehicles offers many choices of how to arrange the loads. In addition, if a vehicle is assumed to have an associated (short) queue of assigned demands, the addition of a new load to a queue could change the order of a queue in many different ways. Figure 4.8 illustrates some of the possibilities in the simplest case. Vehicles v_1 and v_2 are en-route to loads l_1 and l_2 respectively when the request to move load l_3 arrives. If we assume that each vehicle will serve at least one of the loads there are 12 possible ways to (re)assign the three loads. In general, with two vehicles and n loads there are $(n+1)n!$ ways to arrange the loads if all possibilities are allowed and $(n-1)n!$ ways to arrange the loads if we assume that each vehicle must serve at least one load. Although logical methods may be constructed to evaluate the most likely choices, with even a small number of vehicles and loads the possibilities are huge.

Because of the complexity of the problem, a simulation model is used to facilitate the examination of the effect of allowing diversion and other flexible, real-time dispatching strategies

for fleets of vehicles. The simulation model is described in chapter 5, experiments performed are outlined and results discussed in chapter 6.

In the next section the examination of the ability of a fleet to provide timely service continues. Systems without explicit pickup deadline or time windows for service are of interest here too. Instead of focusing on the ability of a fleet to provide immediate service as in the previous section, the expected performance of a fleet providing truckload pickup and delivery services is modeled as a distributed queueing system with Poisson arrivals of requests for service, a general service distribution and k vehicles to provide service. This model allows the estimation of bounds on the expected wait time in the system under various assignment rules. For a single server system with some restrictions on the distribution of service times, extensive performance measures can be obtained.

CONGESTION EFFECTS: CARRIER FLEET OPERATIONS AS A DISTRIBUTED QUEUING SYSTEM

The purpose of this investigation is to examine the effects of congestion in a distributed service system. Queuing models offer a natural approach for analyzing service systems under congestion. This investigation considers a stream of customer requests which arrive at a dispatch center. Each request represents a single load (full truckload) to be moved from its point of origin to its destination. Such a system may be modeled as a queuing system in which requests for service are filled by identical spatially distributed (and mobile) servers. For the purposes of this analysis it is understood that service begins, not when the load is picked up, but when a vehicle begins moving towards the pickup location. Since each service time has two parts, empty and loaded, it is necessary to attribute an empty movement to each loaded move. The empty movement in this case corresponds to the setup time present in many manufacturing processes.

This analysis begins with an M/G/1 (single server) system in which service times are approximately independently and identically distributed (IID) variables. The relationship between such a system and systems in which there exist systematic dependencies between service times and the number and geographic locations of customers waiting for service is examined. It is shown that while analytic approximations for M/G/1 systems with independent travel times do not provide accurate estimates of the performance of the kinds of fleet assignment strategies of interest here, that they do provide bounds on the efficiency of such systems. This relationship is demonstrated directly for a single vehicle and generalized for the multi vehicle, M/G/k system. The discussion in this chapter is limited to assignment rules in which vehicles are assigned at most one load at a time and loads to not have associated pickup deadlines. Strategies in which

vehicles receive multiple load assignments and those addressing time windows and pickup deadlines are discussed in Chapters 5 and 6.

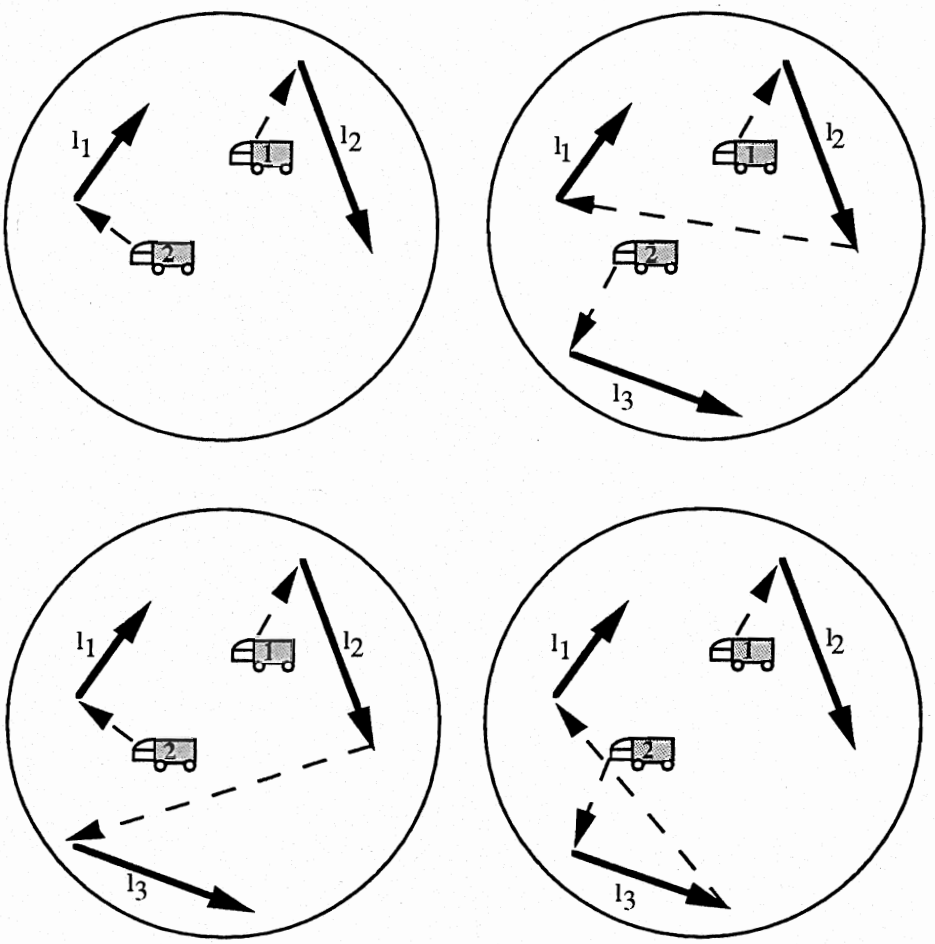


Figure 4.8 Examples of assignment/re-assignment possibilities

Preliminaries

Tijms [1986] employs an M/G/k queuing model of full truckload operations to determine the allocation of vehicles to sub-fleets with the goal of providing "uniform" service to customers. The analysis resorts to approximating the M/G/k queue with either an M/G/∞ or M/G/k queue where the service rate is Erlang distributed. More recently, Gans and van Ryzin [1996b] have investigated the causes and nature of congestion in dynamic dispatching systems by modeling a single vehicle, single origin-multiple destination system as a GI/GI/1 queue. In their investigation, loads arrive in batches at epochs and then are dispatched (again in batches) to routes. A single service completion is marked by the completion of a route, which involves the transport of one or

more loads to one or more destination locations. Dispatching heuristics in which loads are serviced in batches are examined in order to generate an analytically tractable upper bound on the expected work in the system.

The analysis of multiple server queuing systems tends to be very complicated, although simplifying assumptions allow for the approximation of key performance measures. Such analysis typically relies upon the assumption of independence of individual service times. This assumption is difficult to justify when reasonable dispatching strategies are applied in a spatially distributed service system because consecutive service times tend to be systematically dependent. Analytic approximations of system performance measures are derived with the independence assumption (see, for example Nozaki and Ross [1978]). Extending or applying these results to a system in which service times are not independent requires careful analysis and justification. In the next section the significance of the assumption of independent service times is addressed.

Significance of the Independence Assumption - Work in the M/G/1 and M/G/k System. This analysis begins with the M/G/1 case. A single vehicle, truckload trucking system is analyzed under the assumption that loads (service requests) are served in a First Called First Served (FCFS) manner. This assignment strategy possesses the property that consecutive service times are approximately independent. The extent to which service times are only approximately independent is addressed in the section to follow. The independence criterion is important for the following reason (Wolff [1989] p. 278): An important concept in the analysis of queues under general service distributions is the work in system. The work in system at any epoch, $t \geq 0$ is defined as the sum of the service times of all customers in queue and the remaining service times of all customers in service. Following the definition of Wolff [1989], letting V(t) represent the work in the system at time t, when service times are independent, V(t) has a simple representation. At each arrival epoch, V(t) experiences an upward jump. Between arrivals, V(t) decreases continuously as long as V(t) is positive. (See figure 4.9).

The derivation of analytic estimates of performance measures relies upon the predictable behavior of the system and the fact that an additional request may add to the work in the system but does not change the length of time needed to clear the previously accepted requests. The wait time in the system and in the queue as well as the average number of requests in the system and in the queue are calculated based upon this characterization of the work in the system. When service times are not IID variables, the arrival of a new service request can impact the amount of work already in the system.

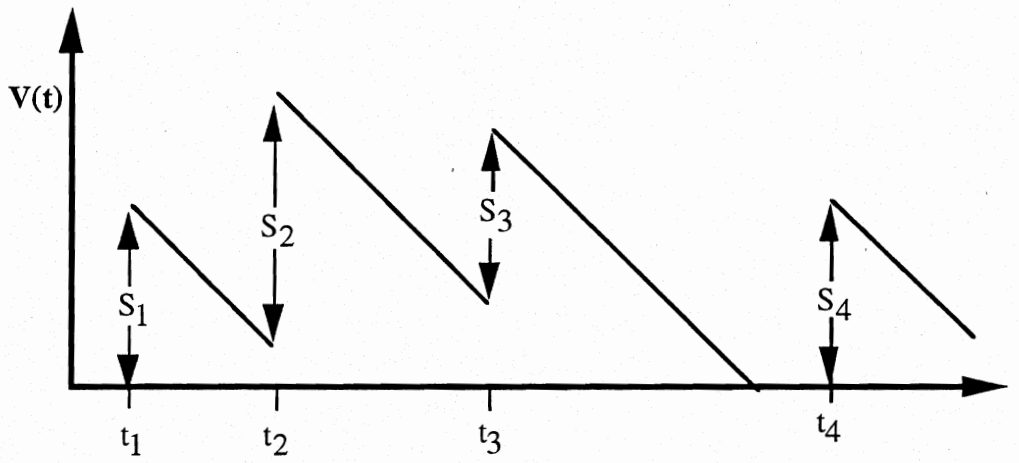


Figure 4.9 Work in M/G/1 queue under independence assumption

(Re-drawn from Wolff [1989] p. 279)

It can be shown that system efficiency under more "intelligent" assignment heuristics is bounded from above by the performance of the system under the FCFS strategy. Empirical estimates of the performance of an M/G/1 or M/G/k system under assignment rules that do not obey the independence assumption may be obtained through simulation. While these estimates allow for the identification of congestion levels that result in reasonable wait times and service costs, and help to characterize system performance, it may be useful at times to have analytic bounds available. Both analytical and empirical estimates of the efficiency of M/G/1 and M/G/k queueing systems are presented in the next several sections.

A Single Vehicle and the M/G/1 Queue

A precise definition of an M/G/1 queuing system is one in which customers arrive according to a Poisson arrival process, there is a single server, all blocked customers wait until served, the server cannot be idle when there is a waiting customer, and, service times are identically distributed, non-negative random variables, independent of the arrival process and each other.

Let us assume that loads are serviced by a single vehicle, operating in a circular work area and that demands arise according to a Poisson distribution over time and a uniform distribution over space (a Poisson point-process). Requests for service arriving when the vehicle is idle are served immediately. E[D] represents the distance between two independently and uniformly generated points inside a unit circle; 1/l is the average time between arrivals and E[S] the expected service time. Assuming that all requests for service are accepted and that service takes place according to a discipline that does not order loads with respect to geographic location, the

time to complete service may be expressed as the sum of two random variables, S_E, the time spent traveling empty to pick up a load and S_L, the time spent traveling loaded.

$$E[S] = E[S_E] + E[S_L]. \tag{4.12}$$

Under the given conditions and the assumption that each unit of distance traveled requires one unit of time, the expected value of each of these two random variables is the time needed to traverse the distance between two points randomly generated in a circle. $E[S_E] = E[S_L] = 0.905$ and may be derived in the following way(Elion, Watson-Gandy and Christofides [1971] p. 154):

Assume that points p_1 and p_2 have polar co-ordinates (r_1, ϕ_1) and (r_2, ϕ_2) respectively, and that in the general case the circle has radius a.

$$\text{Then } E[D] = \int_0^a\int_0^a\int_0^{2\pi}\int_0^{2\pi} (D f_{r_1} f_{r_2} f_{\phi_1} f_{\phi_2}) dr_1 dr_1 d\phi_1 d\phi_2 \tag{4.13}$$

$$= \frac{1}{\pi^2 a^4} \int_0^a r_1 dr_1 \int_0^a r_2 dr_2 \int_0^{2\pi}\int_0^{2\pi} \sqrt{r_1^2 + r_2^2 - 2r_1 r_2 \cos(\phi_1 - \phi_2)}\, d\phi_1 d\phi_2 \tag{4.14}$$

$$= \frac{8a}{5\{\Gamma(5/2)\}^2} = 0.905a \tag{4.15}$$

Continuing along these lines, the variance of the distance σ_D^2 may also be obtained. $\sigma_D^2 = E[D^2] - E[D]^2$ and

$$E[D^2] = \int_0^a\int_0^a\int_0^{2\pi}\int_0^{2\pi} (D^2 f_{r_1} f_{r_2} f_{\phi_1} f_{\phi_2}) dr_1 dr_1 d\phi_1 d\phi_2 \tag{4.16}$$

$$= \frac{1}{\pi^2 a^4} \int_0^a r_1 dr_1 \int_0^a r_2 dr_1 \int_0^{2\pi}\int_0^{2\pi} r_1^2 + r_2^2 - 2r_1 r_2 \cos(\phi_1 - \phi_2) d\phi_1 d\phi_2 \tag{4.17}$$

$$= \left(\frac{4\pi^2}{\pi^2 a^4}\right)\left(\frac{a^6}{4}\right) = 1.0a^2 \tag{4.18}$$

For the special case of the unit circle E[D] = 0.905 and $E[D^2] = 1.0$ so $\sigma_D^2 = \sigma_E^2 = \sigma_L^2 = 0.181$ (easily verified by simulation). Under the assumption that traversing one unit of distance requires one unit of time, $E[S_E] = E[S_L] = E[D]$. While

$E[S] = E[S_E] + E[S_L]$ = 1.81. A mentioned previously, service times that share a point are only approximately independent. This applies in the case of consecutive service to separate customers and also to service to a single customer, which is the sum of two moves which share a point, the origin location of the load served. Simulation of the system under FCFS assignment yields the following value: $\sigma_S^2 = 0.397 \neq \sigma_E^2 + \sigma_L^2$. The covariance is quite small, (0.0175) and the correlation coefficient Corr(S_E,S_L) = 0.048.

We present as further evidence that FCFS service very nearly satisfies the independence assumption the simulation results and presented in figure 4.10, in which an M/G/k system is simulated and values of key performance measures compared to two known approximations developed under the assumption of independent service times. Simulation values very closely represent values obtained under the known approximations. While FCFS service does not precisely preserve the independence assumption, service times are not systematically dependent on the number and location of waiting customers.

Letting E[S] and E[S^2] represent the first and second moments of the service time and r = lE[S], the traffic intensity or utilization rate (sometimes, called congestion level) of the system, some important service performance measures can be approximated (see, for example Larson and Odoni [1981], Ross [1981]).

P_0, the probability that a randomly arriving customer (request for service) finds the system empty is given by:

$P_0 = 1-r.$ (4.19)

$\overline{W}_q$, the average length of time a user spends in the queue

$$\overline{W}_q = \frac{\lambda E[S^2]}{2(1-\lambda E[S])}. \quad (4.20)$$

$\overline{L}_q$ the average number of requests for service already in the queue (not including the request being serviced),

$$\overline{L}_q = \lambda \overline{W}_q = \frac{\lambda^2 E[S^2]}{2(1-\lambda E[S])}. \quad (4.21)$$

$\overline{W}$, the average length of time a user spends in the system,

$$\overline{W} = \overline{W}_q + E[S] = \frac{\lambda E[S^2]}{2(1-\lambda E[S])} + E[S]. \qquad (4.22)$$

L, the number of requests for service already in the system when a randomly arriving request for service arrives,

$$L = \lambda W = \frac{\lambda^2 E[S^2]}{2(1-\lambda E[S])} + \lambda E[S]. \qquad (4.23)$$

Representative values of these performance measures are presented in table 4.2, for varying rates of arrivals of requests for service. These are obtained by applying equations (4.19)-(4.23) but may also be found by simulating an M/G/1 system with the specified characteristics at steady state. Estimates of system performance measures can be used to define the most "attractive" congestion levels for the system, and to turn away loads that can not be served within a reasonable period of time. An "attractive" congestion level results in relatively short empty distances traveled (in this case short service times) and few idle periods, but, at the same time does not result in unreasonable wait times for service. Adding a measure of physical realism to the system, the circle is assumed to have a radius of 250 miles and the wait time is given in hours as well as general "units". The average loaded move is then 226 miles long and the wait time in hours can be estimated under the assumption that travel takes place at a constant speed of 50 miles per hour. We assume in this example that handling times are zero and that the server is never unavailable.

TABLE 4.2 SYSTEM PERFORMANCE MEASURES FOR M/G/1 WITH FCFS ASSIGNMENT

ρ	E[S]	σ_s^2	$\overline{L}$	$\overline{L}_q$	$\overline{W}$	$\overline{W}_q$	$\overline{W}$(hours)	$\overline{W}_q$(hours)
0.73	1.81	0.397	1.88	1.14	4.62	2.81	23.11	14.05
0.76	1.81	0.397	2.12	1.36	5.04	3.23	25.21	16.15
0.79	1.81	0.397	2.45	1.66	5.62	3.81	28.12	19.06
0.82	1.81	0.397	2.90	2.08	6.42	4.61	32.09	23.03
0.85	1.81	0.397	3.51	2.66	7.48	5.67	37.42	28.36
0.88	1.81	0.397	4.46	3.58	9.19	7.38	45.96	36.91
0.91	1.81	0.397	5.98	5.07	11.91	10.10	59.55	50.49
0.94	1.81	0.397	8.00	8.94	17.26	15.40	86.28	77.22

M/G/1 Results and More Efficient Assignment Heuristics

While in a few applications it may be practical to serve customers in a first come first served (FCFS) manner, or even a last come first served (LCFS) manner, in a spatially distributed service system it is more efficient to serve customers (to move loads in this case) in an order which reduces the overall distance traveled. In this section a simple assignment strategy, nearest origin assignment is presented. The efficiency of a single vehicle service system under this rule is compared with that of a similar FCFS service system. It is shown in the next section that the M/G/1 system with FCFS service provides analytically tractable upper bounds on more efficient systems in which performance measures may be simulated but not obtained analytically. The assignment strategy introduced for comparison assumes that the available vehicle is assigned to the load in queue with an origin location nearest to the current location of the vehicle. Before discussing the nearest origin assignment we describe the simulation used to estimate system performance.

The Simulation Framework. The simulation assumes a circular geographic region in which origin and destination locations are uniformly and independently distributed over the region. Demands arrive according to a Poisson process. The simulation horizon is 2600 simulation weeks. The choice of a long simulation horizon is made so that steady state values of key performance measures can be estimated. Values of key performance measures are aggregated over the second half of the simulation horizon, after the system has been in service for 1300 simulation weeks. While not absolutely ensuring steady state, results under the FCFS policy closely match those that can be estimated using known, closed-form approximations of key performance measures (table 4.2). The average number of customers in the system is measured upon the arrival of each new load (customer) and these values aggregated over all arrivals following the half-way point in the simulation. Applying the assumption that Poisson arrivals see time averages (PASTA in Wolff [1989] p. 293)) these estimates are assumed to represent the overall system averages.

Comments on Nearest Origin Assignment . The following nearest origin assignment strategy is analyzed: loads are generated by a Poisson point process on a unit circle; a single vehicle provides service to the loads; a queue of infinite length is allowed to form; all loads in the queue are candidates for assignment when the vehicle becomes available; loads do not have explicit time windows for service or pickup deadlines; if no loads are awaiting service when a vehicle becomes available it remains at the destination location of its last load served until it

receives another assignment. Assuming, as before, that each unit of distance traveled requires one unit of time, the expected service time may be calculated in the following way.

Let $E[S_{fcfs}]$ represent the expected time to complete service under the FCFS policy and $E[S_{no}]$ the expected time to complete service under the nearest origin strategy. Then S_{Efcfs} represents the empty portion of service time and S_L the loaded portion of service time. Let $\mathbf{L_s}$ represent the number of customers in the system and $\mathbf{L_q}$ the number of customers (loads) in the queue. In addition, let $P_\ell = \lim_{t\to\infty} \Pr\{\text{upon arrival of a customer there are l customers in the system}\}$. In a system with Poisson arrivals this can be shown to be equal to the long run proportion of time in which there are l customers in the system and l-1 customers in the queue. Then, the expected service time (with no limit on the number of loads allowed in the queue and all queued loads considered candidates for all assignments) is:

$$E[S] = E[S_E] + E[S_L] \quad \text{(from (4.12))}$$

$$E[S_{no}] = \sum_{\ell=0}^{\infty} P_\ell E[S_{E_{no}} | \mathbf{L_s} = \ell] + E[S_L] \tag{4.24}$$

$E[S_L]$ is known (0.905), and equal in all cases examined, and $E[S_{E_{no}} | \mathbf{L_s} = \ell]$ is a monotonically decreasing function in l, representing the expected distance between a randomly generated point in a circle and l-1 other randomly generated points in a circle. $E[S_{E_{no}} | \mathbf{L_s} = \ell]$ is well approximated by $0.2\sqrt{\frac{2\pi^2}{\ell - 1}}$ (equation (4.2), figure 4.6) when l is greater than one. It will be argued below that this is equal to $E[S_{fcfs}]$ when l is equal to one.

Proposition 4.1: The expected service time $E[S_{no}]$ under the nearest origin assignment rule is bounded from above by $E[S_{fcfs}]$, the expected service time under the FCFS strategy.

Proof:

From (4.12) and (4.23) we have

$E[S] = E[S_E] + E[S_L]$ and

$$E[S_{no}] = \sum_{\ell=0}^{\infty} P_\ell E[S_{E_{no}} | \mathbf{L_s} = \ell] + E[S_L]$$

We wish to show that $E[S_{no}] \le E[S_{fcfs}]$ or, that

$$\sum_{\ell=0}^{\infty} P_\ell E[S_{E_{no}} | \mathbf{L_s} = \ell] + E[S_L] \le E[S_{Efcfs}] + E[S_L] \text{ or, that}$$

$\sum_{\ell=0}^{\infty} P_\ell E[S_{E_{no}} | \mathbf{L_s} = \ell] \leq E[S_{E_{fcfs}}]$ since the expected time spent loaded does not change under different assignment policies.

To show this it is observed once more that for each arriving customer, the expected service time for the next load served is the sum of the expected time spent loaded and the expected time spent moving empty. If the system is empty when the customer arrives, then the expected service time of the next customer (the arriving customer in this case) will be the same as under the FCFS policy. If the vehicle is in service, the expected empty service time for the next load served will be the expected time to traverse the distance between a randomly generated point (the destination of the load currently in service) and the one or more loads in queue when service has been completed. This expected time is less than or equal to the expected time under the FCFS policy. Therefore, the proof is completed and,

$$E[S_{no}] \leq E[S_{fcfs}] \tag{4.25}$$

In order to ensure that the performance measures for the M/G/1 queue with FCFS service can be used to generate bounds on the performance of a similarly configured M/G/1 queue with nearest origin assignment we would need to show that the 2nd moment of the service time is also less than or equal to that under the FCFS assignment. Estimates for the wait time for service, and the expected number in queue (and in the system), rely only on the first and second moments of the service time distribution when requests for service are generated by a Poisson process. A simple simulation of the nearest assignment strategy and the system described yields the results shown in table 4.3. Comparing the standard deviation of service in the simulated system with that of the FCFS assignment (table 4.2) it may be observed that in all cases shown the standard deviation is lower. The coefficient of variation remains essentially constant across experiments with different congestion levels.

Simulation results indicate that under the nearest origin heuristic both E[S] and σ_s^2 decrease as congestion increases. This is because the number of customers waiting for service increases--leading to more selection opportunities at each assignment epoch. The reduction in E[S] as the number of customers in the queue (congestion) increases in displayed in Figure 4.10.

Table 4.3 Simulation estimates of system performance measures for an M/G/1 system under nearest origin assignment

ρ	E[S]	σ_s^2	$\overline{L}$	$\overline{L}_q$	$\overline{W}$	$\overline{W}_q$	$\overline{W}$(hours)	$\overline{W}_q$(hours)
0.73	1.69	0.381	1.57	0.83	3.61	1.92	18.05	9.58
0.76	1.68	0.379	1.72	0.96	3.79	2.11	18.95	10.56
0.79	1.66	0.375	1.91	1.12	4.01	2.35	20.05	11.75
0.82	1.64	0.370	2.13	1.31	4.27	2.63	21.35	13.15
0.85	1.61	0.364	2.43	1.58	4.62	3.01	23.10	15.05
0.88	1.58	0.356	2.81	1.94	5.07	3.49	25.35	17.45
0.91	1.54	0.345	3.35	2.44	5.70	4.15	28.48	20.75
0.94	1.50	0.331	4.12	3.18	6.59	5.09	32.95	25.45

M/G/1 approximations for the wait time in the system and in queue and the expected number of customers in the system and in queue shown in equations (4.20)-(4.23) do not provide accurate estimates of the wait time and queue length because they are derived under the assumption of independence of consecutive service times. For comparison, equations (4.20)-(4.23) are applied to the first and second movements of the service times measured using simulation, and are displayed in Table 4.4. These results clearly do not match the simulation results in Table 4.3 (obtained under the assumption that Poisson arrivals see time averages (PASTA in Wolff [1989] p. 293)). It is assumed as before that one unit of distance traveled requires one unit of time and that these values are converted into hours by assuming that the radius of the circle is 250 miles long and that travel speeds are a constant 50 miles per hour.

Comparison of tables 4.2 and 4.3 illustrate the significant increase in efficiency when a nearest origin assignment strategy is applied over FCFS assignment. Wait time in the FCFS system is one and a half to three times the average wait for service under nearest origin assignment. A reasonable goal might be that the mean time to complete service be less than 24 hours; in that case the system in which the nearest origin assignment is used can run at a utilization rate of 85% while the FCFS system can only safely operate at a rate of 73%.

Table 4.5 offers a comparison of wait times for service displayed in tables 4.2-4.4. Comparing tables 4.2, 4.3 and 4.4 it may be observed that

$$\{\overline{W}, \overline{W}_q\}_{\text{no(from simulation)}} <$$

$$\{\overline{W}, \overline{W}_q\}_{\text{approximation (with E[S] and E[S}^2\text{] from simulation of NO)}} <$$

$$\{\overline{W}, \overline{W}_q\}_{\text{approximation of FCFS}}$$

The comparison indicates that M/G/1 approximations applied using the first and second moments measured through simulation provide tighter bounds on the performance of the system than do the approximations based upon FCFS service and approximate independence of service times.

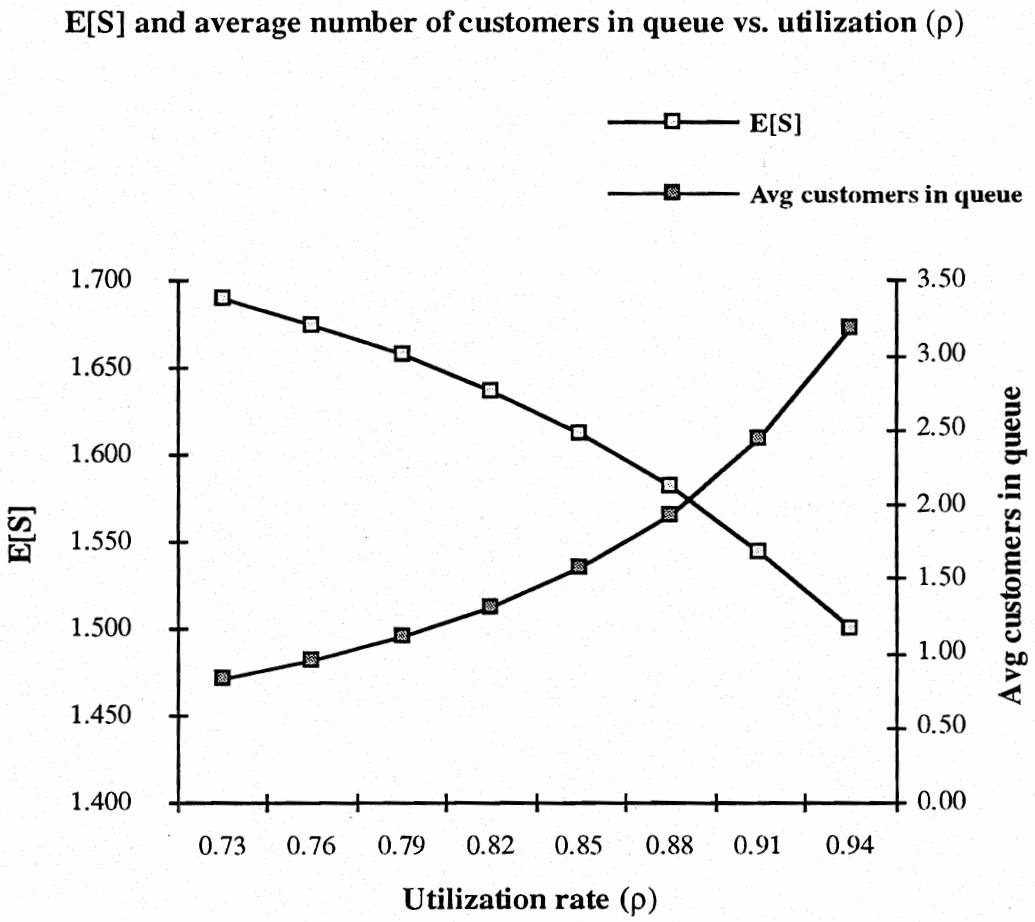


Figure 4.10 E[S] and the average number of customers in queue ($\overline{L}_q$)

TABLE 4.4 APPLICATION OF M/G/1 PERFORMANCE MEASURE APPROXIMATIONS UNDER THE ASSUMPTION OF IID SERVICE TIMES. E[S] AND E[S^2] OBTAINED THROUGH SIMULATION

ρ	E[S]	σ_s^2	$\overline{L}$	$\overline{L}_q$	$\overline{W}$	$\overline{W}_q$	$\overline{W}$(hours)	$\overline{W}_q$ (hours)
0.73	1.69	0.381	1.89	1.15	4.34	2.65	21.72	13.27
0.76	1.68	0.379	2.14	1.38	4.71	3.03	23.54	15.16
0.79	1.66	0.375	2.47	1.68	5.19	3.53	25.96	17.67
0.82	1.64	0.370	2.92	2.10	5.84	4.20	29.19	21.00
0.85	1.61	0.364	3.56	2.71	6.76	5.14	33.78	25.72
0.88	1.58	0.356	4.52	3.64	8.14	6.56	40.69	32.78
0.91	1.54	0.345	6.09	5.18	10.36	8.81	51.79	44.06
0.94	1.50	0.331	9.03	8.07	14.41	12.91	72.05	64.53

TABLE 4.5 COMPARISON OF SIMULATED AND ESTIMATED WAIT TIME FOR SERVICE UNDER NO AND FCFS ASSIGNMENT

$\rho = 0.73$

	E[S]	σ_s^2	$\overline{W}$	$\overline{W}_q$	$\overline{W}$(hours)	$\overline{W}_q$(hours)
I.	1.69	0.381	3.61	1.92	18.05	9.58
II.	1.69	0.381	4.34	2.65	21.72	13.27
III.	1.81	0.397	4.62	2.81	23.11	14.05

$\rho = 0.79$

	E[S]	σ_s^2	$\overline{W}$	$\overline{W}_q$	$\overline{W}$(hours)	$\overline{W}_q$(hours)
I.	1.66	0.375	4.01	2.35	20.05	11.75
II.	1.66	0.375	5.19	3.53	25.96	17.67
III.	1.81	0.397	5.62	3.81	28.12	19.06

$\rho = 0.85$

	E[S]	σ_s^2	$\overline{W}$	$\overline{W}_q$	$\overline{W}$(hours)	$\overline{W}_q$(hours)
I.	1.61	0.364	4.62	3.01	23.10	15.05
II.	1.61	0.364	6.76	5.14	33.78	25.72
III.	1.81	0.397	7.48	5.67	37.42	28.36

$\rho = 0.91$

	E[S]	σ_s^2	$\overline{W}$	$\overline{W}_q$	$\overline{W}$(hours)	$\overline{W}_q$(hours)
I.	1.54	0.345	5.70	4.15	28.48	20.75
II.	1.54	0.345	10.36	8.81	51 79	44 06
III.	1.81	0.397	11.91	10.10	59 55	50 49

I. Nearest Origin (simulated) (table 4.3)

II. Approximation with E[S] and $E[S^2]$ from simulation of NO (table 4.4)

III. Approximation of FCFS (table 4.2)

In the next section the analysis of a single vehicle is extended to a fleet of vehicles. Instead of an M/G/1 queue we examine an M/G/k queue, where k is the number of vehicles in the fleet.

A Vehicle Fleet and the M/G/k Queue

The M/G/k queue presents even more challenge than its singer server counterpart -- approximations are only available for special cases. Again, the examination of such a system under simplifying assumptions may be useful.

We begin with the following assumptions: again, customers arrive according to a Poisson point process; when fleets of varying sizes are examined, we assume that the arrival rate of requests for service is proportional to the number of vehicles in the fleet; service times are independent random variables -- the service rate is directly proportional to the number of servers (vehicles) in the fleet; all customers may be accepted into the queue.

While the convenient expressions corresponding to (4.20)-(4.23) are not known for the M/G/k system, several approximations for $\overline{W}_q$, the wait time in queue (from which other performance measures may be estimated), have been developed. Nozaki and Ross [1978] present several such approximations. These are examined in the next section.

Approximations for the Average Wait in Queue in an M/G/k System. The following approximation is derived in Nozaki and Ross[1978] and is consistent with the assumptions listed. It is suggested (in the paper) that this approximation is also valid when there is a limit on the size of the queue allowed to form. It is applied here with the assumption that all requests for service can be accommodated in the queue and the analysis is restricted to arrival rates that do not (in simulation experiments) result in refusal of requests for service.

$$\overline{W}_q \approx \frac{\lambda^k E[S^2](E[S])^{k-1}}{2(k-1)!(k-\lambda E[S])^2\left[\sum_{j=0}^{k-1}\frac{(\lambda E[S])^j}{j!}+\frac{(\lambda E[S])^k}{(k-1)!(k-\lambda E[S])}\right]} \qquad (4.26)$$

In addition, a heavy traffic approximation offered in Kingman [1965] is provided by Nozaki and Ross [1978]. For the case of Poisson arrivals and for $\lambda E[S] \approx k$, this approximation is:

$$\overline{W}_q \approx \frac{\lambda^k E[S^2]-\lambda^2(E[S])2+k^2}{2\lambda k(k-\lambda E[S])} \qquad (4.27)$$

For smaller vehicle fleets r, the utilization rate is lower to ensure that all requests for service may be accommodated The approximations are evaluated for various values of k (the number of servers-vehicles in the fleet) and the resulting approximations displayed in tables 4.6 and 4.7. Values of E[S} and $E[S^2]$ used in the Nozaki and Ross and Kingman approximation equations are

those obtained through simulation of the system under FCFS assignment. The approximation results are compared to those measured from the simulation of the system under FCFS assignment. Requests (loads) are assigned in the order in which they arrive. When a server becomes idle, it is assigned the first load in the queue. The Nozaki and Ross approximation provides a very accurate approximation for the FCFS system while the Kingman approximation is also quite close.

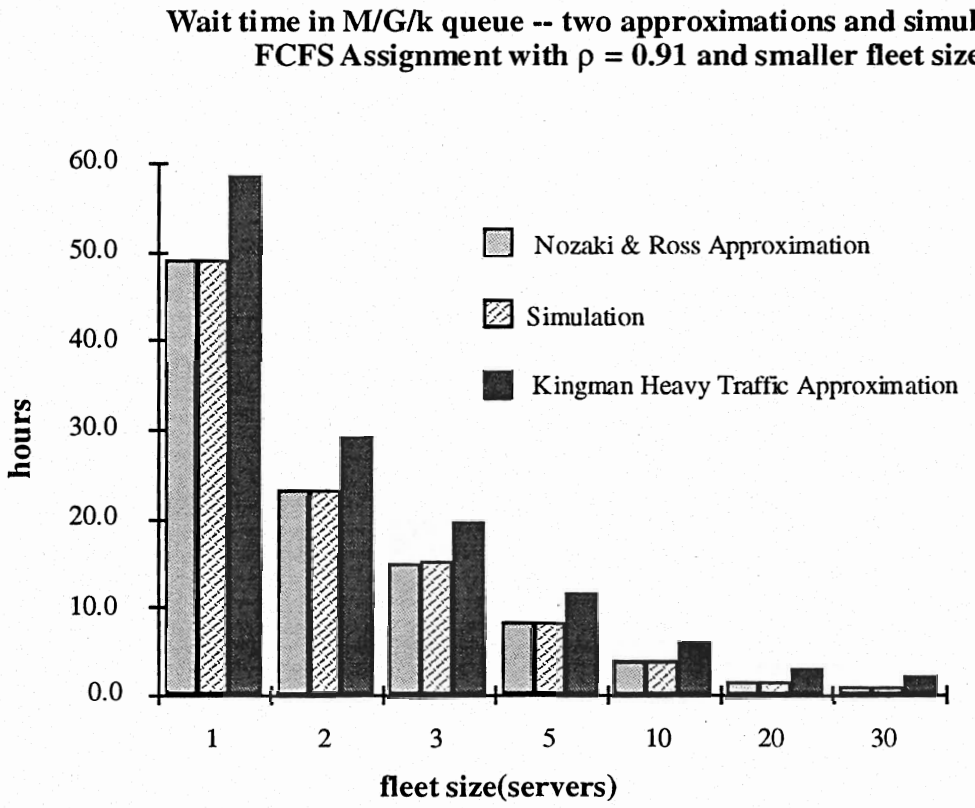


Figure 4.11 Approximation of average wait time in queue with approximations of Nozaki & Ross and Kingman and Simulation of FCFS assignment (smaller fleet sizes)

TABLE 4.6 COMPARISON OF N&R AND KINGMAN WAIT TIME APPROXIMATIONS WITH SIMULATION OF FCFS ASSIGNMENT

k	ρ	E[S]	σ_s^2	λE[S]	$\overline{W}_{N\&R}$	$\overline{W}_{sim}$	$\overline{W}_{King}$	$\overline{W}_{N\&R}$ (hrs)	$\overline{W}_{sim}$ (hrs)	$\overline{W}_{King}$ (hrs)
1	0.91	1.81	0.397	0.91	9.79	9.79	11.70	48.9	48.9	58.5
2	0.91	1.81	0.397	1.81	4.65	4.66	5.84	23.3	23.3	29.2
3	0.91	1.81	0.397	2.72	2.98	3.01	3.89	14.9	15.05	19.5
4	0.91	1.81	0.397	4.53	1.65	1.68	2.31	8.3	8.4	11.6
10	0.91	1.81	0.397	9.07	0.75	0.75	1.18	3.7	3.9	5.9
20	0.91	1.81	0.397	18.12	0.31	0.31	0.59	1.6	1.6	2.9
30	0.91	1.81	0.397	27.17	0.18	0.19	0.39	0.9	1.0	1.9

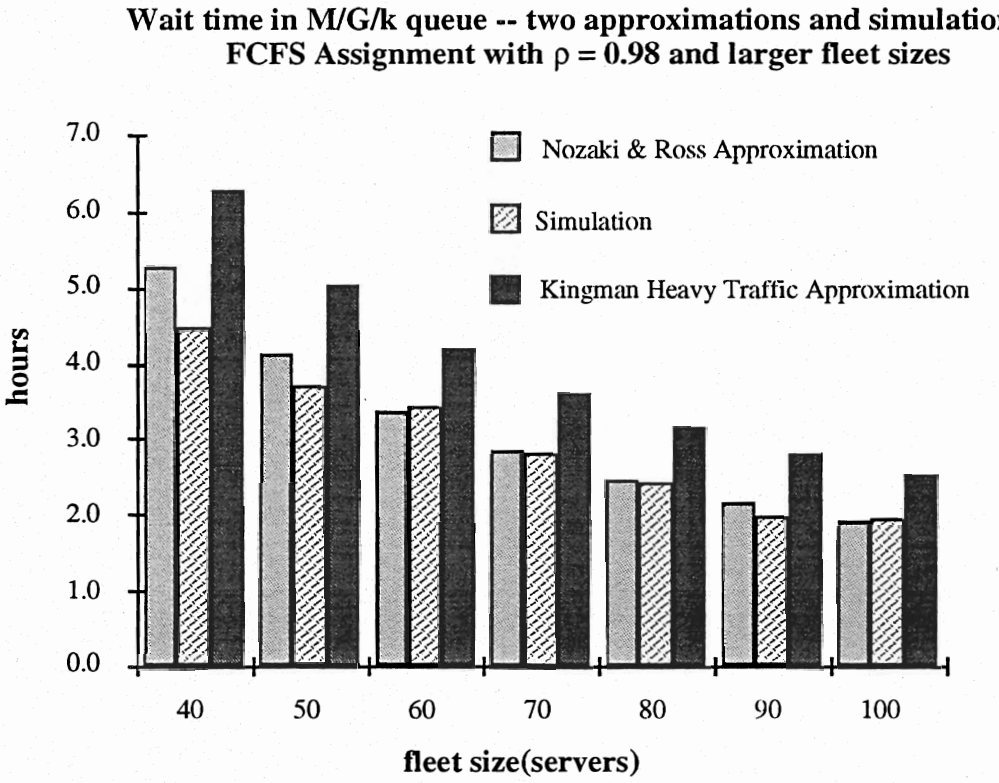


Figure 4.12 Approximation of average wait time in queue with approximations of Nozaki & Ross and Kingman and Simulation of FCFS assignment (larger fleet sizes)

TABLE 4.7 COMPARISON OF N&R AND KINGMAN WAIT TIME APPROXIMATIONS WITH SIMULATION OF FCFS ASSIGNMENT FOR LARGER FLEETS AND HIGHER UTILIZATION

k	ρ	E[S]	σ_s^2	λE[S]	$\overline{W}_{N\&R}$	$\overline{W}_{sim}$	$\overline{W}_{King}$	$\overline{W}_{N\&R}$ (hrs)	$\overline{W}_{sim}$ (hrs)	$\overline{W}_{King}$ (hrs)
40	0.98	1.81	0.397	39.18	1.05	0.89	1.26	5.25	4.45	6.30
50	0.98	1.81	0.397	48.97	0.83	0.74	1.01	4.15	3.70	5.05
60	0.98	1.81	0.397	58.77	0.68	0.69	0.84	3.40	3.45	4.20
70	0.98	1.81	0.397	68.56	0.57	0.56	0.72	2.85	2.80	3.60
80	0.98	1.81	0.397	78.36	0.49	0.49	0.63	2.45	2.45	3.15
90	0.98	1.81	0.397	88.15	0.43	0.40	0.56	2.15	2.00	2.95
100	0.98	1.81	0.397	97.95	0.38	0.39	0.50	1.90	1.90	2.50

M/G/k Results and More Efficient Assignment Heuristics. The comparison of nearest origin assignment to the FCFS strategy for an M/G/1 queue presented mentions that analytic approximation of a queuing system with an assignment rule in which service times are systematically dependent on the state of the system are not apparently possible. This too is the case in the M/G/k system. Approximations are based upon variations of the following approximation assumption (Nozaki and Ross[1978])

" Let G_e denote the equilibrium distribution of G...

Approximation assumption (AA). Given that a customer arrives to find i busy servers, i > 0, then at the time he enters service, the remaining service times of the other i-d(i,k) customers being served has a joint distribution that is approximately that of independent random variables each having distribution G_e."

However, in the nearest origin assignment applied to a fleet of vehicles, we assume that upon completion of service a vehicle is assigned the closest load in the queue. This means that G and G_e, the service time distribution, is highly dependent upon both the number of servers and the number of loads in queue. This implies that the service rate is systematically related to the number of busy servers (which is in turn related to the number of customers in queue). So, the independence assumption stated above and other related assumptions are not valid. Unfortunately these imply that approximations for the M/G/k system cannot be applied in this case. We find however, that just as approximations for key performance measures in the system provided a bound on the performance of the system under the nearest origin strategy, so do the approximations for the average wait time in queue provide bounds on the performance of the M/G/k queue with the nearest origin strategy. The applicability of these bounds depends upon the congestion level of the system. While Kingman's approximation is a heavy traffic

approximation, both the Nozaki & Ross and Kingman approximations provide a tighter bound on the wait time in system when applied to a relatively less congested system under the nearest origin strategy. This may be observed in figures 4.12 and 4.13 for two utilization rates: 0.99 and 0.97. While these are not typical congestion levels, they are selected in order to compare the simulated system under nearest origin assignment to the Nozaki and Ross and Kingman approximations which are explicitly derived for systems under heavy traffic. In a less congested system the performance of the nearest origin strategy approaches that of a system with FCFS assignment.

Applying the Bounds on Wait Time to "Actual " Systems. Despite the lack of IID service times, the approximations of wait time in system may be useful for generating bounds on the wait time in the system. The mean and variance of service times may be measured from actual data, and target utilization regions identified. Load acceptance policies may be adapted to actual conditions in the system and load acceptance or assignments to sub-fleets of vehicles made accordingly. In the absence of explicit pickup deadlines, carrier fleet operators must still provide service to customers fairly quickly. While the development of bounds on wait time, and methods to estimate attractive congestion levels for distributed service systems introduced here is incomplete, it does offer a starting point for this type of analysis.

Behavior of Nearest Origin Assignment as r $\rightarrow$ 1. It is of interest to know how the nearest origin assignment strategy performs in the limit. The expected time to complete service can be shown to be between two bounds:

$$E[S_L] < E[S_{no}] \leq 2E[S_L] \qquad (4.28)$$

where $E[S_L]$ is the expected time to complete the loaded portion of the movement. This relationship is evident from the fact that if the nearest load to a vehicle was always at the exact location of the vehicle then the distance traveled empty would be zero, and by the fact that in the FCFS or other assignment policy not tied to geographic location the expected time to complete the empty movement is equal to the expected time to complete the loaded movement. The discussion to follow shows that in the limit, as the number of loads in queue increases without bound, the expected length of the empty movement goes to zero. Of more immediate interest however, is the fact that the service rate increases with increasing congestion. That is, as congestion increases the service rate also increases (the time needed to complete service decreases). As the queue grows, the service rate increases and at steady state there exists a rather long queue (for a large fleet of vehicles) but the wait time in this queue may be fairly short.

Limited simulation results are presented to demonstrate these findings because of the complexities of solving for $E[S_{no}]$. To approximate $E[S_{no}]$ equation (4.24)

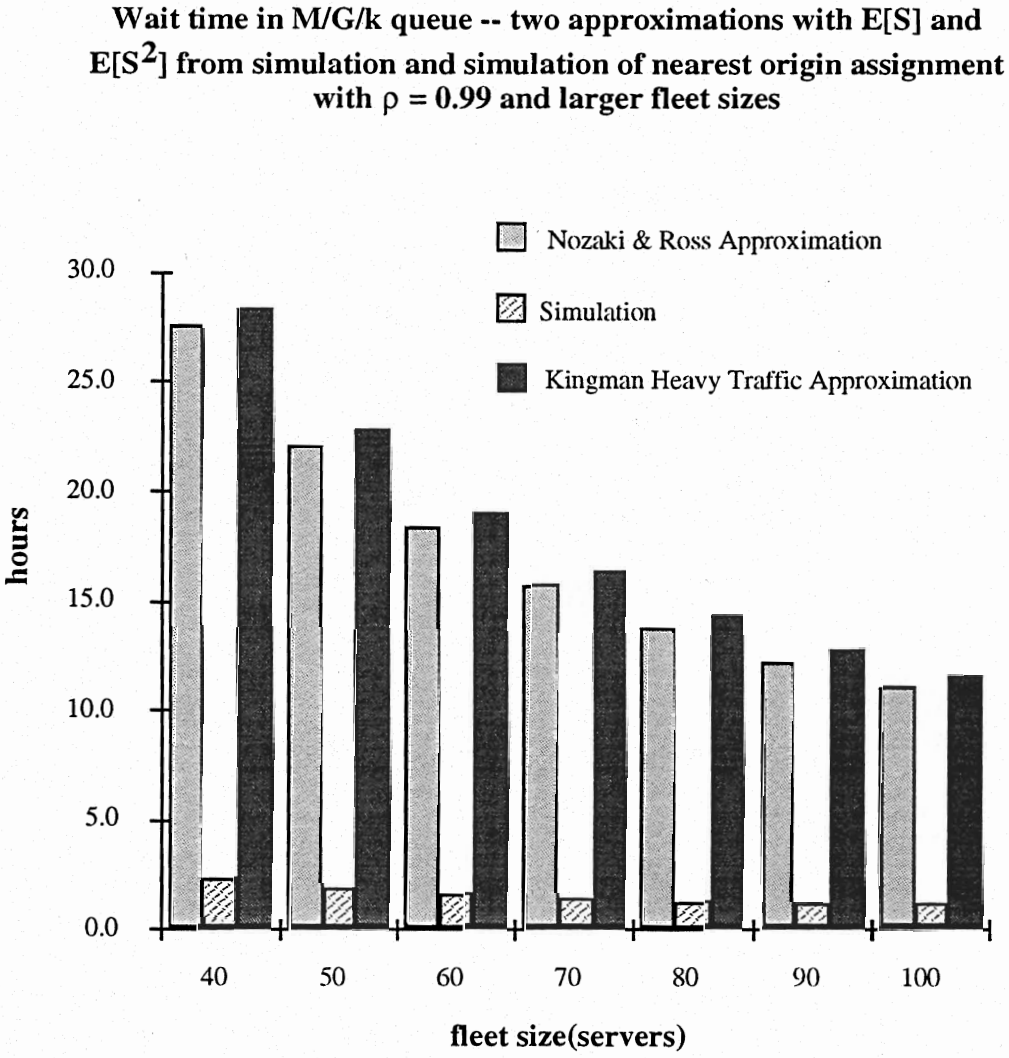


Figure 4.13 Average wait time in queue from simulation of nearest origin assignment and application of N&R and Kingman approximations with utilization = 0.99

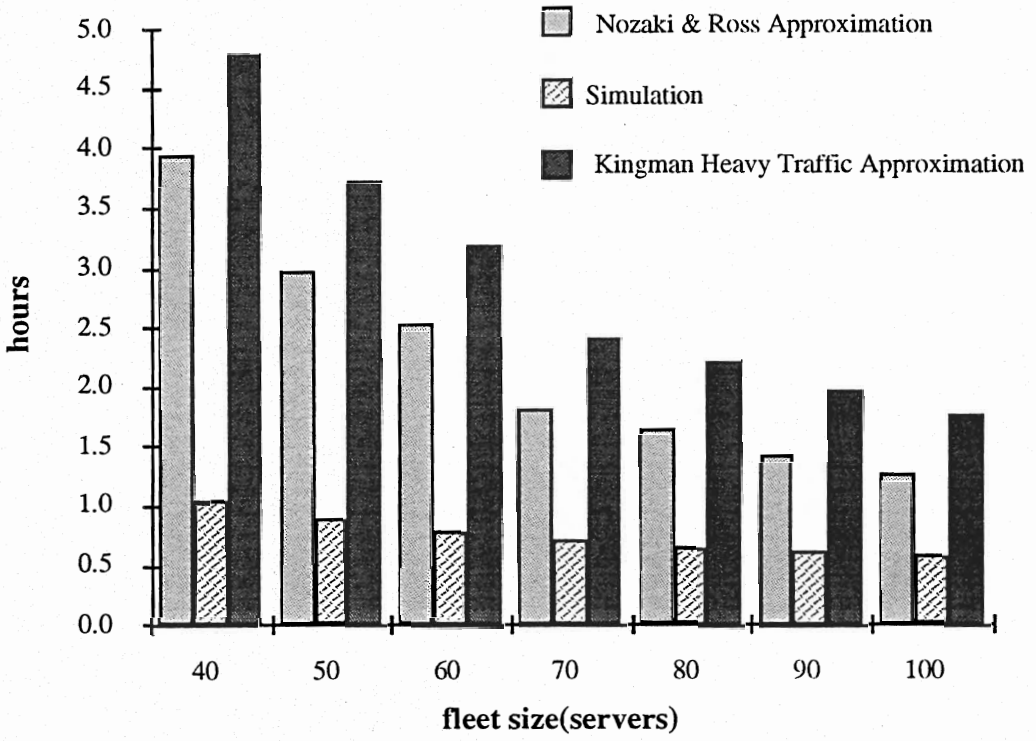


Figure 4.14 Average wait time in queue from simulation of nearest origin assignment and application of N&R and Kingman approximations with utilization = 0.97

$E[Sno] = \sum_{\ell=0}^{\infty} P_\ell E[S_{E_{no}} | L = \ell] + E[S_L]$ would need to be solved.

Although (4.24) was derived for the single server case it may be generalized to those with multiple servers. Solving (4.24) requires the estimation of the long run probabilities that a certain number of loads will be in the system. Estimating these long run probabilities in even an M/G/1 system is not trivial. While in the single server case this may possible, using fairly extensive numerical integration (Van Hoorn [1984] p.101), when the service time function is well defined, and, more easily in the special cases of deterministic, Erlang and Exponential service time functions, the first and second moments of the service time alone provide insufficient characterization of the function. In an M/G/k system these values may be measured but not

approximated in closed form. Without attempting this directly we present the following simulation results.

A fleet of 300 vehicles are simulated in a circular work area. Demands are generated from a Poisson point process at a rate proportional to the number of vehicles. A queue of up to 1500 customers is allowed to form over a horizon in which between 2400 and 3600 customers are served by each vehicle. The simulation is limited to demand arrival rates in which no loads are turned away. The average wait time and number of customers in queue is measured, after a substantial start up period (1200-1800 customers served per vehicle). Figures 4.15 and 4.16 show that E[S] is highly sensitive to average number of customers awaiting service.

While the systematic dependency between the number of customers waiting in the queue and the rate of service under the nearest origin assignment rule may be an extreme case, many systems will exhibit an increase in efficiency as the density of customers and servers increases.

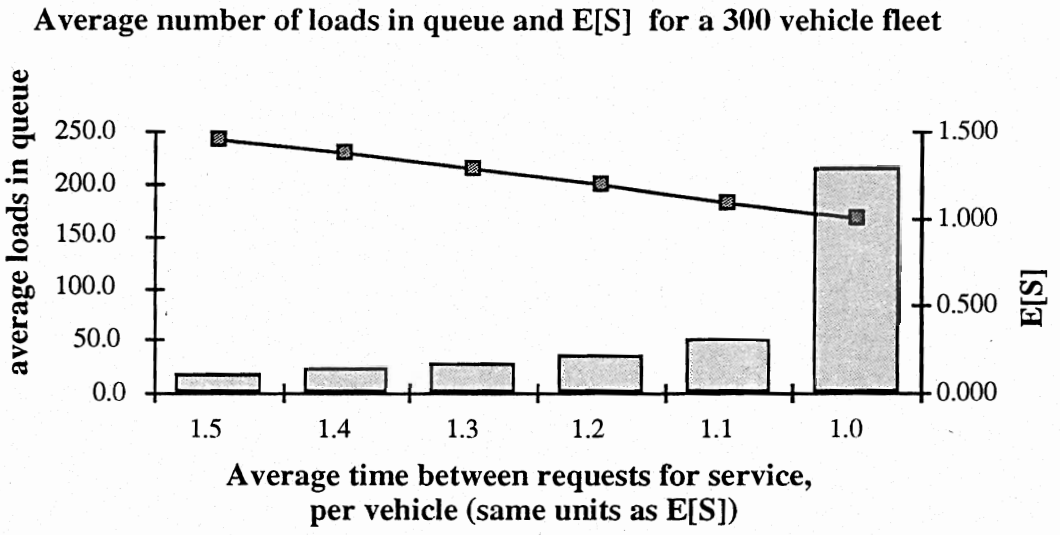


Figure 4.15 $\overline{L}$ and $E[S_{no}]$ for highly congested systems

SUMMARY

A common question throughout this chapter is the question of how, in a system in which customer requests arrive over time, to provide service in a timely fashion to all customers and immediately to those requests that are particularly time-sensitive. The first section of Chapter 4 introduced a strategy of diverting a vehicle en-route to make an immediate pick-up of a more time-sensitive load, or of a load that (when sequenced first) will improve the efficiency of the vehicle's travel route is introduced. Simulation results which extend the analysis are discussed in

Chapter 6. Moving from a single vehicle to fleets (and subfleets) of various sizes, the increase in the ability of a fleet to respond to time-sensitive demands under real time information is estimated, again with the help of simplifying assumptions.

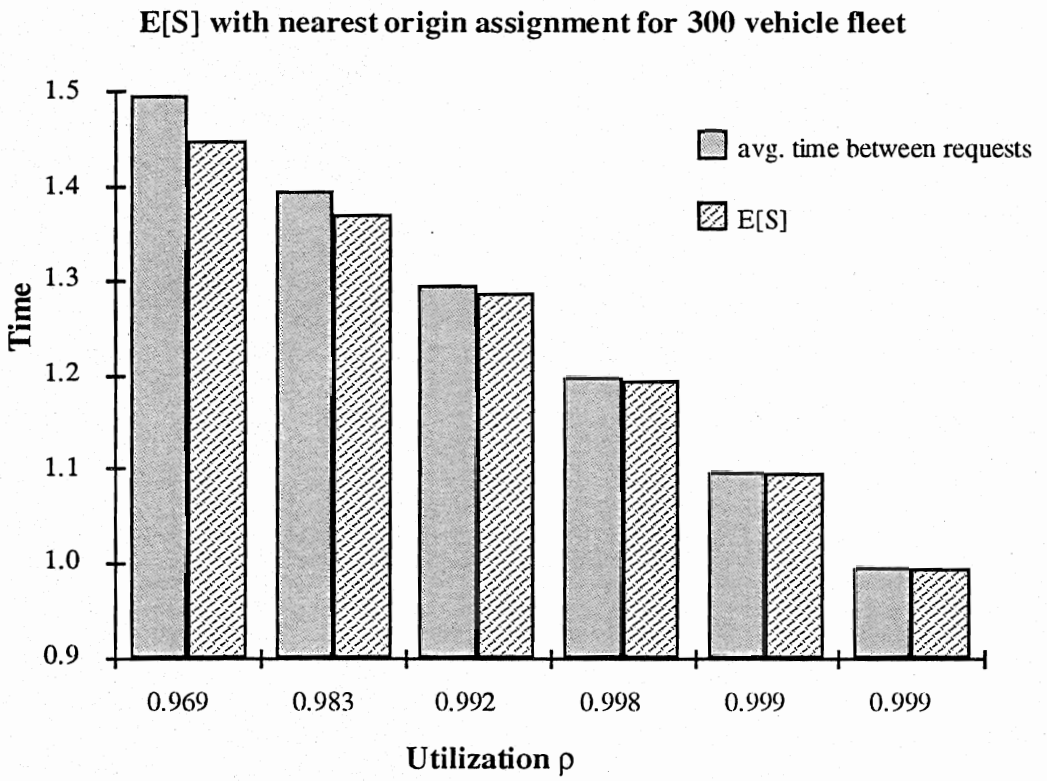


Figure 4.16 $E[S_{no}]$ with nearest origin assignment

Carrier fleet operations are modeled as a distributed queuing system. A system with Poisson arrivals and a general service distribution is examined. An approach for generating an upper bound on the efficiency (measured as the average wait time for service to begin) for a fleet of vehicles is developed and the conditions under which this bound is relatively tight are discussed.

The next chapter describes the design of experiments used to analyze the performance of the operational strategies described in the previous chapter.

CHAPTER 5 EXPERIMENTAL FRAMEWORK AND DESIGN

In this chapter, the simulation framework developed to evaluate the performance of the dynamic dispatching and load acceptance strategies of interest is discussed, and the specific strategies examined are described. The use of simulation makes possible the evaluation of these strategies under a wide spectrum of scenarios with respect to realizations of demands for services, information availability, and communication capabilities. In addition, a carrier fleet operation may be simulated over a long time horizon and the *expected* performance of strategies estimated. While the experiments performed in this research may lack realism in certain respects, the simulation framework provides a testbed for exploring important questions and for defining and investigating operational strategies for fleet management under real-time information. In this analysis, five base case assignment strategies are compared with four local assignment strategies which rely on continuous dispatcher to driver communication and position updates. An assignment strategy, when applied in conjunction with a load acceptance strategy, defines the operational strategy for the given fleet. Both the real-time and base case operational strategies examined are relatively simple, relying solely on current (not forecast) information. These operational strategies were presented in detail in chapter 3.

The simulation framework is deterministic with stochastic inputs. Results for the same input data will be identical in multiple simulation runs. Individual Monte Carlo realizations are randomly generated and the results aggregated over a sufficient number of simulation realizations to provide statistically robust estimates of the performance measures of interest. The performance of vehicle fleets under the operational strategies is studied over long service horizons to gain insight into the average or steady state behavior of each strategy.

Chapter 5 describes the operational strategies examined. Introduced in chapter 3, two of the real-time assignment strategies investigated allow the re-assignment of previously assigned loads to other vehicles in addition to the resequencing of loads already assigned within an individual vehicle's queue of assignments. The others allow the resequencing of previously assigned loads within an individual vehicle's queue only.

Chapter 5 also provides a description of the simulation framework including input parameters and results reported. A more detailed description of the simulation from a procedural point of view is provided in Appendix I.

In Chapter 5 four sets of experiments are outlined. This evaluation seeks to: 1) Investigate the performance of the base case strategies under different demand levels and implementation

scenarios, including, quasi real-time applications of the base case assignment strategies; 2) Evaluate the relative performance of real-time assignment strategies; 3) Compare the real-time assignment strategies to the base case assignment strategies; and, 4) Estimate the benefits of real-time information. The assignment strategies examined include some scenarios that consider explicit service (pickup) deadlines and some that do not. This research is most interested in time-constrained operations. The unconstrained cases are examined to explore the tradeoffs associated with honoring and offer an additional benchmark for measuring system efficiency.

As stated in chapter 1, the main hypothesis is that real-time information on vehicle locations and demands can increase the efficiency of carrier fleet operations with respect to measures of trucking company profitability and responsiveness to customer requests. A related hypothesis is that real-time assignment rules perform well, with respect to those requiring less real-time information and communication, under certain conditions with respect to fleet size, level of demand and pickup deadlines. In order to test these hypotheses and to identify those conditions under which real-time assignment strategies perform better or worse than their less information intensive counterparts, the comparisons described are performed.

OPERATIONAL STRATEGIES EXAMINED

A set of operational strategies are examined; each of these includes one of three load acceptance strategies and one of nine assignment strategies. The load acceptance and assignment strategies are outlined here. Figures 5.4 and 5.5 provide diagrams of the full operational strategies examined under simulation. Assignment strategies (formulated in Chapter 3) are discussed in Chaper 5, before load acceptance strategies (discussed in Chaper 3). This order is chosen because two of the three load acceptance strategies are applied after a feasible, least cost assignment has been identified. As explained in Chaper 5, the rule used to determine the "least cost" solution varies across experiments.

Assignment Strategies

Assignment strategies are classified as either base case strategies or real-time information strategies. The separation is admittedly somewhat arbitrary. A strategy is defined as a base case strategy if, once an assignment is made, it is carried out with no changes in either the vehicle assignment or the order in which service will be provided by the vehicle. The real-time operational strategies require continuous updates on all vehicle locations and demands; in some cases these allow load to vehicle assignments to change as conditions unfold. In the operational

strategies examined, only the real-time information strategies have the ability to take pickup deadlines (or time windows for service) explicitly into account.

Two of the five base case strategies are intended to provide a benchmark for real-time assignment systems. The "first called first served" assignment method should provide an upper bound on reasonable assignment rules; the generation of the optimal traveling salesperson tour through points representative of those served in a week provides a lower bound on the distance traveled to provide service. Descriptions and where applicable, formulations, of the base case assignment strategies were provided in chapter 3.

Base Case Strategies First Called First Served (FCFS) . This strategy assumes that loads are assigned to available vehicles in the order in which they arrive. Service requests are added to a queue of requests upon arrival; when a vehicle becomes available it is assigned the first load in the queue. If one or more vehicles are idle when the request arrives, it is assigned to the vehicle that has been idle longest. Drivers must contact the dispatch center upon completion of service and the dispatch center must be in communication with the driver that has been idle longest.

Nearest Origin Assignment (NO) Accepted service requests enter the pool of unassigned loads. Upon completion of an assignment, the driver contacts the dispatch center for a new assignment, at which time an assignment is made to the nearest accepted and unassigned load. Loads arriving when one or more vehicles are idle are assigned to the nearest idle vehicle. Drivers must contact the dispatch center upon completion of service and the dispatch center must be in communication with all idle drivers.

Classical (Bipartite) Assignment. The fundamental strategy examined is: 1) Loads accumulate over time in a pool of accepted loads; 2) At specified decision points, loads are assigned to idle vehicles, or in some cases to vehicles that will become idle at some time in the future, prior to the next assignment point. Exactly m assignments are made where m = min{ available loads, available vehicles }; the assignment that minimizes the overall distance from the current (or next available) location of the vehicles to the origin locations of the loads is chosen. As discussed in chapter 3, two separate applications of this assignment technique are examined; these are considered separate assignment strategies because their information requirements differ. In the first case, the trigger for assignment is the passage of time, while in the second, assignments are triggered by the state of the system, where the state refers to the ratio of idle

vehicles to available loads. These two cases are referred to as time-based and state-based bipartite assignment (BAT(a) and BAS(b)), respectively.

Time-Based Bipartite Assignment (BAT(a)) Assignments are triggered by fixed, evenly spaced assignment points. Fixed assignment points are separated by a length of time aDL where a is a real number in the interval (0, 2) and DL is the average duration of loaded moves.

State-Based Bipartite Assignment (BAS(b)) Assignments are triggered when the number of loads awaiting service is equal to or greater than a multiplier b times the number of idle vehicles, or, when the number of idle vehicles is equal to b times the number of waiting loads (for example, when the number of loads exceeds {0, 1, 2,..} times the number of idle vehicles). In general, b need not be an integer; however, in our analysis we examine integer values of b only. In the case b = 0, assignment is performed whenever any loads are waiting for service and at least one vehicle is idle. BAS(0) approximates nearest origin assignment.

Cases examined are chosen to highlight the tradeoffs between immediate and delayed assignment of vehicles to loads. Time-based bipartite assignment (BAT(a)) with a close to zero approximates nearest origin assignment. In addition, a system in which "look ahead" is permitted is examined. When "look ahead" is allowed, vehicles that will become idle within a fraction of the time between the current and the next scheduled assignment are included in the current assignment. Comparing the performance of BAT(a), with and without look ahead, allows for the examination of the tradeoffs of including more vehicles in the current assignment, but in which future opportunities may be excluded.

State-based assignment (BAS(b)) represents a quasi-real-time implementation of the bipartite assignment heuristic in which assignments are made when the system reaches a pre-specified state. This adaptive assignment method exhibits several advantages. It leads to better utilization under moderate demands; both the expected value of wait times for service and the variability of these times are lower than in the time-triggered assignment case.

Asymmetric TSP (ATSP). An asymmetric Traveling Salesperson Problem is solved for a set of loads and a single vehicle. This case is used to generate a benchmark for the real-time information cases. A problem based on a set of randomly generated loads of approximately the same number of loads served per vehicle per week is solved and the expected distance traveled under a perfect hindsight (or perfect look-ahead) assignment estimated. The objective is to minimize the empty distances traveled. This bound is compared to the performance of strategies

applied when loads become known over time. A mathematical formulation for the ATSP is provided in chapter 3.

Assignment Strategies Under Real-Time Information. Four different strategies are examined. These all require real-time information on the status and location of vehicles and demands, and differ in the extent to which assignments are flexible with respect to re-ordering and re-assigning loads. Figure 5.1 illustrates the process followed under all four strategies. In all cases, when a service request arrives, the new load is assigned immediately to the least-cost, deadline-feasible vehicle. As discussed in chapter 3, these assume that for each vehicle, a TSPTW sub-problem is solved for the vehicle's current load assignments and the candidate load. The empty distance associated with the feasible ordering which minimizes the empty distance traveled is the cost associated with the assignment. Then, one of three decision rules is applied to make the final load to vehicle assignment. If the load can be served within its pickup constraints, it is assigned to the feasible vehicle for which the addition of the new load results in the:

Least Empty to Loaded Ratio Assignment (ELR). lowest empty to loaded ratio,

Least Overall Empty Distance Assignment (SED). overall empty distance to travel,

Least Additional Empty Distance Assignment (DED). or, the least increase in empty distance to travel.

The four assignment strategies differ with respect to two factors, (1) whether or not en-route diversion is allowed, and (2) whether or not re-assignment of loads from one vehicle to another is allowed. They are described next.

No-en-route diversion or re-assignment of loads **(D^CR^C)**

In the first case, it is assumed that once loads are assigned to a vehicle they will be served by that vehicle and that once a vehicle begins moving empty towards a pickup location it will continue on to make that pickup next.

En-route diversion only **(DR^C)**

This alternative allows for the diversion of an en-route vehicle to pick up another load, provided the same vehicle is able to provide service to the original load within its specified time constraints.

Re-assignment of loads **(D^CR)**

In this case, en-route diversion is not allowed. However the re-assignment of currently assigned loads to alternative vehicles is allowed. The generation of completely new solutions is not examined; rather, a much simpler re-assignment strategy in which loads are considered, one at a time, for re-assignment to other vehicles is examined.

The re-assignment rule is the following: after an assignment of a newly arriving load is made to a vehicle in the system, the last load assigned to each vehicle which has more than two loads assigned is a candidate for re-assignment to another vehicle. The load is removed and becomes a candidate, as though it were a newly arriving load, for re-assignment. This re-assignment process removes at most one load from each vehicle at each re-assignment point (the point directly following the assignment of a newly arriving load).

En-route diversion & re-assignment of loads **(DR)** This alternative allows both the diversion of en-route vehicles and the re-assignment of loads.

Load Acceptance Strategies

The load acceptance strategies, outlined in chapter 3, are discussed here in the context of the experiments performed. In all cases, load acceptance decisions are made immediately after the arrival of the request for service in the system. The load acceptance strategies are listed in order of increasing restrictiveness. In the base cases, only the system capacity check is used. In the real-time strategies, in scenarios without time constraints (pickup deadlines), the system capacity check is all that is required; with time constraints, the feasibility based load acceptance process is invoked. In some cases, profit based load acceptance, an extension of the feasibility based acceptance process is used to reject less attractive loads -- those that cannot likely be served economically.

System and vehicle capacity check prior to load acceptance/rejection. In all operational strategies examined, a simple capacity check is performed. In the base case strategies, if the number of loads waiting in the system exceeds a pre-specified number (generally five times the number of vehicles) then the load is refused service. In the corresponding real-time strategies, the maximum number of loads that may be assigned to an individual vehicle at once is limited by a maximum queue length. In scenarios examined, this number is five, so that, as in the base cases, the maximum number of loads waiting for service is five times the number of vehicles.

The other two load acceptance strategies require that the feasible, least cost, assignment of the candidate load to a vehicle be identified before a decision can be made. Under feasibility based load acceptance, if such an assignment can be found, then the load is accepted. Under profit based load acceptance, an additional requirement must be satisfied prior to the acceptance or rejection of a candidate load.

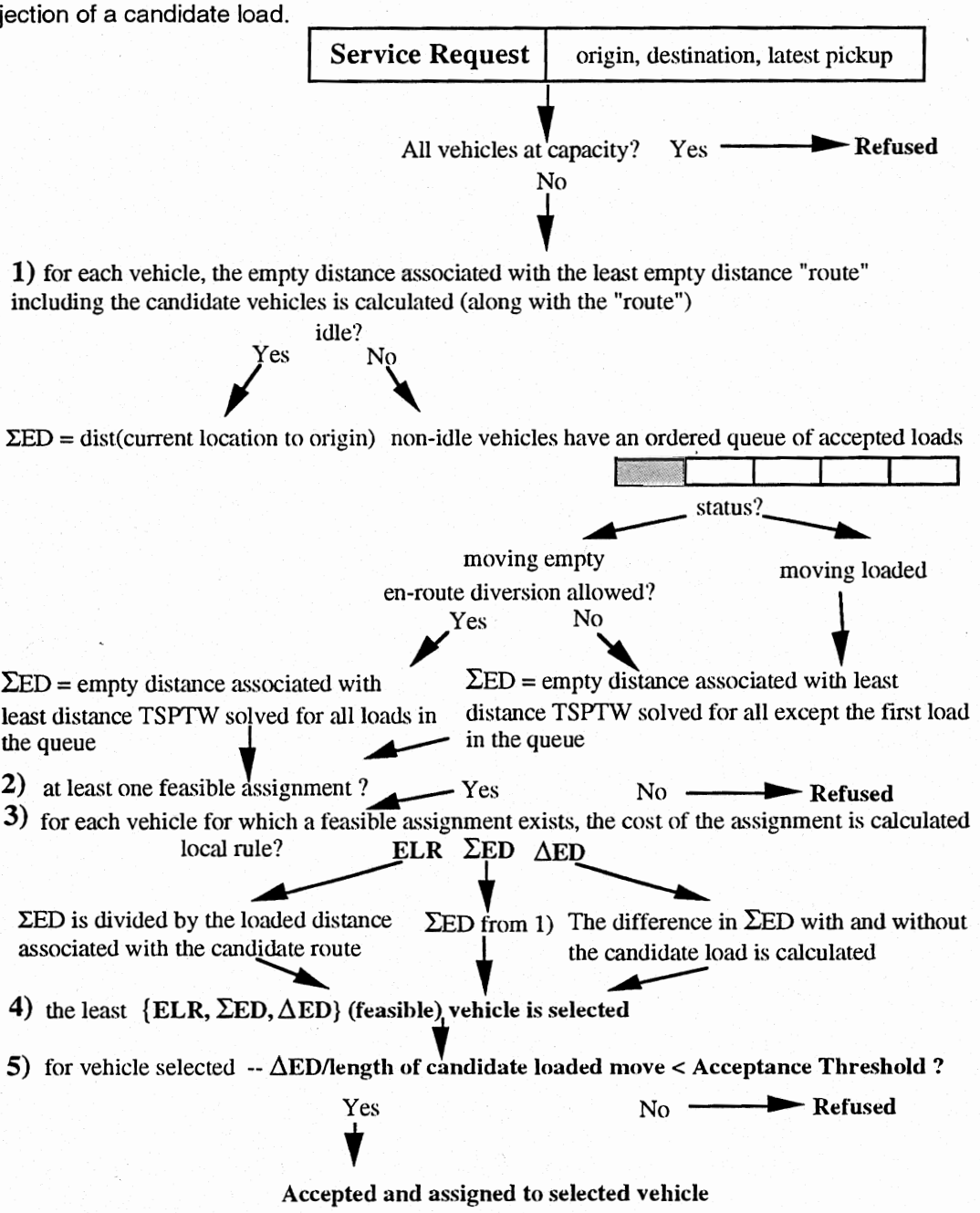


Figure 5.1 The process followed by the real-time operational strategies

Feasibility based load acceptance An explicit feasibility check is performed prior to load acceptance or rejection. Discussed in chapter 3, a deadline-feasible insertion point (including addition to the end of a schedule) is identified for each vehicle for which such an insertion point exists. If no feasible insertion point can be found, the load is refused service. The procedure for finding a feasible insertion point is equivalent to performing an assignment for the load. If the assignment is made successfully the load is accepted (and assigned). If no feasible assignment is found, the load is refused service.

Profit based load acceptance. Once feasibility has been established, in the manner described above, an estimate of the profitability of the candidate load is made in the following way: 1) The "best" deadline-feasible assignment of the load is found. 2) The ratio of the empty distance attributable to the load and its loaded distance is calculated, where the empty distance attributable is defined as the difference between the empty distance associated with the vehicle's route, with and without the candidate load. This estimate is compared to a threshold value, below which loads are refused service. Outlined in chapter 3, several acceptable proxies exist for profit estimation. Simulation experiments included in this analysis use an estimate of the ratio of the empty to loaded distances (E/L) attributable to a candidate load to make the acceptance or rejection decision. Loads with an E/L ratio higher than the threshold value are rejected. The long run E/L ratio for the system varies from 0.08 to 0.50, with most values around 0.30. Threshold values applied should be sensitive to the congestion level of the system The values used in this analysis vary from 0.5 to 1.2. Loads with an (attributed) E/L ratio higher than the threshold value are rejected. A threshold rule with parameter 1.2 would accept all but the most inconvenient loads into the system, while a rule applied with parameter 0.5 would reject all but the loads that fit well with already accepted loads.

Under the real-time operational strategies, whenever a load is accepted into the system it is immediately assigned to a vehicle. In fact, the load acceptance and assignment processes are coupled. As discussed in Chaper 3, while these processes need not necessarily be coupled, this coupling ensures that future load acceptance decisions do not jeopardize the feasibility of already accepted loads.

SIMULATION FRAMEWORK

The simulation allows tests of alternative operational strategies under different scenarios regarding the demand pattern and information availability. In this section, the principal elements and defining parameters of the simulation framework are described. In addition, the profit model

used to evaluate the profitability of the operational strategies is discussed, followed by an explanation of those simulation parameters that vary across experiments and the method used to verify the statistical significance of results.

The principal elements of the simulation are the dispatch center, from which all decisions are made, a set of vehicles, and a set of requests for service.

Principal parameters of the simulation are:

- Demand arrival pattern
- Number of vehicles
- Load acceptance strategy
- Assignment strategy

The dispatch center operates according to an operational strategy, including a load acceptance strategy and an assignment strategy (which may also include the re-assignment of previously assigned loads).

Service requests are randomly generated according to a specified space-time stochastic process (or according to a pre-set schedule), and are characterized by the following attributes:

time of request

a load origin location (x, y or lat, long coordinates)

a load destination location

a service time window (w_e, w_l, for earliest and latest pickup times). In all cases it is assumed that the earliest time of pickup is the time at which the request was received.

At each instant, each vehicle has an associated status, which may be one of the following:

moving loaded,

moving empty,

idle and available to accept assignments,

or unavailable to accept assignments.

Only the first three of these are used in the simulation experiments discussed. Vehicles are assumed to be in service continuously.

High Level Specifications

The simulation experiments performed are governed by the following high level specifications.

i) The geographic region is a circle of radius 250 miles.

ii) The travel metric is Euclidean.

iii) The demand arrival process is a Poisson point process, origin and destination locations are uniformly and independently distributed over the circular work area.

iv) It is assumed that travel takes place at a constant speed of 50 miles per hour.

v) As a result of i) - iii) the average distance traveled loaded is approximately 226 miles. Applying iv) the average duration of a loaded movement is 4.525 simulation hours.

vi) Each week contains 70 hours of work time. All vehicles are in service (but sometimes idle) at all times. The maximum distance that can be driven in a week is therefore 3500 miles.

The profit model applied is another high level specification of the simulation framework. Because it requires more clarification than i) - vi), a section in the next chapter is devoted to its explanation.

Implementation of the Profit Model

Chapter 3 introduces a general operating profit model for carrier fleet operations in which profit is equal to revenue earned (which includes both a fixed portion and a portion proportional to loaded distances traveled) minus a set of operating costs. A simple expression for the operating profit is given below.

$$\text{Profit} = \sum_{\text{over all loads served}} \begin{bmatrix} \text{Revenue} \\ -\ \text{(empty travel cost)} \\ -\ \text{(loaded travel cost)} \\ -\ \text{(handling cost)} \\ -\ \text{(daily vehicle charges)} \\ -\ \text{(daily driver charges)} \end{bmatrix}.$$

The model implemented in the simulation experiments is closely related to this model but makes the assumption that fixed revenue earned per load served is approximately equal to handling costs. This simplifies the model so that only revenue generated per distance traveled, empty and loaded travel costs, and weekly vehicle and driver costs are included. The following parameter values are assumed.

Revenue earned per loaded distance traveled = $1.20 per mile

Cost incurred per distance traveled (empty or loaded) =$0.57

(It is assumed that $0.30 would be paid to the driver and that $0.27 would be the marginal cost per mile for using the vehicle).

The fixed cost per week for each vehicle is $300.0

The fixed cost per week for each driver is $300.0

While a serious effort has been made to develop a simple but reasonable cost model, many different values within a range would have been acceptable. The most recent version of American Trucking Trends (American Trucking Associations [1996] p. 22) estimates that in 1995 the overall cost per mile to provide service was $1.302. Excluding interest, depreciation, management and overhead, $0.915 of this amount can be attributed to operating cost, while the remaining $0.387 is attributable to fixed cost or overhead. Values provided are aggregated over both the truckload and less-than-truckload (LTL) segments of the market. With industry profit margins typically two to four percent (American Trucking Associations [1996], Association of American Railroads [1992]), average revenue earned per (loaded) mile, would need to be in the range of $1.60 and $1.80 for an operation to remain profitable.

Both the cost and revenue associated with truckload operations are less than those associated with LTL operations. In the cost model implemented, a fleet working at %50 utilization would incur operating costs of $0.91 per mile while at %100 utilization the cost is $0.74 per mile.

The $1.20 value selected as a representative revenue earned per mile is reasonable for a truckload operation.

In the simulated system, a driver at 100% utilization would travel 3500 miles per week. Figure 5.2 shows how the break-even point for operating profitability varies with utilization and the fraction of time spent performing revenue generating work. Time spent moving empty is included in the fraction of time utilized, but only the fraction of time spent loaded generates revenue. For example, under this model a fleet that is 100% utilized can remain profitable when only 62% of its time is spent moving loaded (an E/L ratio of 0.61), while a fleet that is only 50% utilized must spend about 78% of its travel time loaded (an E/L ratio of 0.28).

Premiums for Meeting Pickup Deadlines Systems in which loads have associated pickup deadlines that must be met will usually operate at lower utilization levels (for the same overall demand) than those in which deadlines are either non-existent or non-binding, since some service requests must be refused. In those systems a premium is assumed to be earned depending upon how the stringent the deadline is. In the simulation experiments systems investigated have the following deadlines {2, 4, 6} or {4, 8, 12} hours where the average loaded move has a duration of approximately 4.5 hours. Premiums are fixed charges that vary with the pickup deadline requested. Premium charged are: {$68.88, $54.30, $40.73, $27.15, $13.58} for {2, 4, 6, 8, 12} hour pickup deadlines. These values correspond to a percentage {25, 20, 15, 10, 5} of the price charged for the average loaded move.

Variable Parameters

The vector {la, a, k, h, r, **w**, i} specifies the variable parameters of the simulation.

la is the load acceptance rule applied,

a is the assignment rule used,

k is the number of vehicles in the fleet,

h is the simulation horizon (in simulation weeks),

w is a vector of pickup deadlines and associated fraction of requests in each deadline category,

r is the rate of arrival of requests for service per vehicle, and,

i is the minimum number of simulation iterations over which results are aggregated.

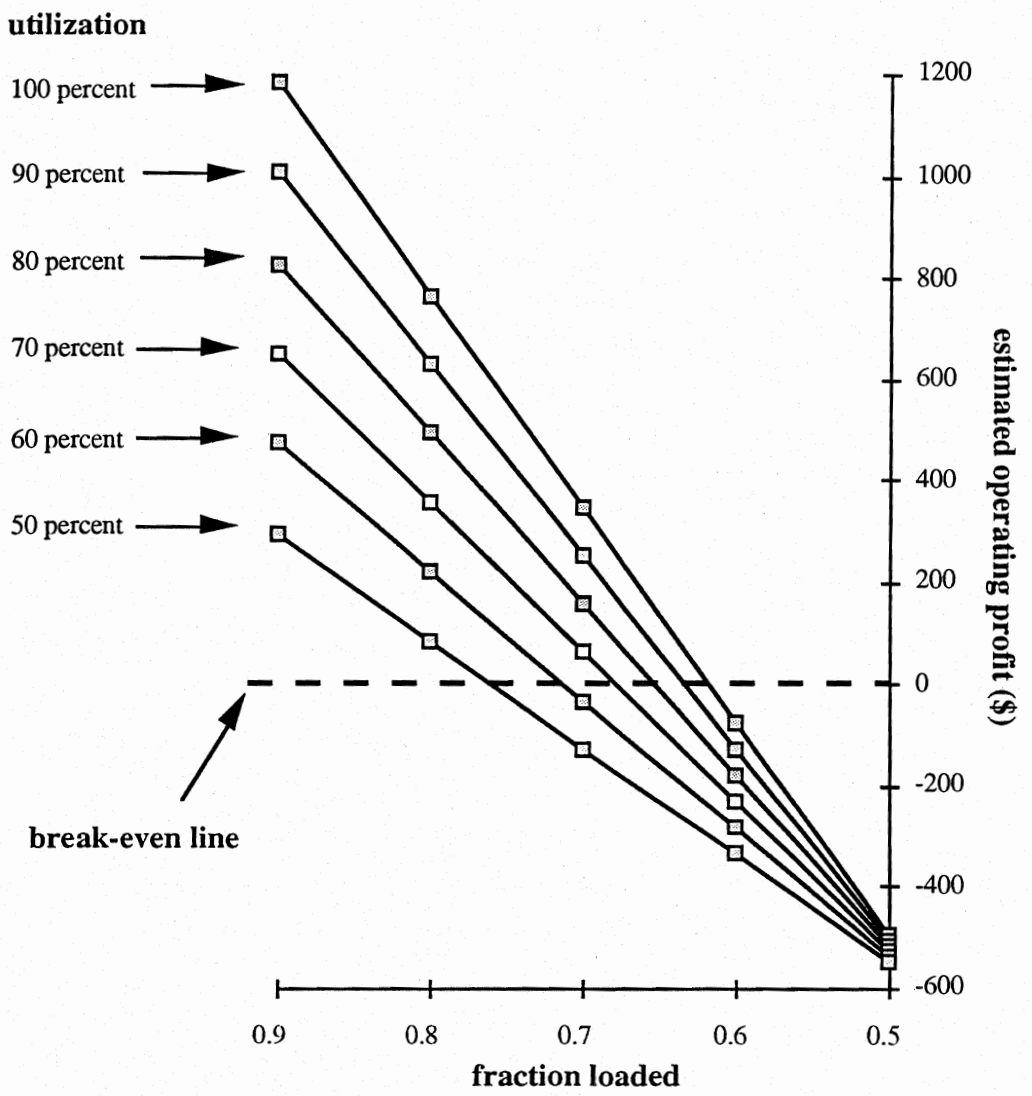


Figure 5.2 Variability of the break-even point for operating profitability as a function of the utilization and revenue producing work performed

Each of the input parameters may take on several values. While many values are possible, analysis presented here is limited to the following set of values. The range of values is intended to provide opportunities for examining the operational strategies of interest under different conditions, primarily with respect to intensity of demand. In addition, time constraints imposed

range from none at all to moderate to very restrictive. These are designed to lend insight into the behavior of systems with more or less underlying flexibility and to allow for the testing of flexible dispatching strategies under varying conditions.

la: Three choices are examined, these represent capacity check, feasibility check, and profit based load acceptance.

a: Nine choices are defined in chapter 5 as

{FCFS, NO, BAT(a), BAS(b), ATSP, D^CR^C, DR^C, D^CR, DR}.

k: The fleet sizes examined are {1, 2, 5, 10, 20, 50, 100, 300}

vehicles.

h: The length of the horizon is limited to {26, and 2600 } simulation weeks.

w: The vector of pickup deadlines and associated fraction of demand in each category takes on three values, {tight, medium, none} corresponding to $\{(\text{no deadlines};\ 1.0),$

$(2,\ 4,\ 8 \text{ hours};\ \frac{1}{3},\frac{1}{3},\frac{1}{3}),$

$(4,\ 8,\ 12 \text{ hours};\ \frac{1}{3},\frac{1}{3},\frac{1}{3})\}$

r: The average rate of arrival of service requests is assigned one of three values, {rapid, medium, slow}. In the case with Poisson distributed arrivals the rates selected are: {0.200, 0.1333, 0.100} service requests per vehicle per hour. Alternatively expressed as {5, 7.5, 10} simulation hours between requests, on average, per vehicle.

i: The minimum number of simulation iterations (consecutive, independent realizations) is 100. Where possible simulations are performed for a larger number of iterations.

Results Reported

Chapter 3 discusses a set of objectives held by carrier fleet operators and table 3.1 outlines associated performance measures of interest. The performance measures in the three categories of interest, namely, profit, customer service and other operational considerations are listed. These measures are:

Profit Measures

- Revenue generated by the fleet per week and by each vehicle per week.
- Operating costs incurred, per week and per vehicle per week.
- Operating profit generated by the fleet per week and per vehicle per week.
- Ratio of time spent empty to time spent loaded and ratio of time spent idle and empty to time spent loaded.

- Overall fraction of time vehicles spent moving empty, moving loaded and idle.
- Mean and standard deviation of the length of loaded, empty and combined loaded and empty movements.
- Fraction of loads served falling into each of the pickup deadline categories. For example, if $\mathbf{w} = \{4,\ 8,\ 12 \text{ hours};\ \frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$ then this result might be reported as {0.321, 0.339, 0.339} if more tight pickup deadline loads were refused service than those with less tight pickup deadlines.

Customer Service Measures

- Mean and standard deviation of the wait time for service. (Time between arrival of the request and pickup)
- Mean and standard deviation of the number of customers waiting for service.
- Overall fraction of service requests accepted.
- Fraction of loads served falling into each of the pickup deadline categories.
- Fraction of pickup deadlines missed (for assignments not respecting deadlines explicitly)

Other Measures

- Mean, and standard deviation of loaded and empty distances traveled for vehicles across the fleet. This is a measure of the equality of assignments across vehicles.
- Fraction of assignments or service requests effected by particular sub-strategies. For example, if loads are rejected, not for infeasibility exclusively but because a profit based load acceptance rule is applied, then the fraction of requests for which this is the case are reported. Similarly, if en-route diversion is allowed, the fraction of assignments involving en-route diversion are recorded.

A subset of these are the focus of analysis of simulation results addressed in chapter 6. These are listed here, roughly in order of importance.

- Mean and standard deviation of the average length of loaded, empty movements.
- Mean and standard deviation of the average wait time for service.
- Ratio of time spent empty to time spent loaded.
- An estimate of the operating profit generated per vehicle per week.
- Fraction of loads served falling into each of the pickup deadline categories.
- Fraction of pickup deadlines missed (for assignments not respecting deadlines explicitly).
- Mean, and standard deviation of loaded and empty distances traveled for vehicles across the fleet.
- Fraction of assignments or service requests effected by flexible assignment strategies.

EXPERIMENTS PERFORMED

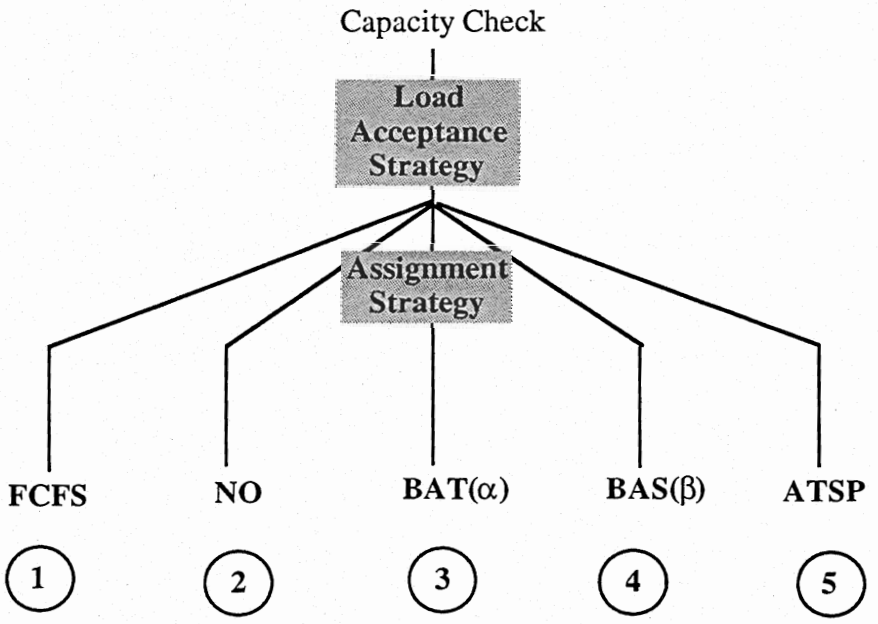
This section contains a description of the experiments performed. Figure 5.3 contains a diagram of the so-called base case operational strategies examined while figure 5.4 shows the real-time operational strategies examined. Each branch on the trees displayed shows a different operational strategy. Most of the branches in the tree uniquely define a method of handling the acceptance and assignment of requests for service. Notable exceptions to this rule are the base cases BAT(a), BAS(b), which actually specify a family of assignment strategies which vary with different values of a and b, the parameters that determine the frequency with which assignments are made.

All of the experiments described here were performed with a simulation horizon of twenty-six simulation weeks. Unless otherwise specified, simulations are performed for a minimum of one hundred iterations.

Comparison A - Four Base Cases

In this analysis, the first four base cases are compared. Experiments extend the results discussed in chapter 4 in which first called first served assignment is compared with nearest origin assignment. Chapter 4 also describes a model of carrier fleet operations as an M/G/k queue. Results are drawn from experiments conducted with extremely high utilization levels and for fleets of up to three hundred vehicles. A simulated system is compared to analytical heavy traffic approximations for the M/G/k queue. The utilization rate in the system takes on values very close to one hundred percent in order to perform the comparison. While performance measures can be obtained at these utilization rates, there is little indication that such a system is stable or that the values are obtained at steady state. Simulation experiments here complement that analysis. They allow the estimation of performance measures in a more stable range, and, facilitate the comparison of base case assignment rules to other heuristics. Overall efficiency and the effects of moderate congestion on the wait time for service, size of queue and the length of service time (related to the empty distances traveled) are of interest. While it is assumed that serving customers as quickly as possible is a goal, explicit service deadlines are not enforced.

Tree of load acceptance strategies and assignment strategies (base cases)



FCFS = first called first served
NO = nearest origin
BAT(α) = time triggered bipartite assignment
BAS(β) = state triggered bipartite assignment
ATSP = assymmetric traveling salesperson problem

Figure 5.3 Tree of base case operational strategies

Tree of load acceptance strategies, assignment strategies and decision rules (real-time cases)

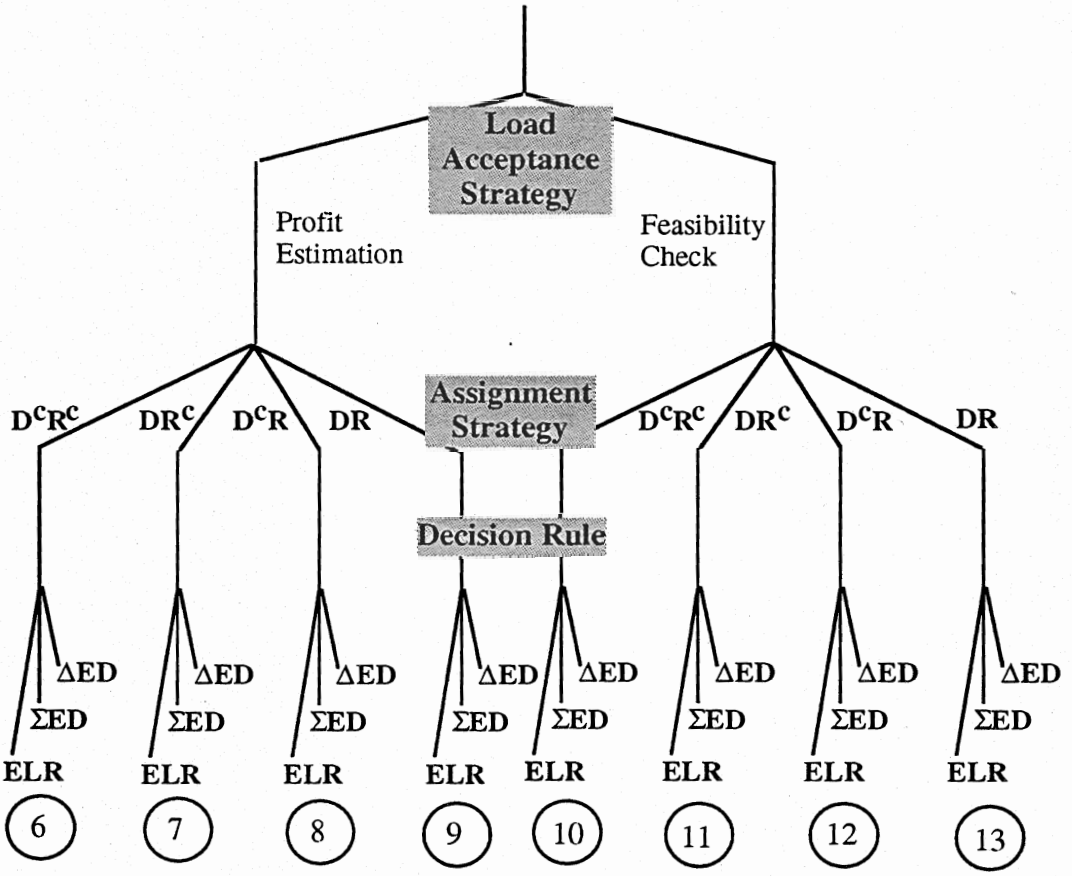


Figure 5.4 Tree of real-time operational strategies

The cases examined which include pickup deadlines are of interest as a comparison to cases in which deadlines must be satisfied. Cases examined for alternatives 1-4 (in figure 5.3) are shown in figure 5.5. Fleet sizes are limited to 10, 20 and 50 vehicles, not because of an inherent limitation in the simulation framework; rather, because the compact region served and the

demand arrival pattern which specifies that origin and destination locations are generated uniformly (and independently) on a circle makes an analysis with smaller fleets more meaningful than one with larger fleets.

The Effect of Limiting Pool Size. The capacity check for load acceptance specifies that the number of loads waiting for service is no more than 5 times the number of vehicles in service. In the scenarios chosen for examination, this is a binding limit in the heavy demand scenarios only. However, in heavy demand cases the limit on pool size can have significant consequences. The effect of this limit is examined in a set of additional simulation experiments which are discussed in chapter 6.

Immediate Versus Delayed Assignment. The tradeoffs between immediate and delayed assignment of loads to idle vehicles is another topic addressed in chapter 6 that does not fit neatly into the experimental design presented in chapter 5.. Both bipartite assignment strategies are of interest in this analysis and results of simulation experiments in which a, b, fleet sizes and intensity of demand vary are discussed.

Comparison B - Local Assignment Strategies Requiring Real-Time Information

The four local assignment rules outlined in chapter 5 are compared. In each case the performance of three decision rules which assign loads to drivers in which empty to loaded ratio, additional empty distance, and, the overall empty distance is least are examined relative to the performance measures outlined in chapter 5. Local assignment rules that incorporate real-time information on vehicle locations, the status of the fleet the location and characteristics of service requests and explicit pickup deadlines are of interest here. The performance of decision rules based on a set of objective function proxies is examined relative to a set of higher level objectives. Cases 6 to 9 in figure 5.4 are the focus of this examination. The complete set of experiments shown in figure 5.5 are conducted for alternatives 6 and 10, with selected experiments conducted for alternatives 7, 8 and 9 and for alternatives 11, 12 and 13. Alternatives 11, 12 and 13 are identical except that load acceptance thresholds based on the predicted profit associated with candidate loads are applied. As in the experiments described in the previous section, fleet sizes are limited to 50 vehicles because the service region selected for this analysis is compact. In addition, the computational requirements (time requirements) of simulation experiments, which are conducted for a 26 week horizon and are repeated a minimum of one hundred times for each scenario examined limit the size of the fleets. Incorporating promising re-

assignment heuristics into actual fleet management systems (which must accommodate fleets as much as two orders of magnitude larger) would not be computationally or practically infeasible.

Comparison C - Local Assignment Strategies Requiring Real-Time Information and our Base Cases.

In this analysis the performance of the real-time information cases is compared to the first four base cases outlined in chapter 5. As mentioned in the discussion of comparison A, although pickup deadlines are not explicitly handled in the base case scenarios, (alternatives 1-4 in Figure 5.3), experiments in which pickup deadlines are assigned but not honored are examined. Experiments conducted are shown in figure 5.5.

Comparison D - Local Assignment Rules and Solutions To Corresponding Asymmetric Traveling Salesperson Problems (Examination Of a Single Vehicle)

A single vehicle is examined here. Assignments do not have deadlines for pickup. Sets of randomly generated loads of the size of the number of loads typically served per vehicle per week are assigned and the average empty distance traveled to provide service compared to the average empty distance traveled under the real-time assignment rules, but in which demands become known over time.

CONVERGENCE CRITERIA

Two separate convergence criteria are used in the simulation. These serve different purposes. The first criterion is used to determine the point at which the system is approaching steady state; this criterion is applied within every iteration. The second convergence criterion is used to determine the point at which a sufficient number of iterations have been run, ensuring that the values of performance measures, aggregated over all iterations, have converged to their true average values. The iterations are independent. Random variables are drawn from a distribution beginning with a different (randomly generated) seed. Performance measures are obtained for each iteration and are aggregated over all or part of the simulation horizon. Then, they are aggregated over the set, or a subset, of iterations. If an individual iteration fails to meet the steady state convergence test, values of the average and standard deviation of wait time and of the average number of customers in the system and awaiting service are not included in the aggregate statistics provided at the end of the full set of iterations.

If convergence was not reached in any other individual iterations, the simulation would stop soon after the minimum number of iterations because the value of aggregate performance measures would never change.

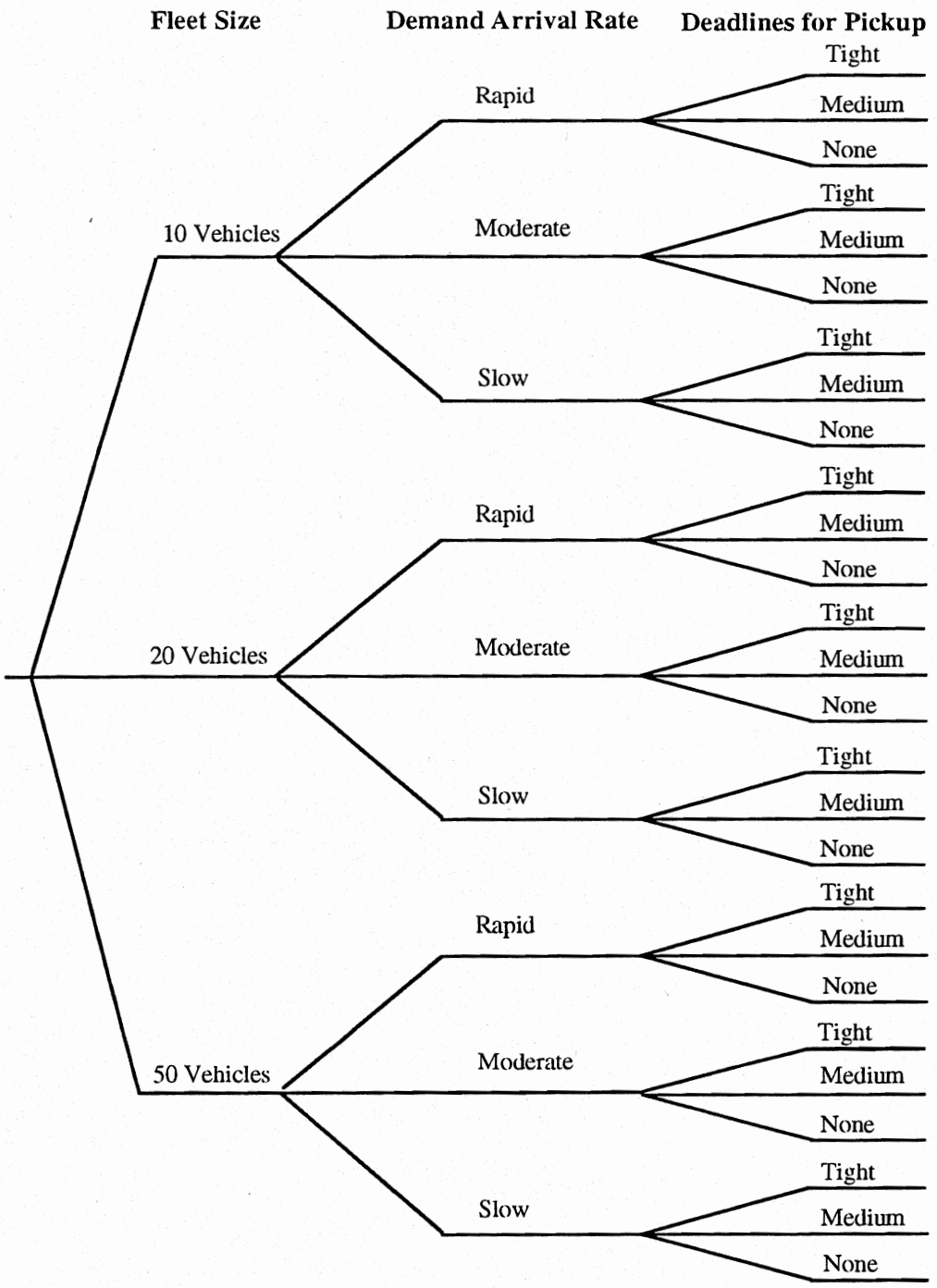


Figure 5.5 Set of experiments for comparisons A, B and C.

The steady state convergence criterion is the following:

1) The simulation horizon is split into ten even slices. The average wait time for service, a key performance measure, is calculated for each time slice. If the difference in the average wait time for service in consecutive time slices is less than five percent in three time slices, it is assumed that the system is *approaching* steady state and performance regarding the average and variability of the wait time for service and average number of customers in the queue are calculated from that time slice to the end of the horizon.

Figure 5.6 illustrates the convergence checking process.

Letting $\overline{w}_{i+1}$ represent the average wait for service in the $(i+1)^{st}$ time slice, and $\overline{W}_i$ the overall average over the first i time slices.

If $(\overline{W}_i - \overline{w}_{i+1}) / \overline{W}_i < 0.05$ a counter is incremented.

When the counter has been incremented three times then the span for the observation period for the average and standard deviation of wait time for service and average number of customers in the system and awaiting service begins. This observation period lasts until the end of the simulation horizon (the end of the iteration).

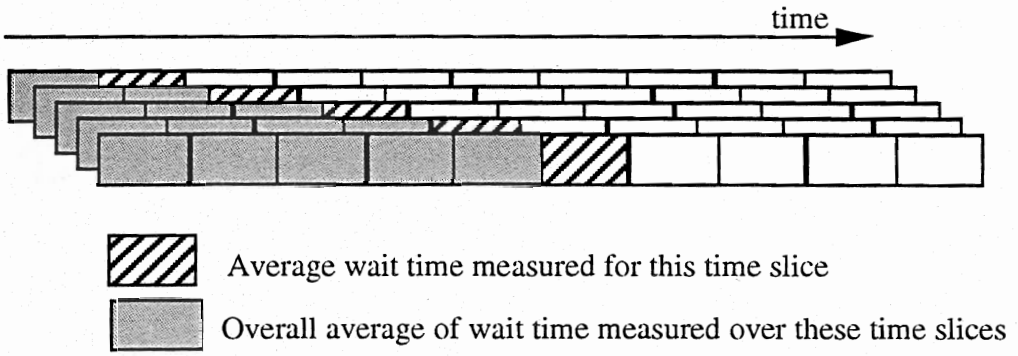


Figure 5.6 Diagram of application of first convergence criterion

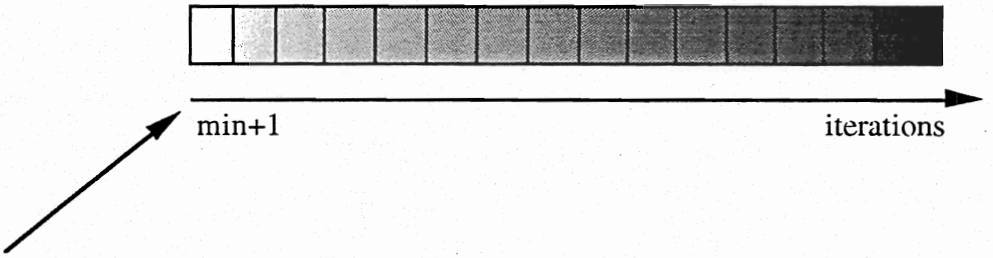
The stopping criterion is the following:

2) If, after the minimum number of iterations (typically 100), the values of key performance measures, aggregated over iterations performed up to that point, do not change by more than one tenth of one percent in five consecutive iterations, it is assumed that the system has converged. The performance measures tested are the average empty and loaded distances traveled over the iterations run so far. Typical simulation experiments converge within ten to thirty iterations after the initial one hundred. A diagram is provided for this second convergence criterion in figure 5.7.

Letting $\overline{E}_{i+1}$ and $\overline{L}_{i+1}$ represent the average empty and loaded distances traveled in the first i+1 iterations, a counter is incremented if

$$(\overline{E}_i - \overline{E}_{i+1})/\overline{E}_i < 0.001 \text{ and } (\overline{L}_i - \overline{L}_{i+1})/\overline{L}_i < 0.001 .$$

If this relationship is observed in five consecutive iterations then the simulation stops.



convergence checking begins after
the minimum number of iterations

Figure 5.7 Diagram of application of second convergence criterion

The second convergence criterion is binding less often than the first. Experiments that converge with respect to the second criterion do not necessarily converge with respect to the first.

For certain interarrival times and less consistent assignment strategies, only a small fraction of iterations pass the test of steady state convergence. The factors that affect whether the system reaches steady state or not are, level of congestion, fleet size, pickup deadlines, and, stability of the assignment rule. If the rate of arrival of requests for service is such that r, the ratio of the rate of arrival of requests (per vehicle) divided by the average service rate, is very close to 1.0, there is no guarantee that steady state will ever be reached. On the other hand, if the rate of arrival of requests for service is so high that the system is always running at its maximum allowable capacity and requests are being turned away then ($r > 1.0$ but $r_{experienced} \approx 1.0$), and as a result the system does display steady state behavior. Under heavy demand intensity, the limit on the number of loads that may be waiting for service at any time eliminates oscillatory behavior in the number of loads in the system.

Results discussed in chapter 6 are all drawn from experiments in which steady state convergence was obtained (in addition to the simulation-stopping criterion). Although it should be possible to construct scenarios in which convergence, as defined, is not obtained, for any number of iterations, no such cases were observed. While in some experiments steady state

convergence was not obtained when the minimum number of iterations was set at 100, when this was raised to a higher limit, generally 250, no non-convergent scenarios were observed.

TESTS OF STATISTICAL SIGNIFICANCE OF RESULTS

Where applicable the statistical significance of results is reported. In each set of experiments, the mean value of the performance measures is reported. The significance of the differences if these mean values for different scenarios is tested at significance levels of five and one percent with the following test for the difference between two means:

Letting m_1 and m_2 represent the mean values of a performance measure under the two different scenarios, $\bar{x}_1$ and $\bar{x}_2$ the sample means of the corresponding performance measures, s_1 and s_2 the sample standard deviations of the performance measures and m_1 and m_2 the number of simulation iterations performed.

Null Hypothesis: $H_o: \mu_1 - \mu_2 = 0$

Confidence Interval $P\left(\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1}{m_1} + \frac{s_2}{m_2}}} < z_\alpha \right) = 1 - \alpha$

With the exception of the number of miles driven over the simulation horizon, the standard deviation of all quantities of interest (expressed in simulation units, rather than physical units (miles, hours) is less than one. For simulations performed over one hundred iterations the denominator of the test statistic is no larger than 0.141. Differences in mean values of measured quantities (measured again, in simulation units, rather than physical units) of 0.328 units, or, (82 miles, 1.64 hours) are significant at a level of one percent. In some experiments clear trends may be inferred from the simulation results but the differences in some measured quantities (for example, the average empty distances traveled to provide service) may not be statistically significant. Where applicable this is noted.

SUMMARY

This chapter provides a description of the load acceptance, assignment and re-assignment strategies examined and the simulation framework used to investigate them. The four primary sets of experiments are outlined. These are designed to facilitate an investigation of the performance of locally oriented real-time assignment rules relative to other locally oriented real-time assignment rules; of the performance of four bases cases rules that do not rely on real-time information updates; of the performance of these locally oriented real-time assignment rules

relative to the base cases; and, of the performance of real-time assignment rules relative to an assignment rule in which a full week's loads (for a single vehicle) are known a priori. This last case is comparable to an analysis of assignments with perfect hindsight. The discussion of the simulation and experiments performed in this chapter was primarily descriptive in nature. A detailed description of the procedural details is provided for the interested reader in Appendix I.

CHAPTER 6 ANALYSIS OF ASSIGNMENT STRATEGIES: EXPERIMENTAL RESULTS

INTRODUCTION

In this chapter the results of the experiments described in chapter 5 are examined. This introductory section, in addition to providing a blueprint of the examination of results, is used to highlight some of the more significant findings.

Beginning with a comparison of the first four base case assignment strategies, FCFS, NO, BAT(a), and BAT(b), the effects of imposing a limit on the number of loads that may reside in the pool at one time, and, the effects of delayed vs. immediate assignment of loads to waiting vehicles are examined. Introduced in chapter 5, results discussed in Chapter 6 pertaining to the application of the bipartite assignment technique illustrate the impact of the pool size limit and of varying the length of time between the generation of consecutive assignments. Performance varies significantly across differing assignment periods.

In chapter 6, the real-time operational strategies are examined. The performance of the four assignment strategies, each requiring varying degrees of operational flexibility are investigated with and without deadlines for service (pickup) and, in the case of heavy demand, with and without profit based load acceptance rules. In chapter 6, the relative performance of the three local rules, defined in chapter 5, for assigning loads to vehicles (within the four real-time information strategies) is shown to vary across demand levels, pickup deadlines and system flexibility. None of the three rules dominates absolutely. Results of an extended set of simulation experiments performed over additional levels of demand intensity are examined. The relative strengths and weaknesses of the three local assignment rules are also discussed.

An examination of the effect of allowing en-route diversion and re-assignment of loads on operational efficiency is presented in chapter 6. A section later in the chapter 6 is devoted to the benefits of applying profit based load acceptance rules, particularly in an environment where demand exceeds operating capacity.

Chapter 6 also compares the real-time strategies to the base cases. Without deadlines for service, and under heavy demand, the two closely related base case assignment methods, time and state based bipartite assignment, out-perform the real-time cases with respect to both wait time for service and the average empty distance driven to provide service. In fact, when demands are high and no pickup deadlines are in place, even the nearest origin strategy out-performs the real-time information cases. The variability of wait times, however, is somewhat lower under the real-time assignment strategies. These very high demand scenarios are

intended to provide an upper bound on the efficiency of the bipartite and nearest origin assignment cases.

Even in the absence of pickup deadlines, under moderate and low demand, results of the comparison of the base cases to the real-time cases differ from results generated under (artificially) high demand. Under moderate and low demand levels, the real-time cases, in which multiple loads are assigned to a single vehicle, can be more efficient than the base cases, with respect to the empty distances driven to provide service and hence the profitability of the system. In some cases, wait time for service is higher.

The chief advantage of the real-time cases is that pickup deadlines are met. As designed, the base cases do not explicitly handle pickup deadlines. If loads are assigned such deadlines at least thirty percent and sometimes closer to one-hundred percent of the loads are served long after their deadlines have passed.

Discussed first in chapter 6, the relative performance of the real-time strategies under high demand is significantly improved by the application of a profit based load acceptance rule.

The combined effect of allowing en-route diversion, real-time load re-assignment and profit based load acceptance rules can be shown to improve efficiency, measured by the distances traveled to provide service, while wait times for service and associated variability are kept low by pickup deadline constraints. Under real-time information fleets can be at once profitable and responsive to customers needs for rapid response to service requests.

COMPARISON A - FOUR OF FIVE BASE CASE ASSIGNMENT STRATEGIES

This section is concerned with a comparison of four of the five base case strategies: 1) first called first served assignment, 2) nearest origin assignment, 3) time triggered bipartite assignment and 4) state triggered bipartite assignment. The experiments outlined are intended to test the performance of the base case strategies against each other. The first called first served strategy is inefficient in spatially distributed service system applications but it is included here as a benchmark case. The expected empty and loaded distances under the assumption of uniform origin and destination locations, and the variance of these distances were derived for the FCFS policy in chapter 4. As discussed in the same section, in the single vehicle case, performance measures can be estimated with a high degree of accuracy using fundamental queueing relationships and the first and second moments of the service time. The nearest origin assignment also has some attractive analytic properties, discussed in chaper 4. As will be clear from results presented in the next few sections, nearest origin assignment performs well under very high demand but less well in less congested systems. The performance of bipartite assignment varies widely depending upon the intensity of demand and the specific

implementation selected, particularly with respect to the length of time between consecutive assignments.

FCFS Assignment

As mentioned above, first called first served assignment is included here as a benchmark because it is well-defined, possesses analytically known properties, and its performance is predictable: the average lengths of the empty and loaded moves are equal, the variability of wait time is low relative to other assignment strategies examined, and, the length of time unserved loads have been in the queue at the end of the simulation horizon is approximately half of the average wait time for service. This last result confirms that wait times for service are not longer for loads left in the pool than for those served, an observation that follows directly from the character of FCFS assignment, in which no loads are preferred over any others. This is not the case under the other strategies examined. As may be observed in figure 6.1, the average wait time over all loads accepted into the queue and the average wait time for loads left in the queue at the end of the simulation horizon are consistent across fleet sizes. This assignment method is not practical under most realistic cost models as the cost to provide service exceeds potential revenues.

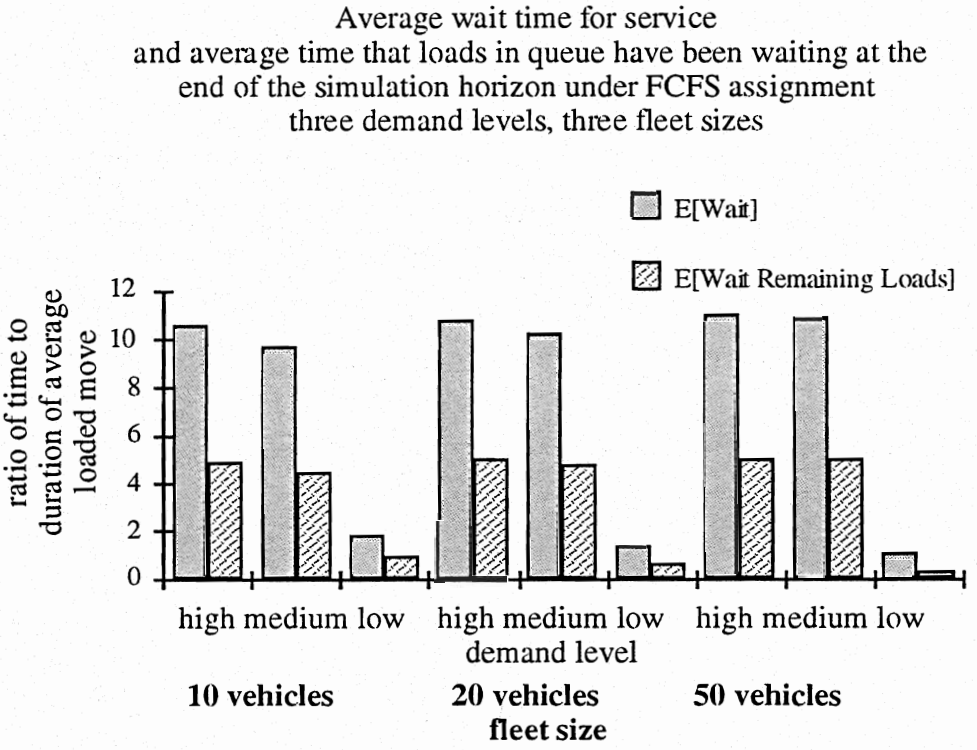


Figure 6.1 Average wait time for service and average time in queue for loads not served at the end of the simulation horizon under FCFS assignment

Nearest Origin Assignment

The nearest origin assignment strategy was introduced in chapter 4 and defined again in chapter 5. The purpose of the discussion in chapter 4 was to compare the performance of a system under nearest origin assignment to heavy traffic approximations for system performance measures derived for M/G/k queueing systems. Nearest origin assignment performs best in high demand environments where the number of choices of loads to assign each vehicle is high; it performs fairly well in low demand environments where the number of idle vehicles available to serve each incoming service requests is high; it fares least well, both in absolute and comparative terms when demands are not so low that many vehicles are available to be assigned loads and not so high that the pool of unassigned demands is fairly large. Unlike the classical assignment strategy in which at pre-defined times, or, alternatively, when the system reaches a certain state, all accumulated loads are candidates for simultaneous assignment to available vehicles, nearest origin assignment assigns loads to the nearest idle vehicle upon arrival to the system when the pool is empty; when the pool is not empty each arriving load joins the pool and becomes a *candidate* for assignment *immediately.* However, when demands are heavy, the variability of wait times can be high. Without safeguards to guarantee that all loads are served within a reasonable length of time, an unattractive load (say on the periphery of the circle in this case) can languish indefinitely. The average length of time that loads not served by the end of the simulation horizon have been waiting can be more than twice the wait time for served loads; the associated standard deviation more than three times that of served loads. Figure 6.2 shows the average wait times for service and wait times in pool for loads not served at the end of the simulation horizon, under different demand intensities and different fleet sizes. It may be observed that although the average wait times under NO are less than those under FCFS the average length of time in pool for loads not served by the end of the simulation horizon may be higher, under heavy demands, than under FCFS assignment. In addition, while the average wait time for service decreases with an increase in fleet size, the wait times for loads remaining in the pool at the end of the simulation horizon do not decrease.

Performance of Classical (Bipartite) Assignment: Tradeoffs Between Immediate and Delayed Assignment of Loads to Vehicles

Central to this research is the issue of immediate versus delayed assignment of loads to vehicles. When service requests arrive to the system over time it may be advantageous to avoid assigning loads until more information is available; on the other hand, throughput may be increased by serving requests as soon as possible, thereby freeing drivers to accept new assignments. In this section, the third and fourth base cases are examined in order to explore

these tradeoffs in detail. Mentioned in chapter 5, these experiments do not fit neatly into the experimental design constructed for comparisons of different operating strategies. The examination here is intended to bring to the forefront issues that are of importance to the overall analysis and to clarify the choices of parameter values (a and b) used in the simulation experiments to follow.

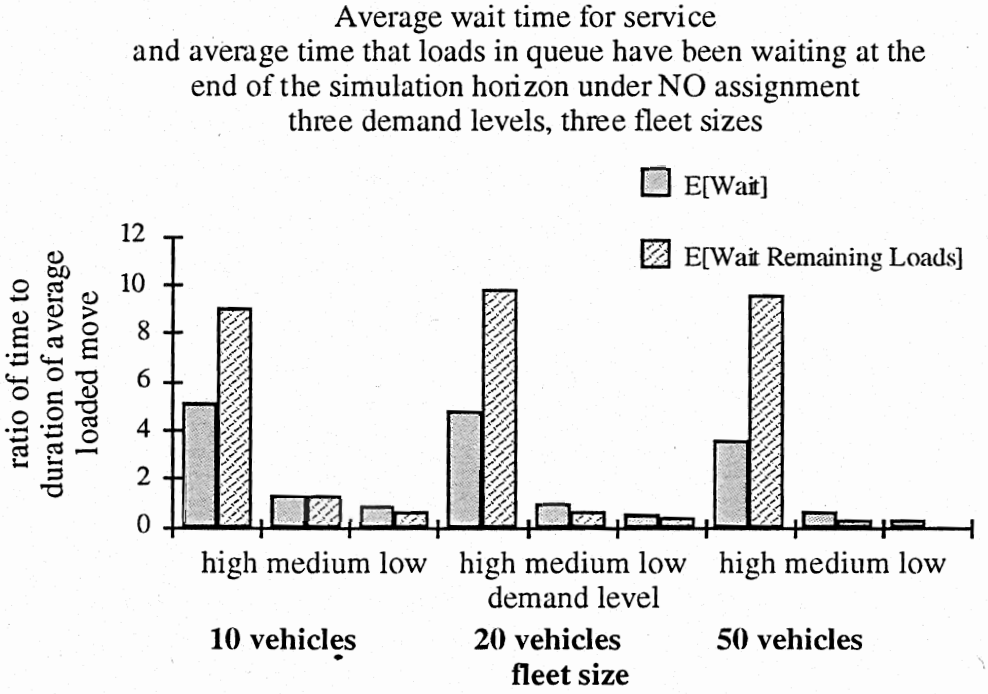


Figure 6.2 Average wait time for service and average time in pool for loads not served at the end of the simulation horizon under NO assignment

The cases examined are chosen to highlight the tradeoffs between immediate and delayed assignment of vehicles to loads. BAT(a) with a close to zero approximates nearest origin assignment. In addition, a system in which "look ahead" is permitted -- that is, vehicles that will become idle within a fraction of the time between the current and the next scheduled assignment are included in the current assignment -- allows for the examination of the tradeoffs associated with including more vehicles in the current assignment. The latter would on one hand lead to a superior solution in terms of greater efficiency (due to more loads and more vehicles and hence lower cost assignments) while, on the other hand excluding future opportunities. BAT(a) with look ahead could properly be termed a quasi real-time assignment strategy since accurate information about the current and near term status and location of vehicles is needed. BAS(b) represents another quasi-real-time implementation of the bipartite assignment heuristic in which assignments

are made when the system reaches a pre-specified state. This adaptive assignment method offers advantages under moderate and low demands and can reduce the variability of wait times for service.

Heavy Demand and the Restriction on Pool Size. As described in chapter 5, the maximum allowable number of loads in the system is restricted to a fixed number. The limit chosen for the experiments conducted is five times the number of vehicles in the fleet. One reason for this particular limit is that the resulting maximum number of loads in the system corresponds to the limit in the real-time strategies examined. In addition, and more importantly, operational issues favor a limit on the number of loads in the pool. The limit is non-binding under moderate and low demand but causes rejection of requests under heavy demand. When less restrictive limits are imposed the average distance traveled empty goes down, but the wait time for service, the variability of that wait time and the length of time loads not served at the end of the service horizon have been waiting is very high. Figure 6.3 shows the first and second moments of wait time and the average length of time remaining loads had been in the pool at the end of the simulation horizon for pool limits of 5, 10 and 15 times the number of vehicles. Results for both time based and state based bipartite assignment are shown. Parameters a and b are assigned representative values chosen for simulation experiments under heavy demand. Numbers are shown relative to the length of the average loaded move. The 18- 30% increase in profitability that results from a reduction in the average empty distance does not justify the increase in wait times experienced.

Throughput Maximization: a $\rightarrow$ 0 and Nearest Origin Assignment. In a high demand environment, and in a compact service area, the goal of maximizing throughput dominates. When loads to not have associated pickup deadlines and demands for service exceed the ability of the fleet to provide service, the pool of loads waiting to be served will contain many candidate loads. In the system examined in this study, it appears that the purely "greedy" (because it looks for the best next load for one vehicle at a time and takes only the distance to the origin location of loads into account) nearest origin assignment, in which vehicles are assigned loads as soon as they complete service, provides the best overall utilization. Empty distances traveled can be reduced slightly by allowing loads to accumulate, but overall throughput decreases. Chapt er 6 presents simulation results that examine the system with a modification in which vehicles that will become idle a fixed fraction of the between the current and next assignment period are included as candidates for assignment. This look ahead can be shown to significantly improve

throughput under high demand. In this section we do not consider look ahead policies and merely examine the performance of the system when the parameter a takes on various values.

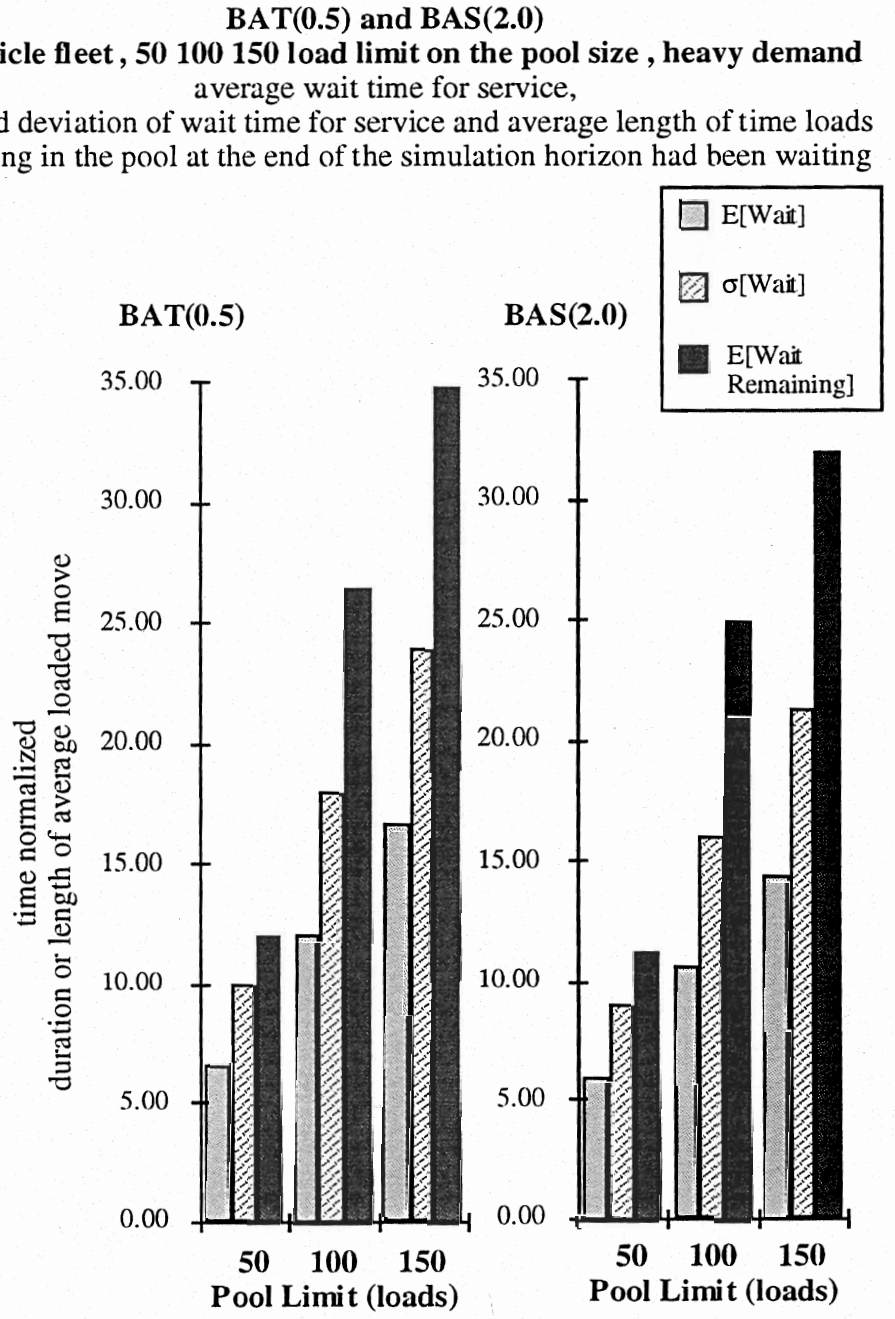


Figure 6.3 Wait time and variability of wait time - pool limits of 5, 10 and 15 times the fleet size

Figure 6.4 shows how empty distance decreases as the time between assignments increases. Of course, there is an associated increase in wait time, shown in figure 6.5. The performance of the nearest origin assignment is included in all figures as a benchmark for comparison. Under the conditions described here, larger values of a (longer times between assignments) means that more loads are turned away and that on average vehicles spend more time idle. Figure 6.6 shows how the operating profit generated decreases as a increases. Figures 6.7 and 6.8 illustrate the corresponding increase in idle time and decrease in the fraction of service requests accepted as the time between assignments increases.

More Choices: More Efficient Solutions. In contrast to the very high demand scenario, when the demand level allows all requests to be served, the goal of serving efficiently dominates. This may require waiting to assign loads until an efficient assignment is possible. In some cases, throughput may actually be increased by waiting until a sufficient number of loads are available for assignment (an increase in idle time may be traded off for reduced empty travel time). If demand for service is very low, more of the fleet is available to provide service to requested loads so the likelihood that an immediate efficient assignment can be found is higher than in the moderate demand case. Figure 6.9 shows the reduction in average empty distance traveled as the number of loads included in an assignment increases. The curves shown in the diagram provide a lower bound on the distances traveled in the system examined. The results are generated (through simulation) in the following way: 1) before an assignment, exactly PL loads are generated, uniformly and independently over a circle, where PL is the limit on the pool size. 2) An assignment is generated, matching V vehicles with the PL loads. The loads chosen in the best solution are served, the rest are discarded. 3) When the V loads have all been served, each by a different vehicle, steps 1) - 3) are repeated. Vehicles remain at the destination point of loads served between assignments periods; they are assigned a new load in every assignment period. Empty distances traveled are presented as the ratio of the average empty distance traveled to the average distance traveled loaded.

Moderate Demand: Conflicting Criteria. In a moderate demand environment conflicting criteria should be taken into account. Satisfying as many service requests as possible is important, as is providing service within a reasonable amount of time. As mentioned in the previous section, allowing demands to accumulate for short periods can lead to more efficient driver to load assignments but results in an increase in wait time for service and in some cases, a decrease in the ability of the fleet to provide service to future requests - resulting in a reduction of operating revenues and profits. Figure 6.10 shows the reduction in the length of the average

empty move as the time between consecutive assignments increases, while figure 6.11 illustrates the corresponding increase in wait time for service that accompanies this reduction. Results for the nearest origin assignment are included here too. It may be observed in figure 6.12, that operating profit increases as the time between assignments increases, up to a point. The decrease that is observed for a values of 1.5 and 2.0 is a result of turning loads away -- a decrease in throughput. The fraction of loads accepted is shown in Figure 6.13.

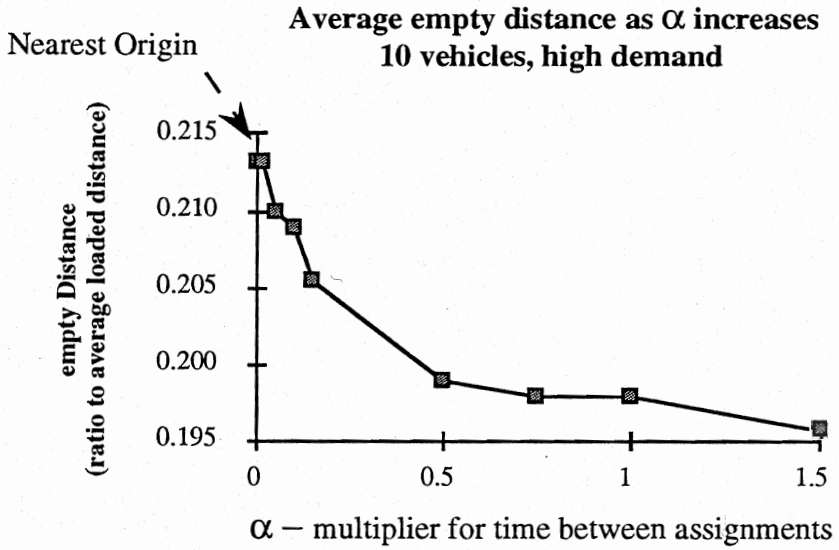


Figure 6.4 Empty distance as the time between assignments increases

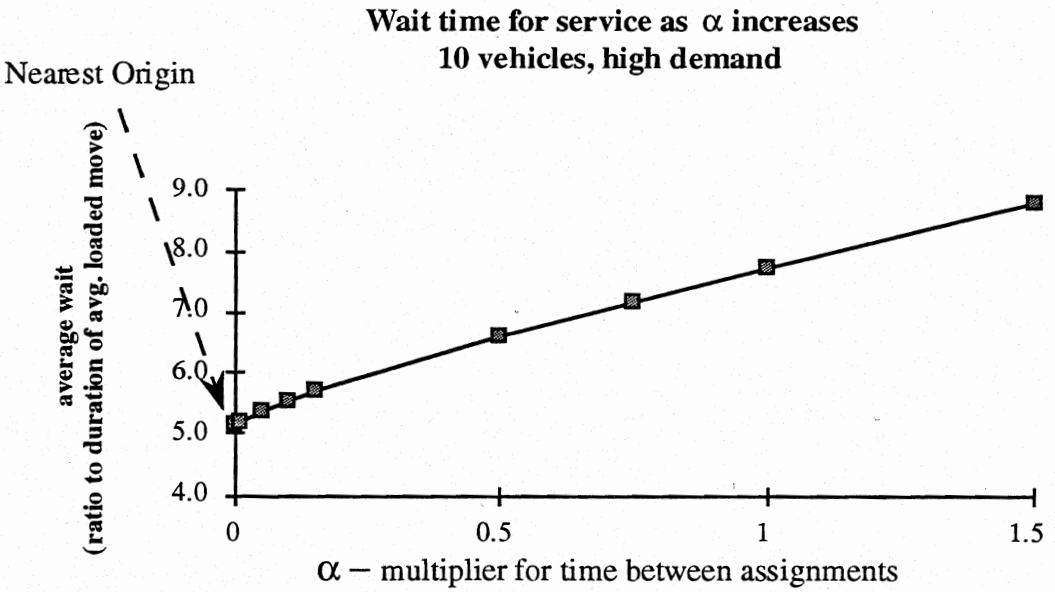


Figure 6.5 Wait time for service as time between assignments increases

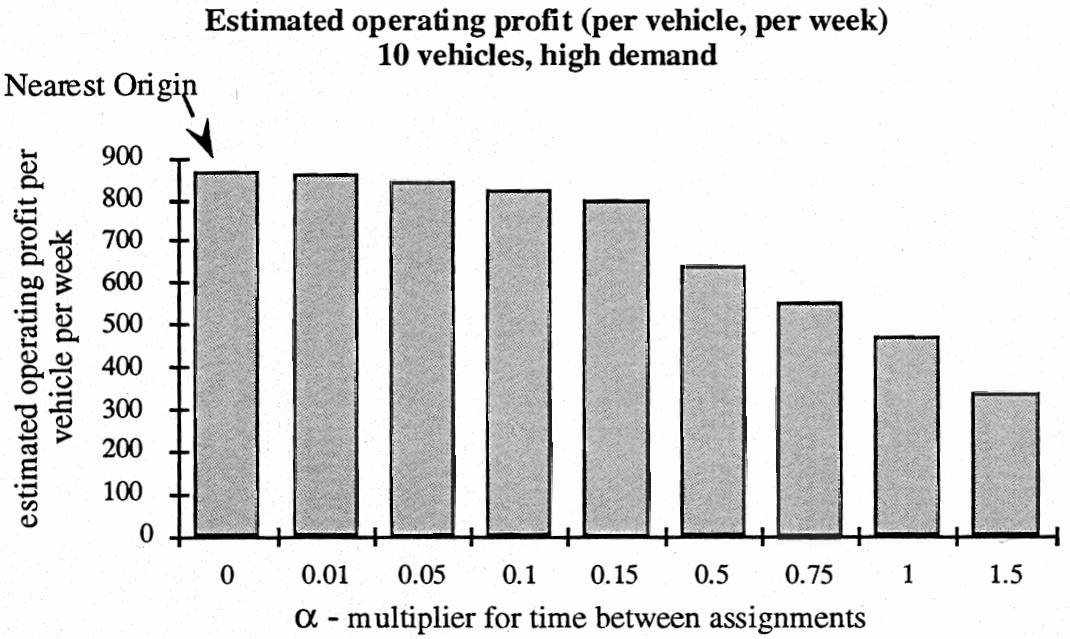


Figure 6.6 Operating profit as time between assignments increases

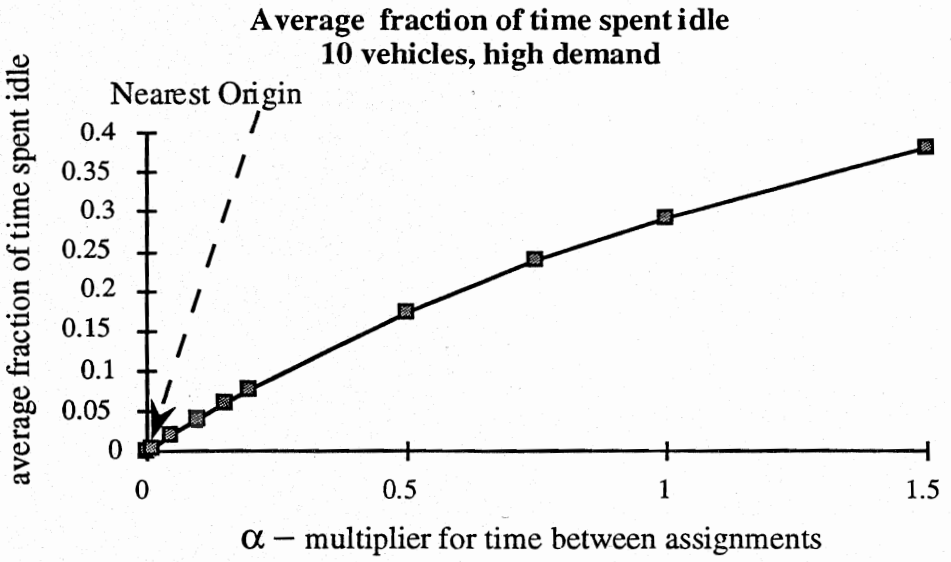


Figure 6.7 Average fraction of time spent idle as time between assignment increases

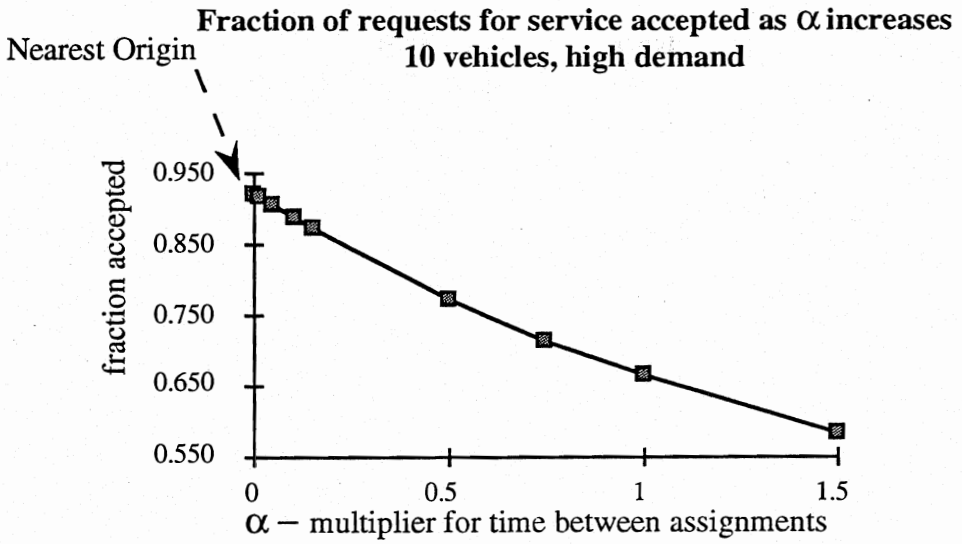


Figure 6.8 Fraction of requests accepted as time between assignments increases

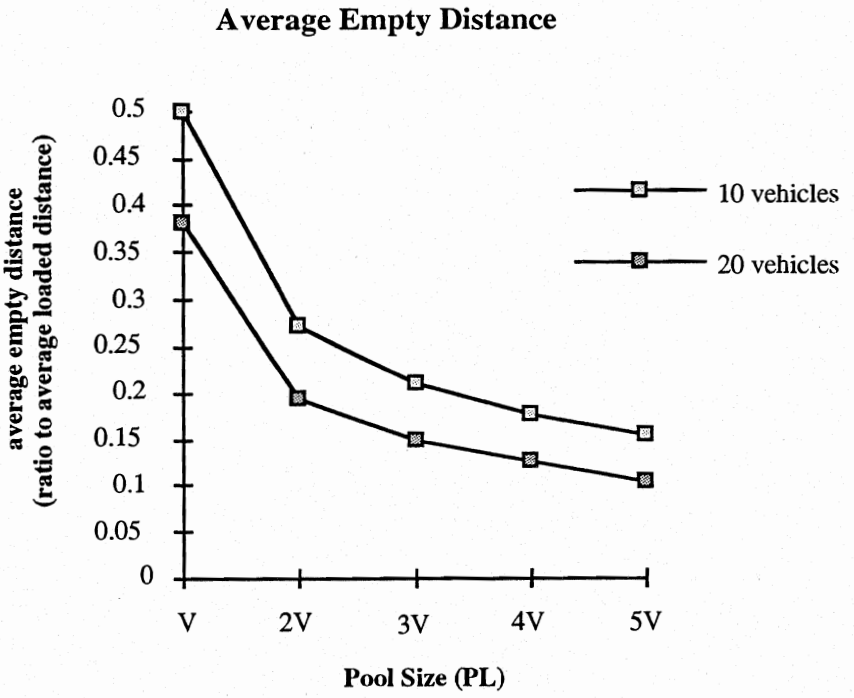


Figure 6.9 Average empty distance traveled when nV randomly generated loads are candidates for assignment to a fleet of V vehicles.

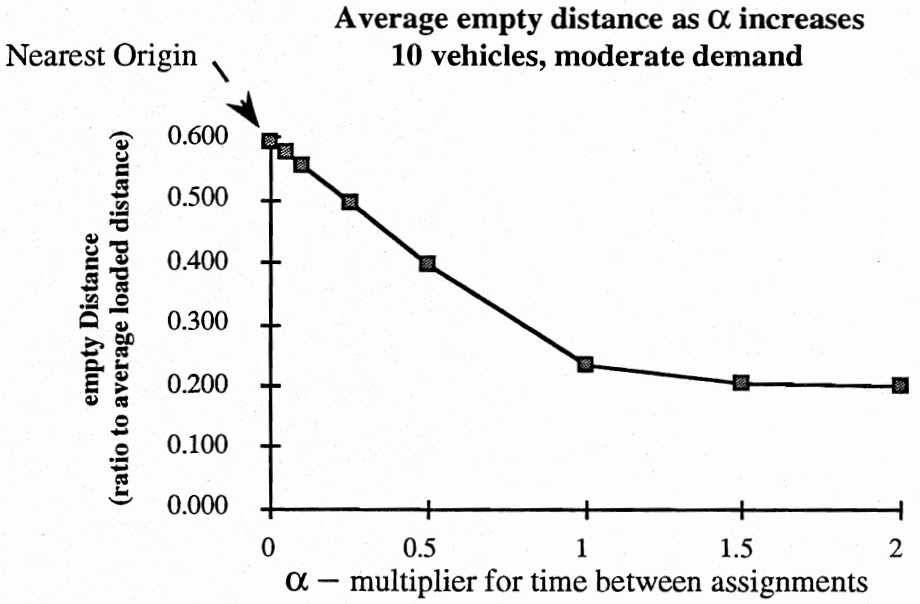


Figure 6.10 Empty distance as the time between assignments increases

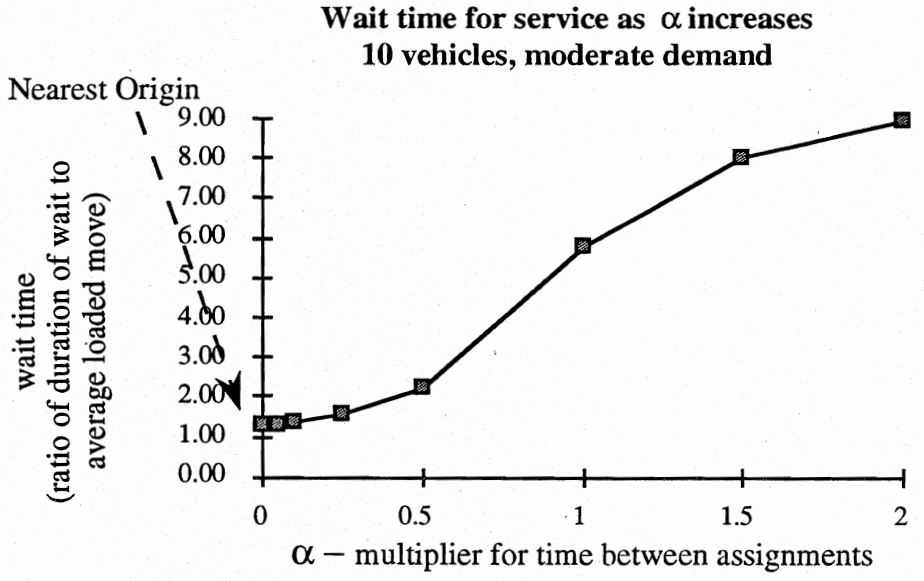


Figure 6.11 Wait time for service as time between assignments increases

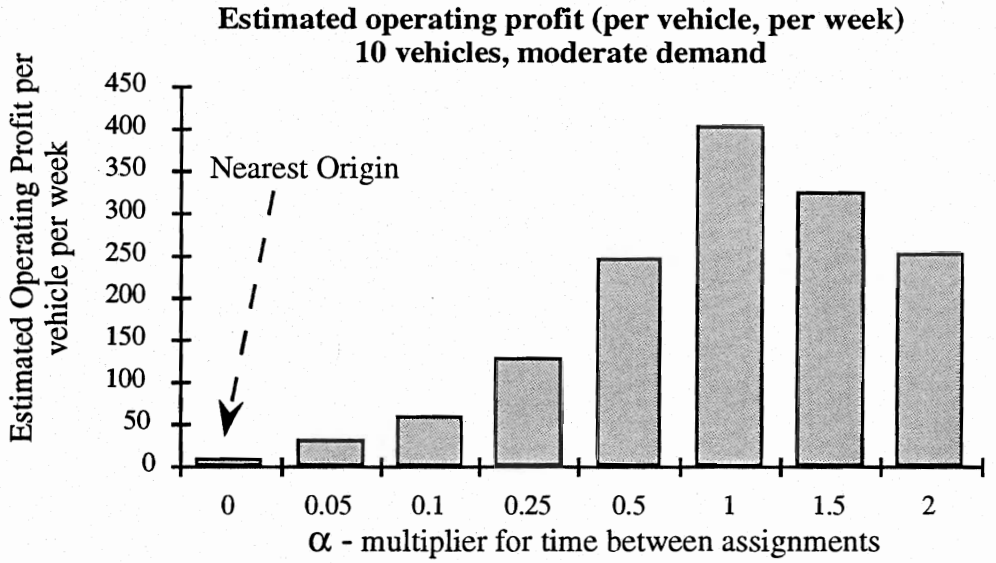


Figure 6.12 Operating profit as time between assignments increases

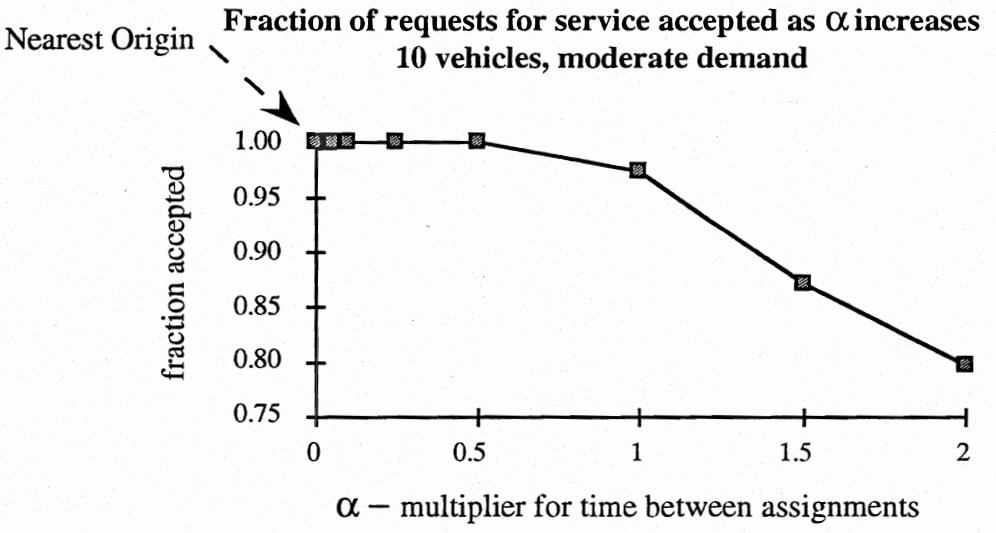


Figure 6.13 Fraction of requests accepted as time between assignments increase

"Look ahead" Policies. In a moderate demand environment, including vehicles that will become available within a fraction of the time between the current and the next scheduled assignment time can lead to a reduction in the average wait time for service. This reduction comes at a price however, as future opportunities for cost effective assignments may be missed. Results of numerical experiments involving "look ahead" policies suggest that a look ahead period near one half of the assignment period yields the most attractive combination of a reduction in the mean and standard deviation of wait time and a relatively small increase in the average distance traveled to provide service. Figure 6.14 offers a comparison of the average empty distance traveled to provide service and average wait time under three policies. In the first, only idle vehicles are considered for assignment; in the second, vehicles that are to become idle halfway between assignment periods are also considered; while in the third, vehicles that will become idle any time before the next assignment period are considered for assignment. Figures 6.15 and 6.16 compare the half look ahead policy to the no look ahead policy over a wide range of values of a. While the half look ahead policy is dominated, in this moderate demand scenario, with respect to operating profit generated by the no look ahead policy, it performs nearly as well for large values of a. For a = 1.5, for example, the operating profit generated is quite close to the no look ahead case and the average wait time is less than half the wait time in the no look ahead case.

Under heavy demand, employing a half look ahead policy appears to improve performance with respect to all criteria of interest except the average distance traveled empty, which increases. The reason for its success is that when demand is heavy, assigning more loads in the current assignment does not significantly reduce future opportunities. By the time the next assignment period arrives most loads assigned in the last period will have been replaced by new arrivals. Under heavy demand, vehicles spend less time idle under the look ahead policy, resulting in higher operating profits. The increased throughput achieved more than makes up for the additional time spent moving empty and results in a decrease in the average time spent waiting for service. Figures 6.17 and 6.18 compare the profitability and average wait time under the half-look ahead policy when demand is heavy to one in which only idle vehicles are candidates for assignment. The time between assignments in each case is half the duration of the average empty move (a = 0.5). The improvements in performance of the system examined under high demand and the half look ahead policy are consistent across fleet sizes. Figure 6.19 shows the improvement in profitability, reduction in wait time and increase in empty distance traveled to provide service that results when the half look ahead policy is compared to the no look ahead policy for fleets of 10, 20 and 50 vehicles.

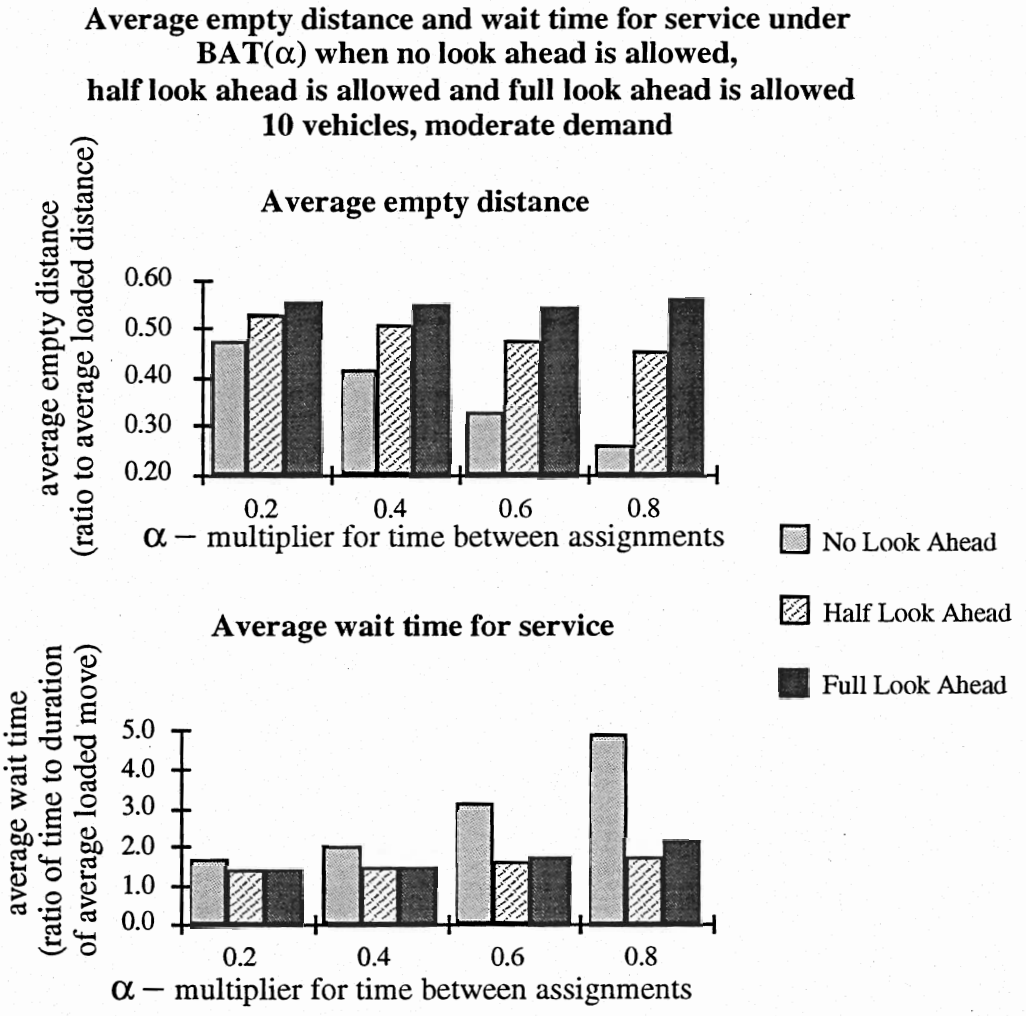


Figure 6.14 Comparison of performance of BAT(a) relative to empty distances traveled and wait time for service under no, half and full look ahead policy

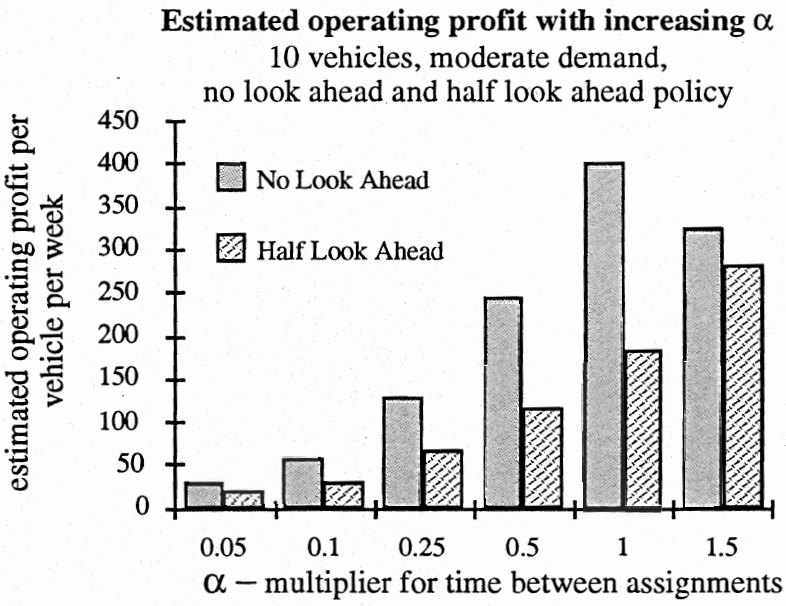


Figure 6.15 Operating profit - look ahead policy vs. no look ahead policy as the time between assignments increases - moderate demand

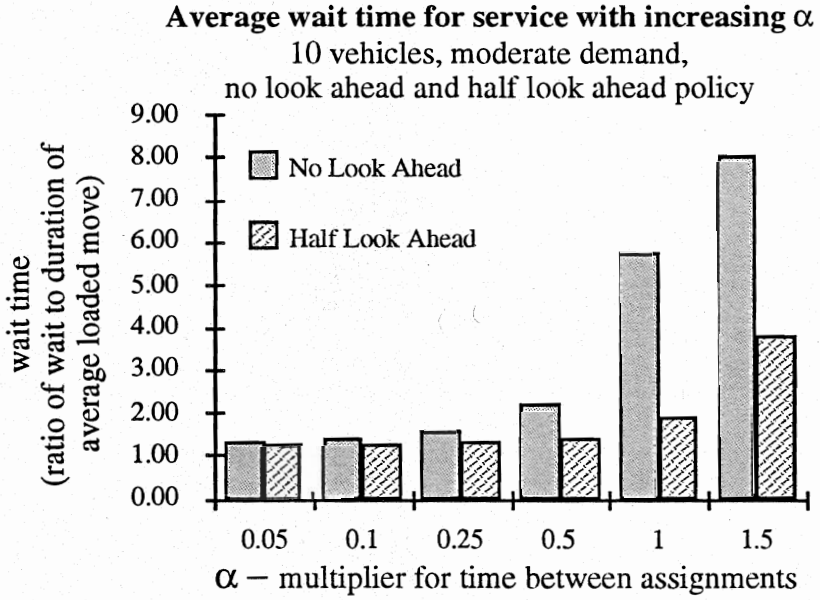


Figure 6.16 Wait time for service - half look ahead policy vs. no look ahead policy as the time between assignments increases - moderate demand

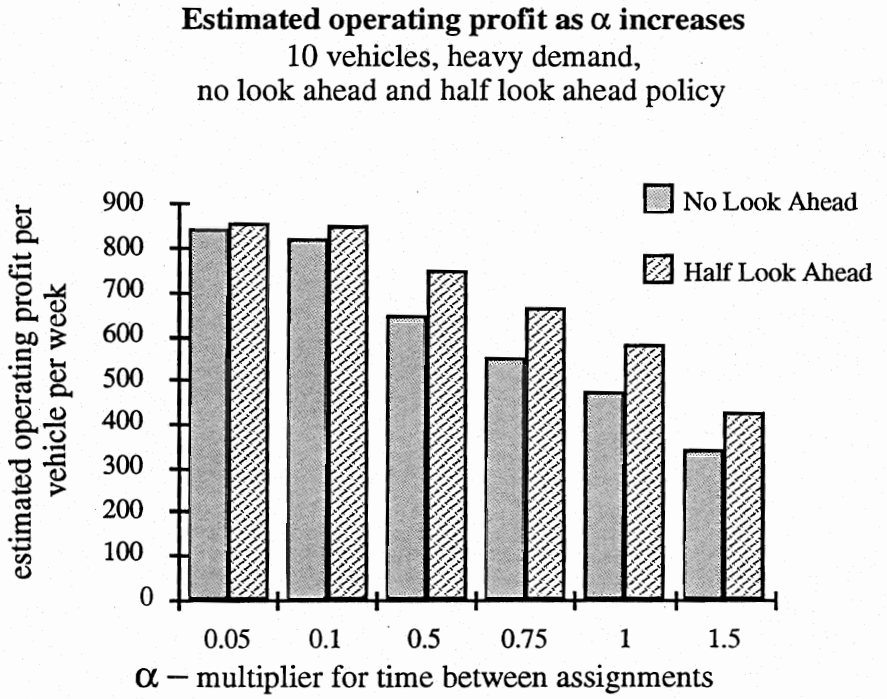


Figure 6.17 Operating profit - look ahead policy vs. no look ahead policy as the time between assignments increases - heavy demand

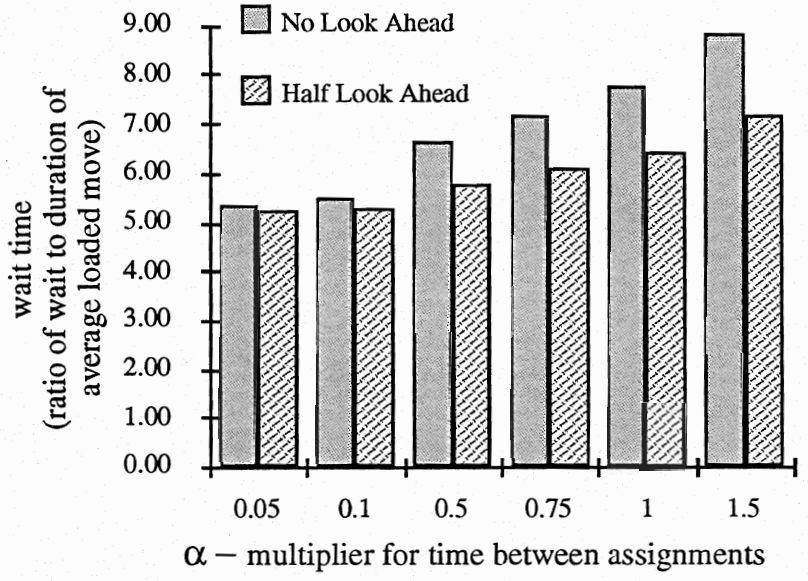


Figure 6.18 Wait time for service - half look ahead policy vs. no look ahead policy as the time between assignments increases - heavy demand

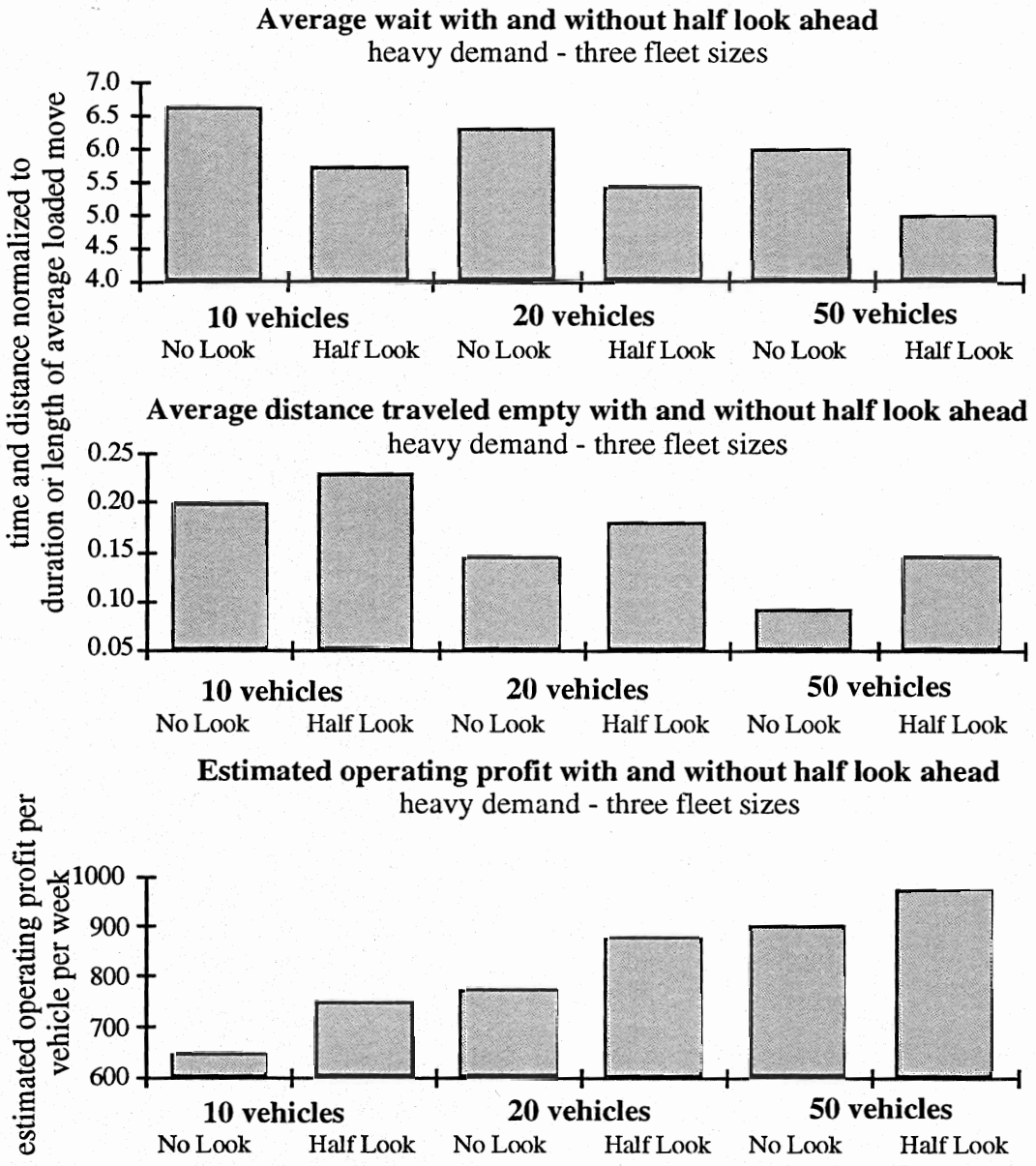


Figure 6.19 Three performance measures - half look ahead policy vs. no look ahead policy under BAT(0.5), three fleet sizes, heavy demand

Assignments Triggered by Excess Idle Vehicles or Waiting Loads . An examination of state-based bipartite assignment (BAS(b)), outlined in Chapter 5, is presented here. Assignments are triggered when the number of loads awaiting service is equal to a multiplier b times the number of idle vehicles. Assignments are also triggered when the number of idle vehicles is equal to b times the number of waiting loads. The combination of two triggers is designed to limit the time loads wait for service, both when demand is moderate or heavy (the former trigger is active) and when it is low (the latter trigger is active). The choice of a single trigger, rather than two different parameters, is admittedly arbitrary. It may well be that sensitivity analysis would reveal a choice of two parameters that would perform better than the single one. We leave that a subject of future research.

The simulation results discussed focus primarily on scenarios where demand is moderate. Because of the limitation placed on the number of loads allowed in the pool, the pool is always nearly full under high demand. As a result, the ratio of vehicles to waiting loads is nearly always the same; under high demand, the steady state performance of BAS(b) does not vary with b. Figure 6.20 illustrates the uniformity of performance of BAS(b), relative to four performance measures, for three values of b.

In the moderate demand environment examined, simulation results echo those in which fixed assignment periods of varying length are examined - the average distance driven to provide service is lower when the assignment is triggered less often (higher b), the average wait time for service, higher. Figures 6.21 and 6.22 illustrate this relationship. The stable region for this assignment rule is limited to cases where b is less than or equal to the multiplier which specifies the maximum number of loads in the pool. In this investigation, that number is five, that is, the maximum number of loads allowed in the pool is five times the number of vehicles. The reason for this limitation is that when the system begins all vehicles are idle; the first assignment is made when the number of loads in the pool is equal to b times the number of vehicles. If b is greater than the multiplier that specifies the pool size, a situation quickly arises in which the moderate and heavy demand trigger (PL greater than or equal to bV) is never initiated. The operating profit, shown in figure 6.23, does not drop off after a point, as it does when the time between assignments is increased in the moderate demand case. Because assignments are triggered by an accumulation of loads, in the stable region examined, loads are not turned away because of lack of capacity.

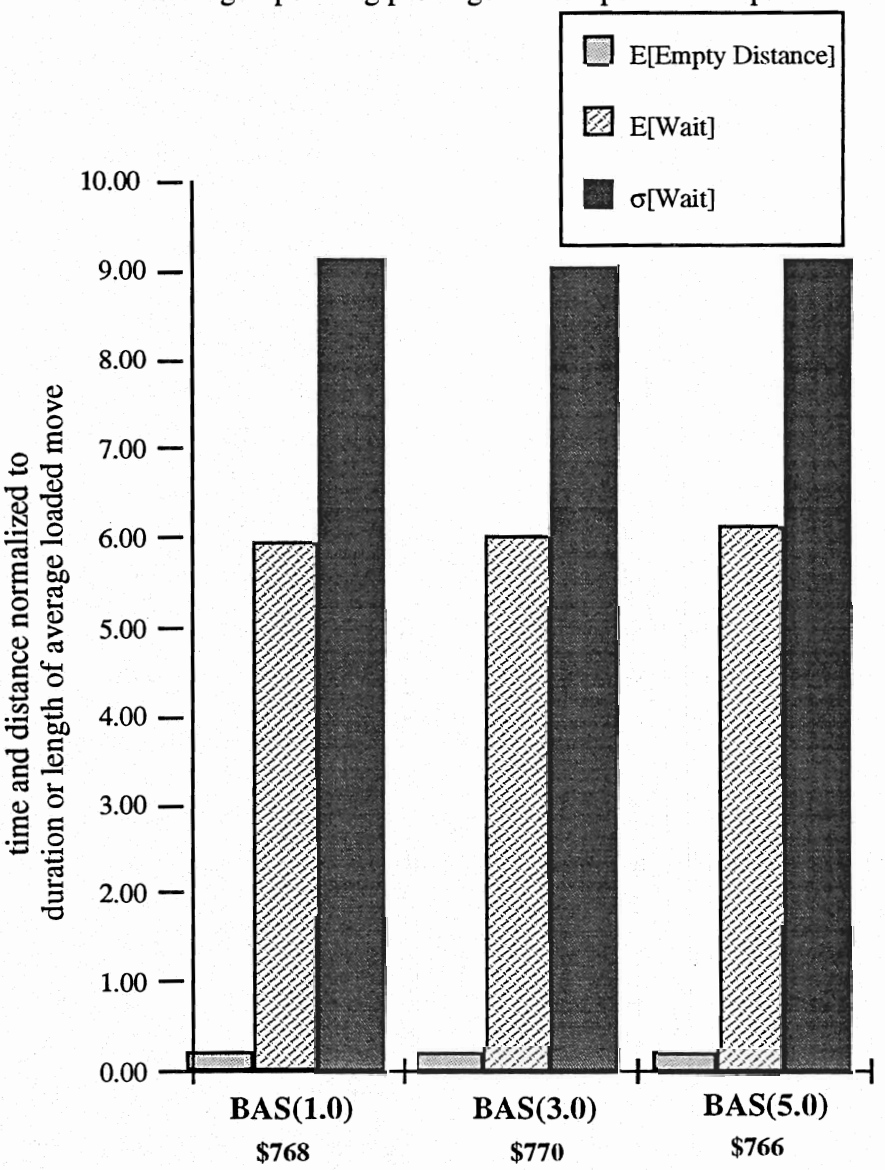


Figure 6.20 Performance of BAS(b) under high demand relative to four measures

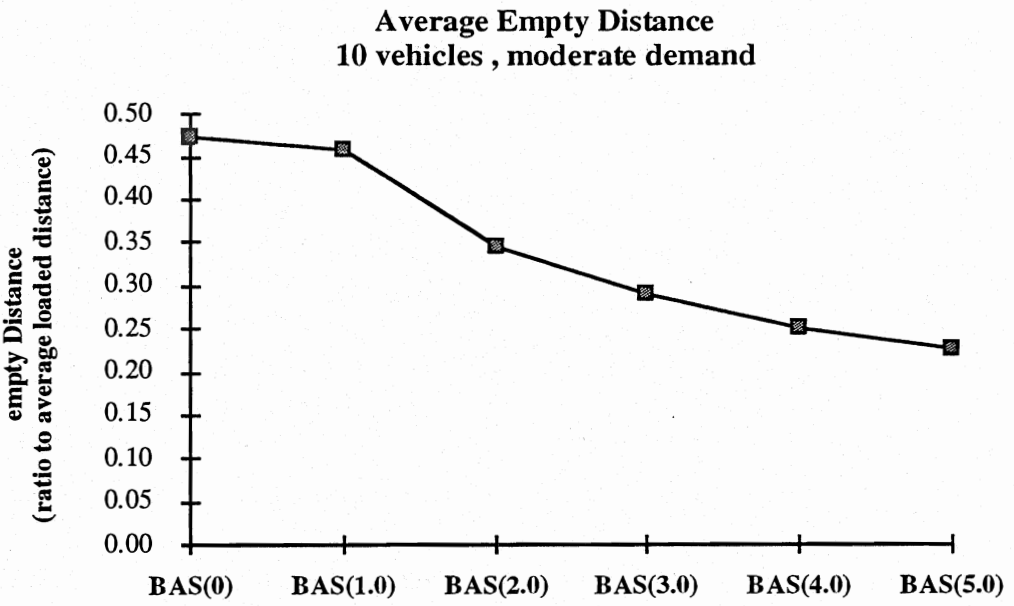


Figure 6.21 Average empty distance as accumulation of loads increases

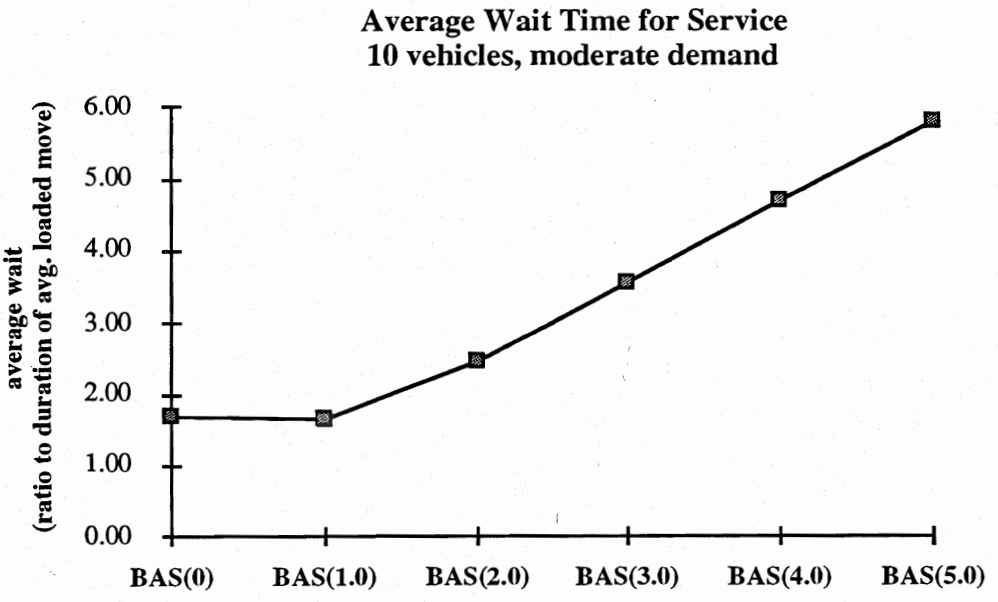


Figure 6.22 Average wait for service as accumulation of loads increases

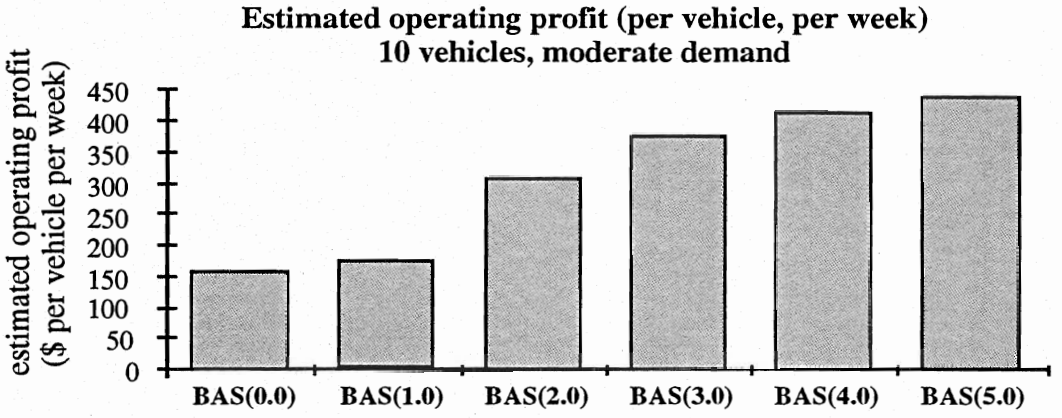


Figure 6.23 Operating profit under BAS(b) as accumulation of loads increases

Summary of Bipartite Assignment Performance and Selection of Cases for Further Comparison. The discussion and simulation results presented in this section point out the extent to which the performance of the system under these closely related assignment strategies varies. Figures 6.24-6.27 present results from simulations of both time-based and state-based assignment periods and show that neither rule strictly dominates the other. Assignment based on the state of the system possesses clear advantages in systems in which demands fluctuate in an unpredictable manner, while time-based assignment which requires less unplanned driver to dispatcher communication may perform quite well, when an appropriate assignment period is chosen and under a stable request arrival pattern. As examined in chapter 6, a look ahead policy can improve the efficiency of time based assignment under high demand.

The cases chosen for comparison to the real-time information cases are BAT(a) and BAS(b), with b = 2.0 in all cases and a = 0.50, 0.75 and 1.25 under high moderate and low demand.

Summary of Base Cases Comparisons

Figures 6.28 to 6.36 present a comparison of the performance of FCFS, NO and bipartite assignment strategies with respect to four criteria, average empty distance, average wait time for service, standard deviation of wait time and operating profit generated. In may be observed that under heavy demand the nearest origin strategy performs best with respect to all four criteria, but that its relative advantage over BAS(b) decreases with larger fleet sizes. In the limit, as a approaches zero, BAT(a) approximates NO. Similarly, under heavy congestion BAS(b), for any value of b approaches NO. The value of 0.50 is chosen for a in the high demand scenario because the corresponding frequency of assignments, generated every 2.26 simulation hours, is a more representative value than a close to zero. A half look ahead policy is beneficial in all of

the high demand scenarios examined, but does not achieve the efficiency (and high throughput) of NO and BAS(b).

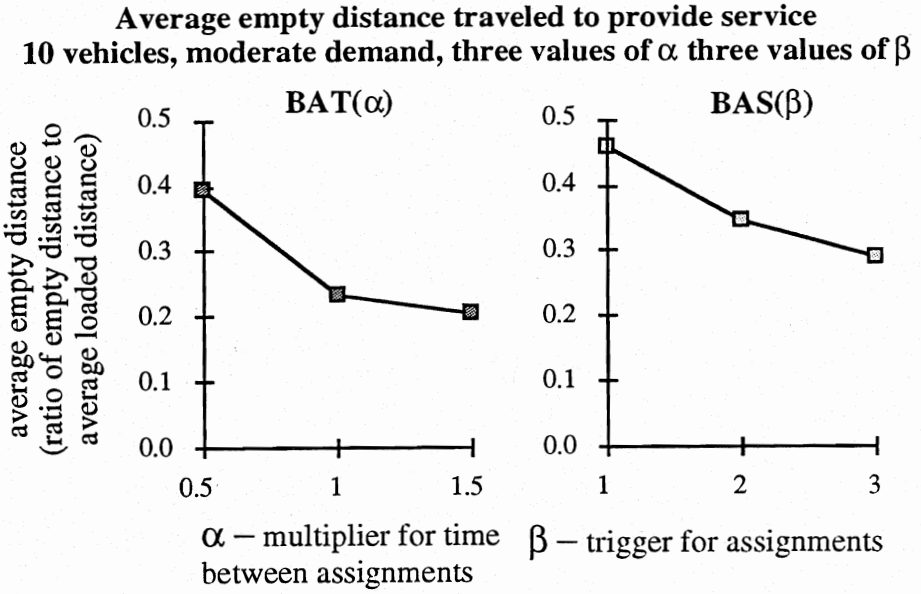


Figure 6.24 Average empty distance traveled - BAT(a) and BAS(b)

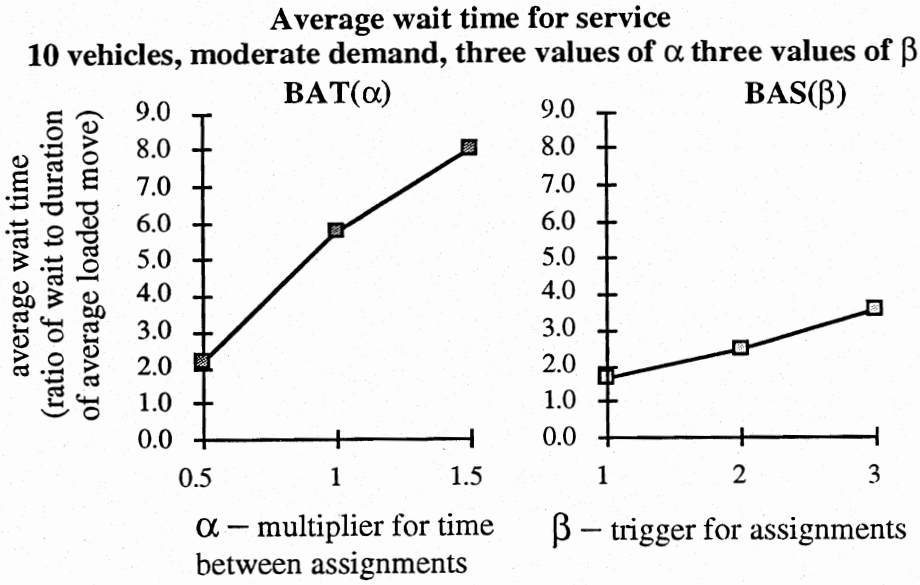


Figure 6.25 Average wait time for service - BAT(a) and BAS(b)

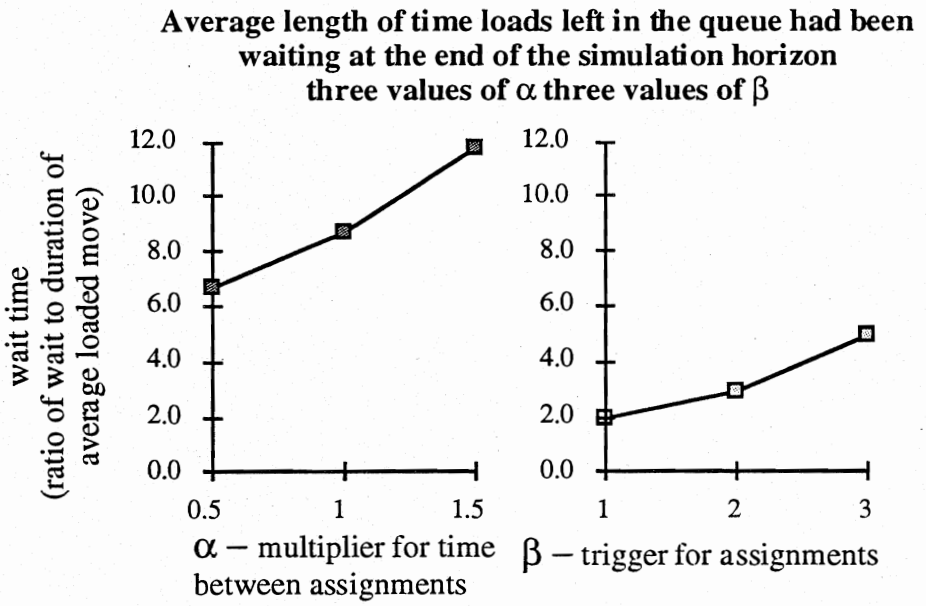


Figure 6.26 Average length of time loads not served had been in pool at the end of the simulation horizon - BAT(a) and BAS(b)

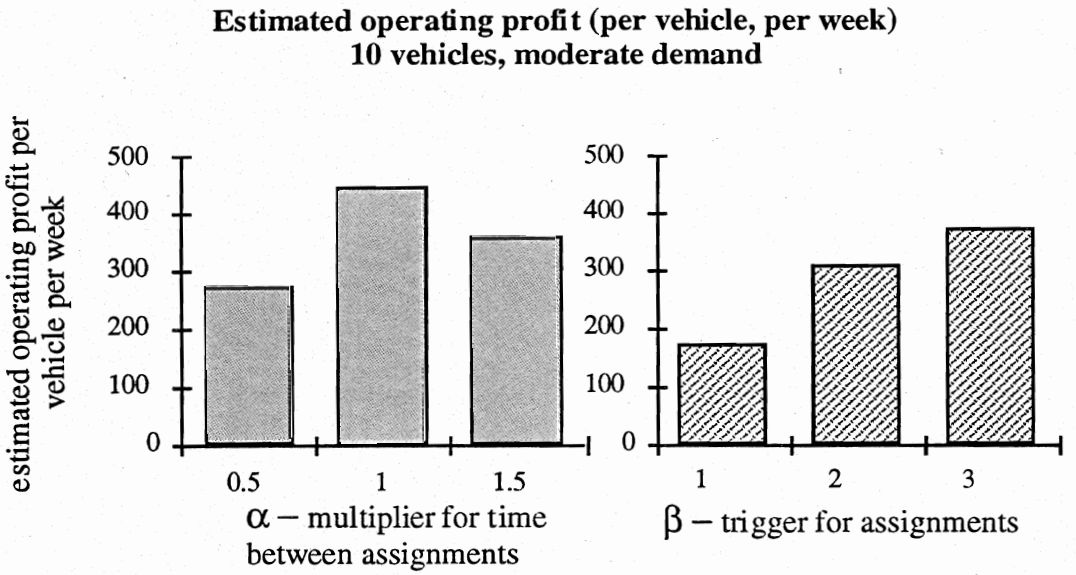


Figure 6.27 Operating profit - BAT(a) and BAS(b)

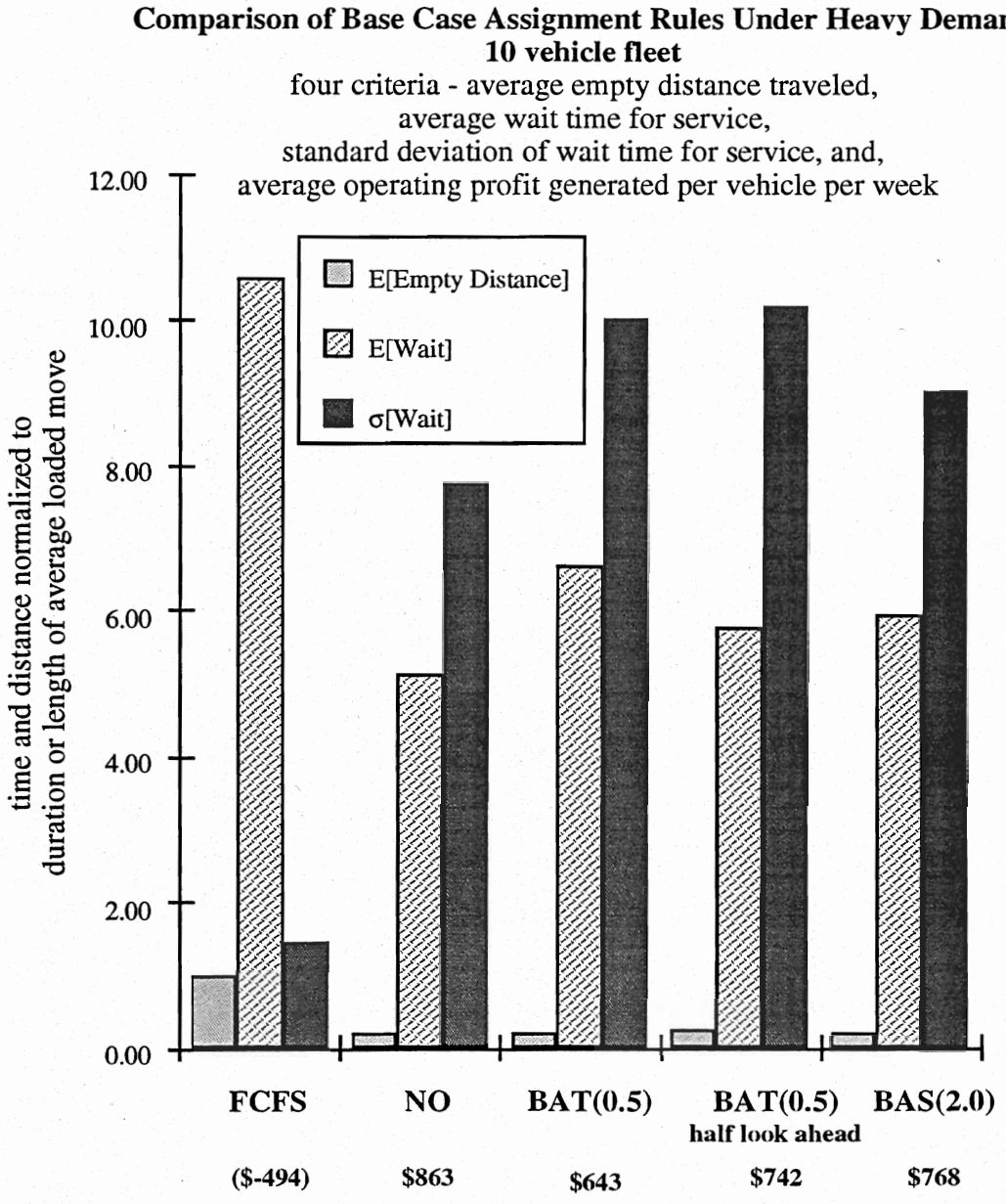


Figure 6.28 Comparison of base cases - heavy demand - 10 vehicles

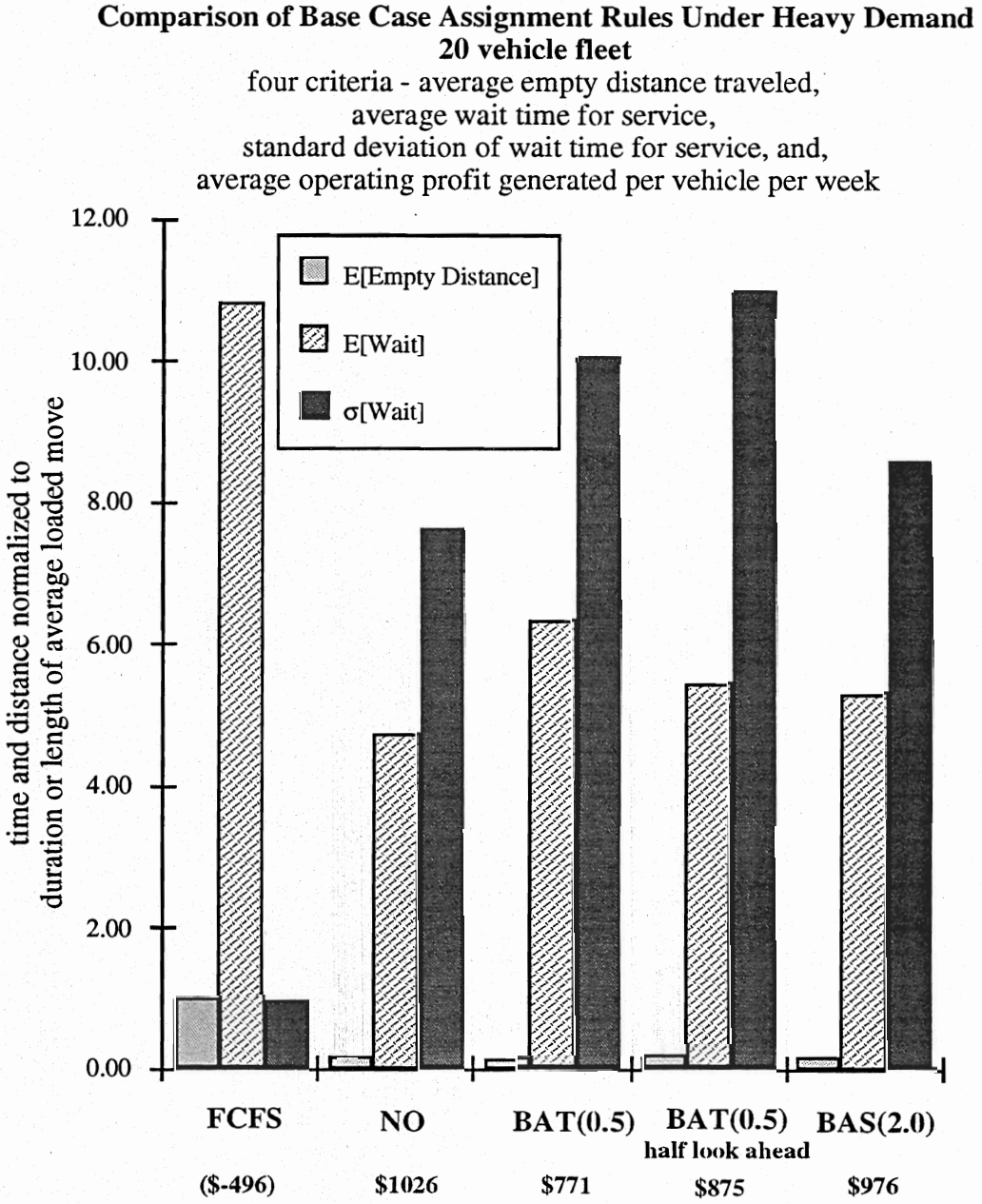


Figure 6.29 Comparison of base cases - heavy demand - 20 vehicles

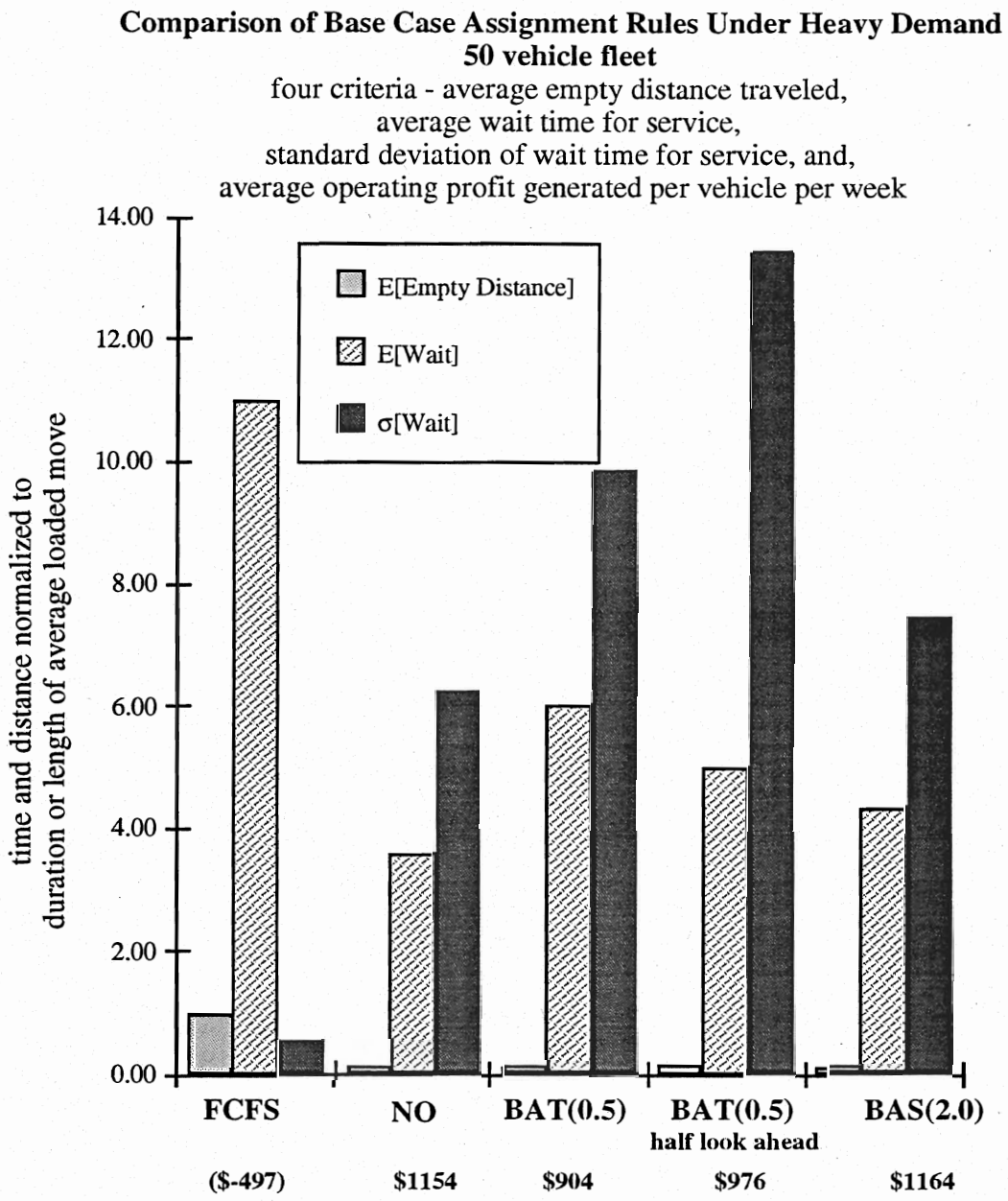


Figure 6.30 Comparison of base Cases - heavy demand - 50 vehicles

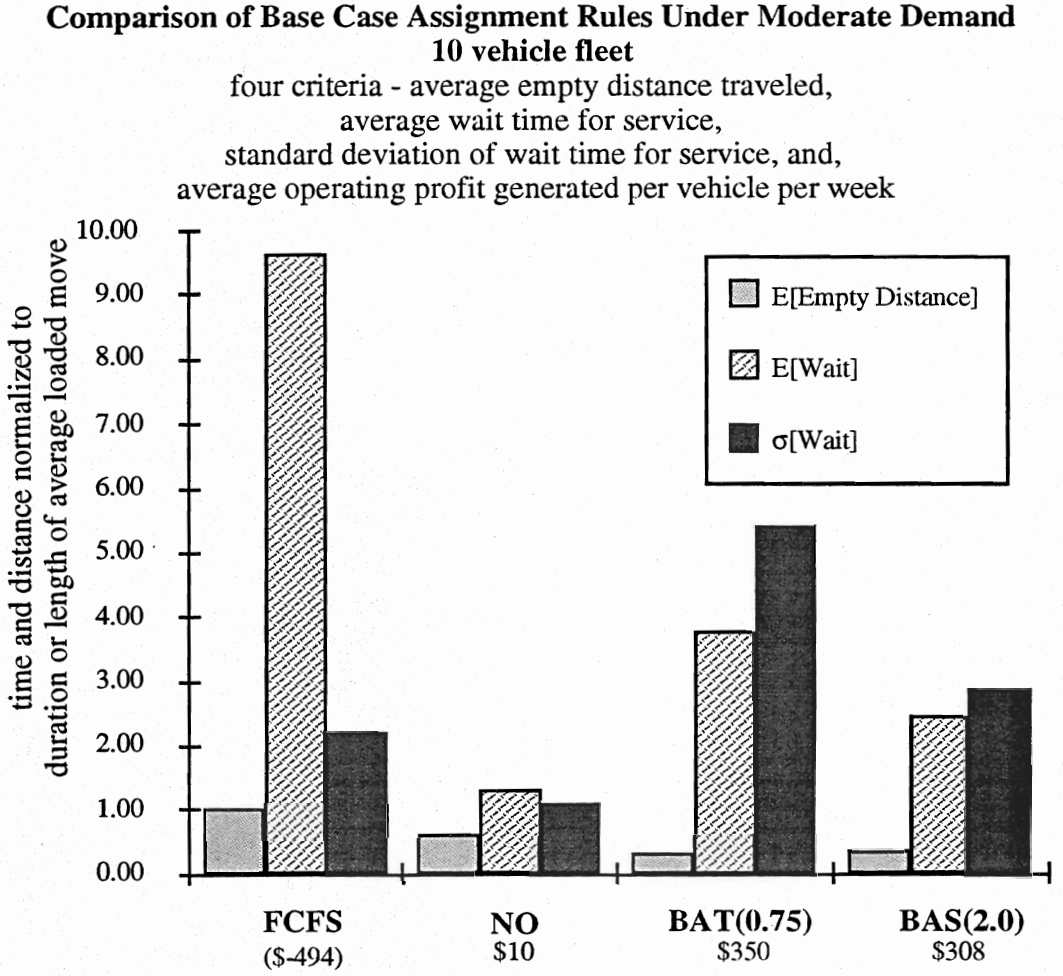


Figure 6.31 Comparison of base cases - moderate demand - 10 vehicles

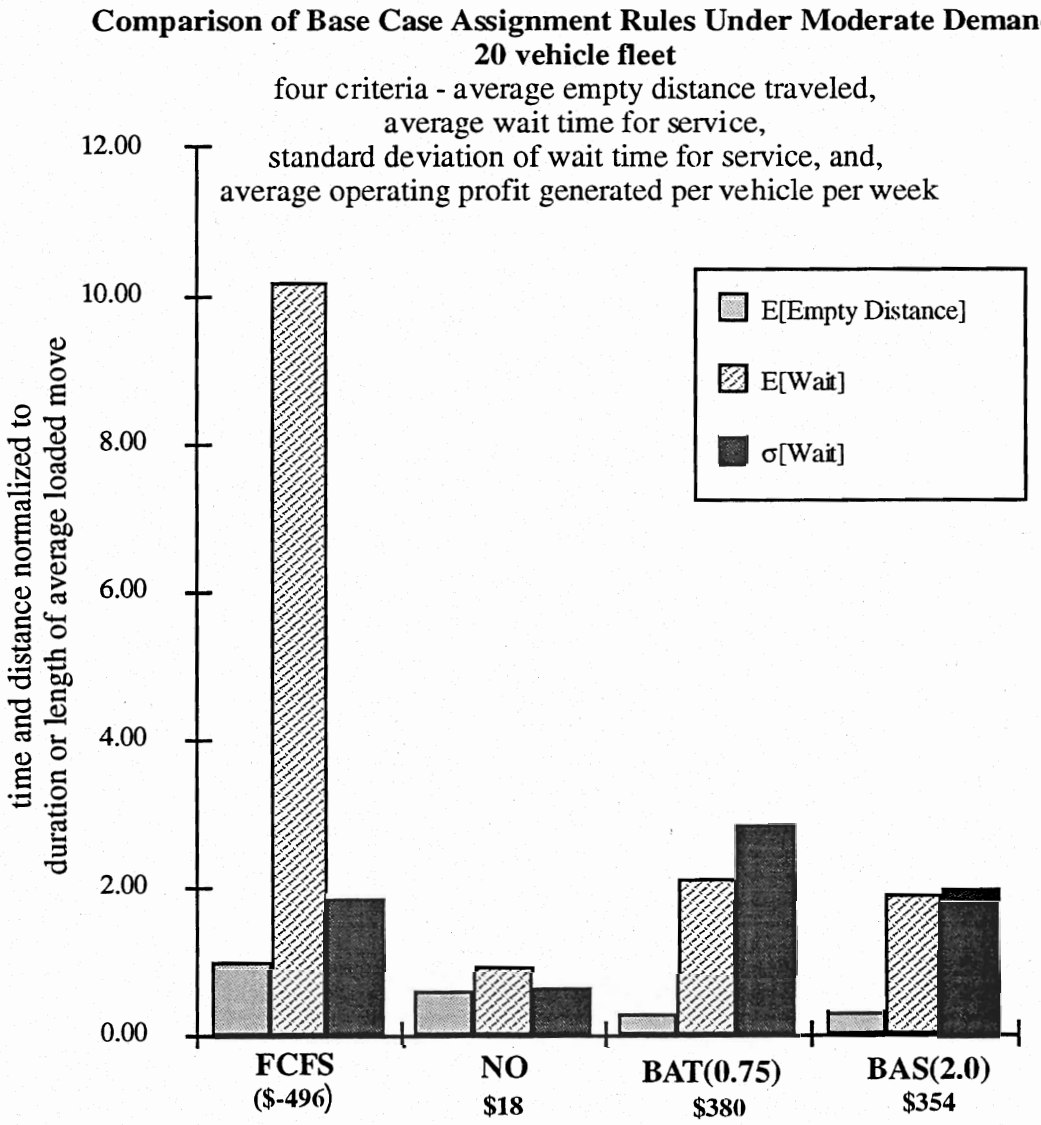


Figure 6.32 Comparison of base cases - moderate demand - 20 vehicles

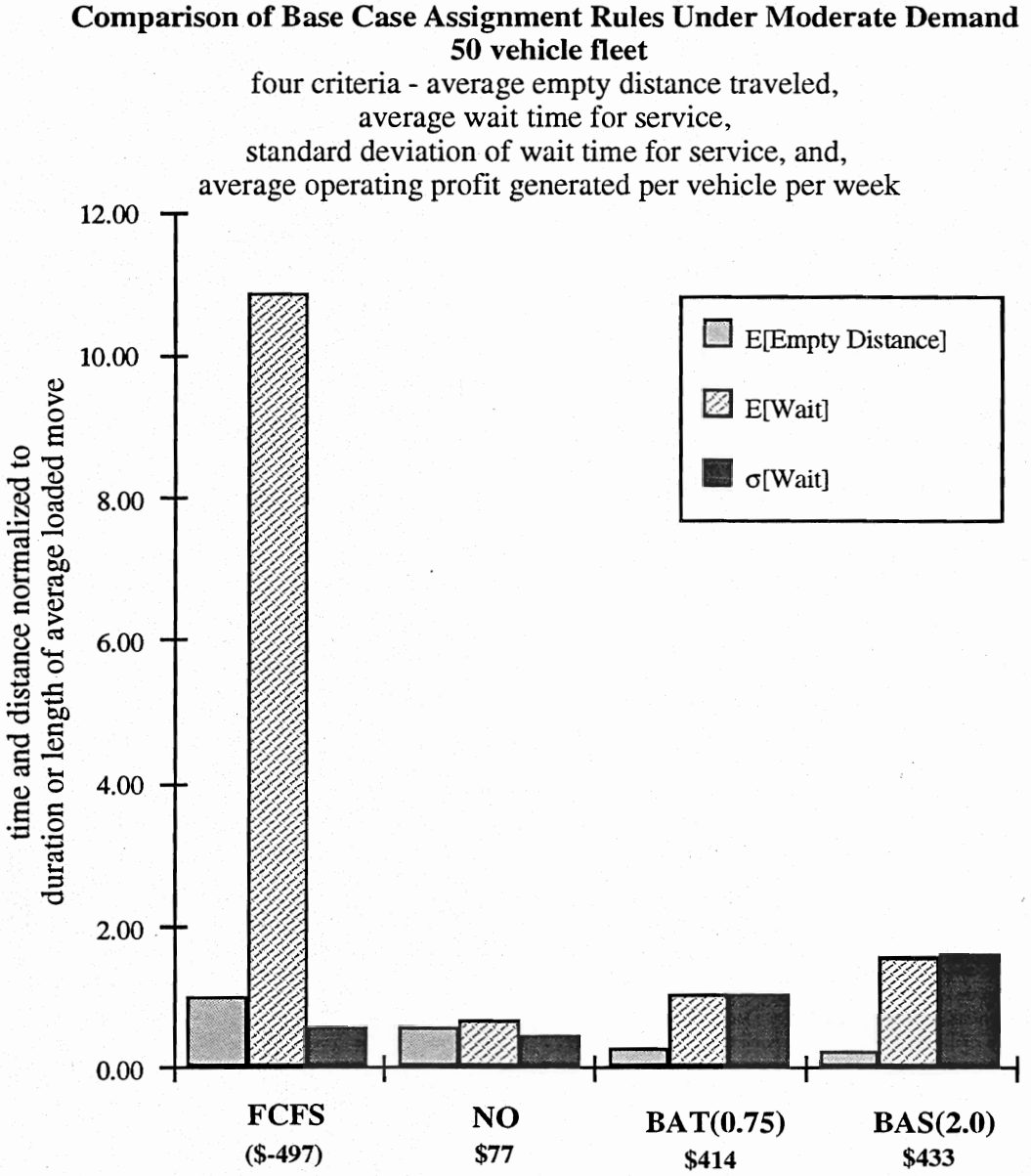


Figure 6.33 Comparison of base cases - moderate demand - 50 vehicles

Comparison of Base Case Assignment Rules Under Low Demand
10 vehicle fleet
four criteria - average empty distance traveled,
average wait time for service,
standard deviation of wait time for service, and,
average operating profit generated per vehicle per week

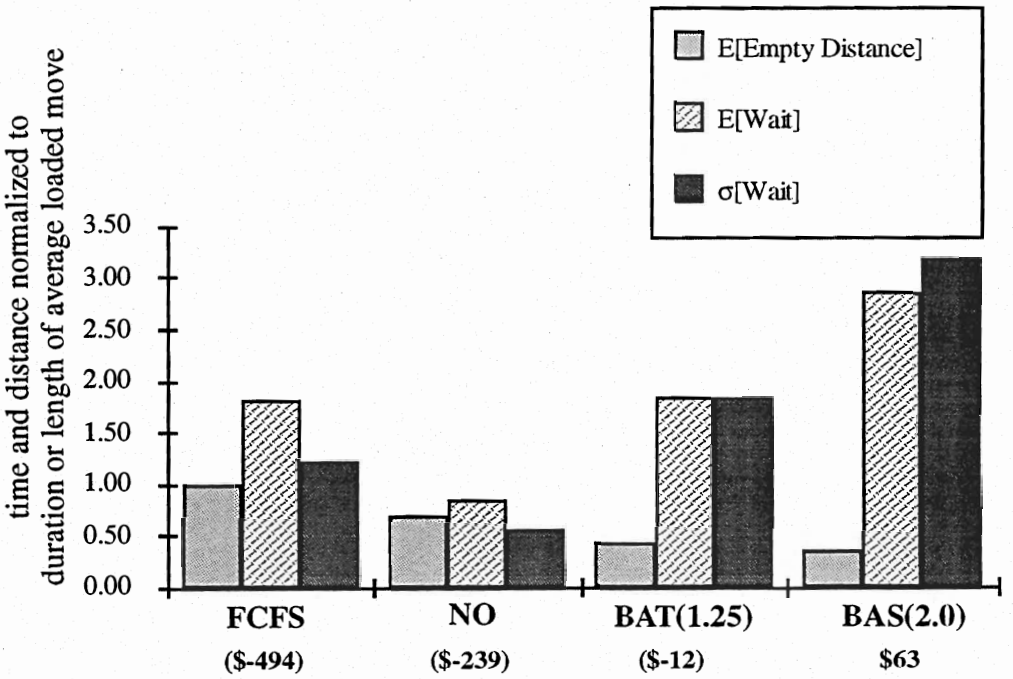


Figure 6.34 Comparison of base cases - low demand - 10 vehicles

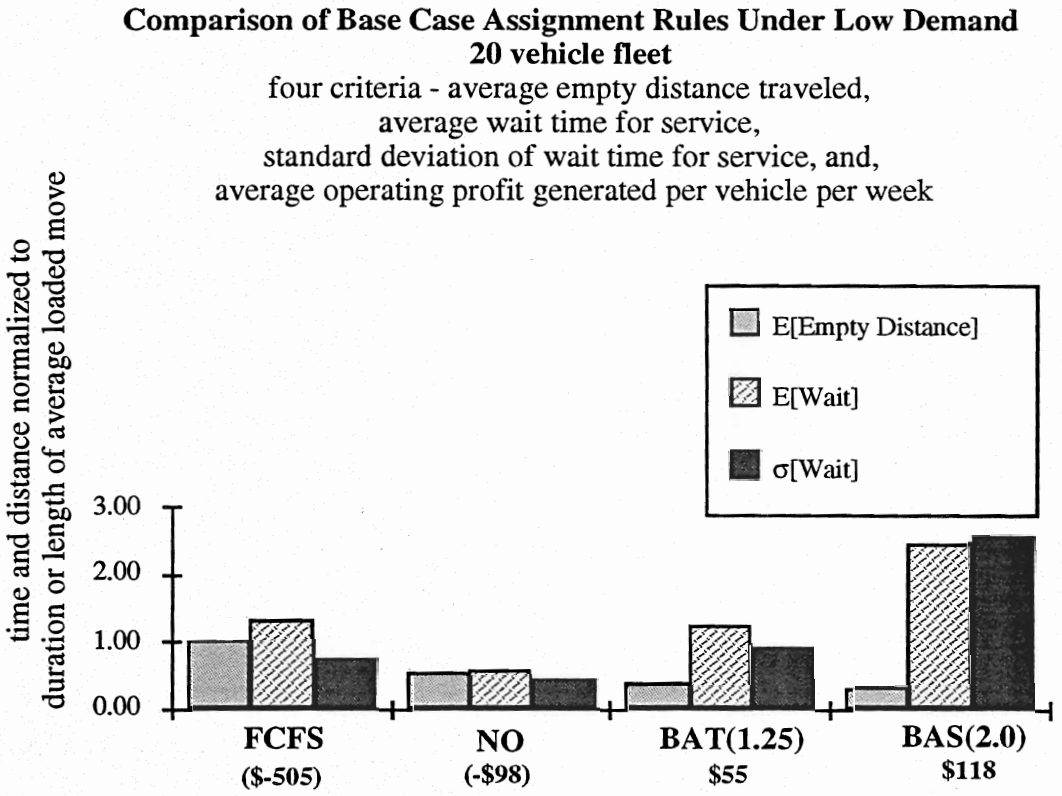


Figure 6.35 Comparison of base cases - low demand - 20 vehicles

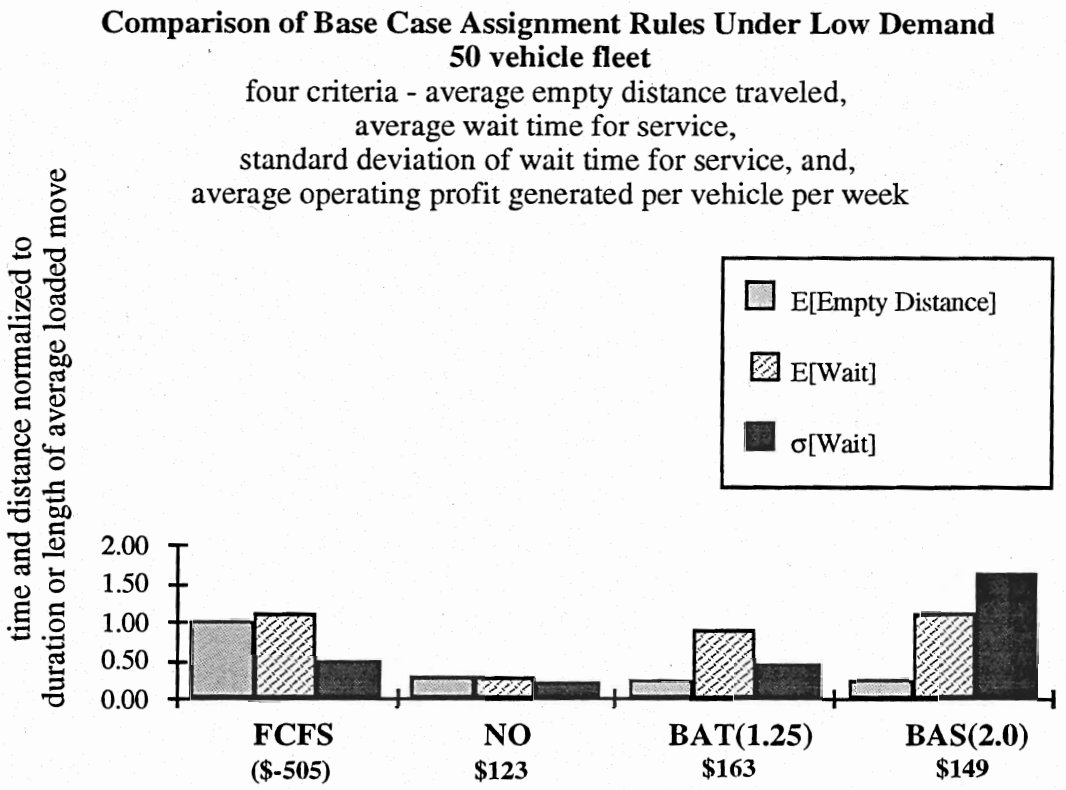


Figure 6.36 Comparison of base cases - low demand - 50 vehicles

Under moderate demand, both BAT(a) and BAS(b) are significantly more efficient, from a distance traveled point of view, than nearest origin assignment. However, under moderate demands, wait times under these strategies exceed those under nearest origin assignment. BAT(0.75) is more profitable than BAS(2.0) but results in longer and more variable wait times for service for fleets of 10 and 20 vehicles but is less profitable and results in shorter and less variable wait times than BAS(2.0) for fleet of 50 vehicles. The reason for the switch is that the 50 vehicle system is more efficient, so its effective congestion (measured as the ratio of loads to vehicles), is less than in the 10 and 20 fleet systems at the same level of demand. So for the same value of a, the systems perform somewhat differently. A slightly larger value for a would result in the same relative performance as seen with a = 0.75 and 10 and 20 vehicle fleets.

When demand for service is low, BAS(2.0) performs well, especially with smaller fleet sizes, because it allows sufficient demands to accumulate before making an assignment. A time-triggered assignment rule that would achieve the same efficiency with respect to distance traveled would have higher variability with respect to wait times for service. When the fleet size is

larger, BAT(1.25) performs better because the larger fleet increases the chance of finding a good solution in each assignment period.

Statistical Significance of Observed Differences. Tables 6.3, 6.4 and 6.5 show the conditions under which differences in observed values of the average empty distance traveled to provide service and the average wait time for service are statistically significant when a test with a level of 1%. It is important to note that the performance of the bipartite assignment strategies varies widely. Statistical significance of performance relative to two primary criteria, average empty distance and wait time for service. are provided.

TABLE 6.1 STATISTICAL SIGNIFICANCE OF OBSERVED DIFFERENCES IN TWO KEY PARAMETERS, UNDER HIGH DEMAND

Average Empty Distance

High Demand	FCFS	NO	BAT(0.5)	BAS(2.0)
FCFS		Y	Y	Y
NO	Y		N	N
BAT(0.5)	Y	N		N
BAS(2.0)	Y	N	N	

Wait Time For Pickup

High Demand	FCFS	NO	BAT(0.5)	BAS(2.0)
FCFS		Y	Y	Y
NO	Y		Y	N
BAT(0.5)	Y	Y		N
BAS(2.0)	Y	N	N	

TABLE 6.2 STATISTICAL SIGNIFICANCE OF OBSERVED DIFFERENCES IN TWO KEY PARAMETERS, UNDER MODERATE DEMAND

Average Empty Distance

Mod Demand	FCFS	NO	BAT(0.75)	BAS(2.0)
FCFS		Y	Y	Y
NO	Y		Y	Y
BAT(0.75)	Y	Y		N
BAS(2.0)	Y	Y	N	

Wait Time For Pickup

Mod Demand	FCFS	NO	BAT(0.75)	BAS(2.0)
FCFS		Y	Y	Y
NO	Y		Y	Y
BAT(0.75)	Y	Y		Y
BAS(2.0)	Y	Y	Y	

COMPARISON B - LOCAL ASSIGNMENT STRATEGIES REQUIRING REAL-TIME INFORMATION.

In this section we compare the performance of the four real-time assignment strategies outlined in chapter 5.

Chapter 6 examines the relative performance of the three local decision rules used to make final assignment decisions under each of the four assignment strategies. Following that analysis, the effects of allowing en-route diversion, reassignment of loads, and profit based load acceptance are examined.

Three Local Decision Rules

Within each of the four local assignment strategies (assignment without en-route diversion or load re-assignment (D^CR^C), assignment with en-route diversion alone (DR^C), assignment with load re-assignment alone (D^CR), and assignment with both en-route diversion and load re-assignment (DR)), three local decision rules are examined. The local decision rules assign loads to the feasible vehicle for which an assignment including the candidate load has the: lowest

empty to loaded ratio (ELR); least overall empty distance to travel (SED); and, the least increase in empty distance to travel (DED).

TABLE 6.3 STATISTICAL SIGNIFICANCE OF OBSERVED DIFFERENCES IN TWO KEY PARAMETERS, UNDER LOW DEMAND

Average Empty Distance

Low Demand	FCFS	NO	BAT(1.25)	BAS(2.0)
FCFS		Y	Y	Y
NO	Y		Y	Y
BAT(1.25)	Y	Y		N
BAS(2.0)	Y	Y	N	

Wait Time For Pickup

Low Demand	FCFS	NO	BAT(1.25)	BAS(2.0)
FCFS		Y	N	Y
NO	Y		Y	Y
BAT(1.25)	N	Y		Y
BAS(2.0)	Y	Y	Y	

No En-route Diversion, No Re-assignment of Loads. The performance of these rules varies much less in deadline-constrained scenarios that in unconstrained scenarios. When pickup deadlines are moderate, rule SED out-performs the others in all cases; when pickup deadlines are tight it out-performs the other rules but differences are much smaller. Figures 6.37 and 6.38 illustrate the relative performance of these three rules under different levels of demand, with and without pickup deadlines. The criterion for evaluation shown in figure 6.37 is the average empty distance moved. This appears to be the most robust indicator of the overall effectiveness of the system when comparing cases with the same pickup deadline distribution and the same demand arrival patterns when a capacity or feasibility only load acceptance rule is used. (Under a profit based load acceptance rule the ratio of empty to loaded distances can be better estimator of efficiency). The criterion for evaluation in figure 6.38 is the average wait time for service.

It may be observed in figures 6.37 and 6.38 that the relative performance of the three assignment rules is not consistent over varying levels of demand in the unconstrained systems. Simulation experiments across a set of more finely discretized demand levels illustrate this behavior more clearly. Figure 6.39 present the simulation results, for a fleet of 10 vehicles, and demand levels ranging from low to high, where the low and high demand levels correspond exactly to the definitions used elsewhere in the chapter. Three evaluation criteria are presented: average empty distance, average wait time for service, and, average operating profit generated (per vehicle, per week).

These results show that rule SED results in uniformly lower wait times for service. This follows directly from the fact that SED loads the vehicles more evenly, in what might be described as a "round robin" fashion, rather than loading only a subset of the fleet when demand is moderate or light.

What is most interesting, is the variability of the performance of SED with respect to the criteria of empty distance driven and hence, operating profitability (displayed in the graph on top in figure 6.39). It may be observed in figure 6.40, which presents corresponding results for the 10 vehicle, moderate deadline constrained case, that this variability is not observed when pickup constraints are in place. In the time constrained system, SED performs better than competing assignment rules with respect to all criteria. The reason SED is well suited to the pickup constrained assignment is the same reason the wait times are lower in the unconstrained cases; in pickup constrained cases the number of loads assigned to each vehicle "route" will be limited, even in the high demand scenarios. The behavior dictated by the constraints is the same behavior followed by the SED assignment rule in scenarios without time constraints.

The question remains: why does SED perform well at high demands, and so poorly at moderate demand levels in the unconstrained scenario? The answer may be found in figure 6.41, which displays the average empty distance driven as a function of r, the average rate of arrival of requests (per vehicle), divided by the average service rate. Under moderate demand, ELR and DED perform well by capitalizing on route building opportunities for a subset of vehicles. These compact routes tend to result in lower distances traveled. However, as r approaches 1.0, the queue limit on each vehicle becomes a binding constraint and fewer vehicles are available to take new requests. SED, on the other hand, loads vehicles evenly so that even as the system nears its maximum capacity there are several feasible choices for each new service request. SED takes a steep downturn in efficiency (increase in empty distance) when the capacity limit is reached -- in other words for $r > 1.0$. At that point none of the assignment rules performs much better than the others -- all are equally constrained.

Average empty distance under ELR, ΣED and ΔED assignment rules

strategy D^CR^C, three demand levels, three distributions of pickup deadlines, 10 vehicle fleet. feasibility only load acceptance rule

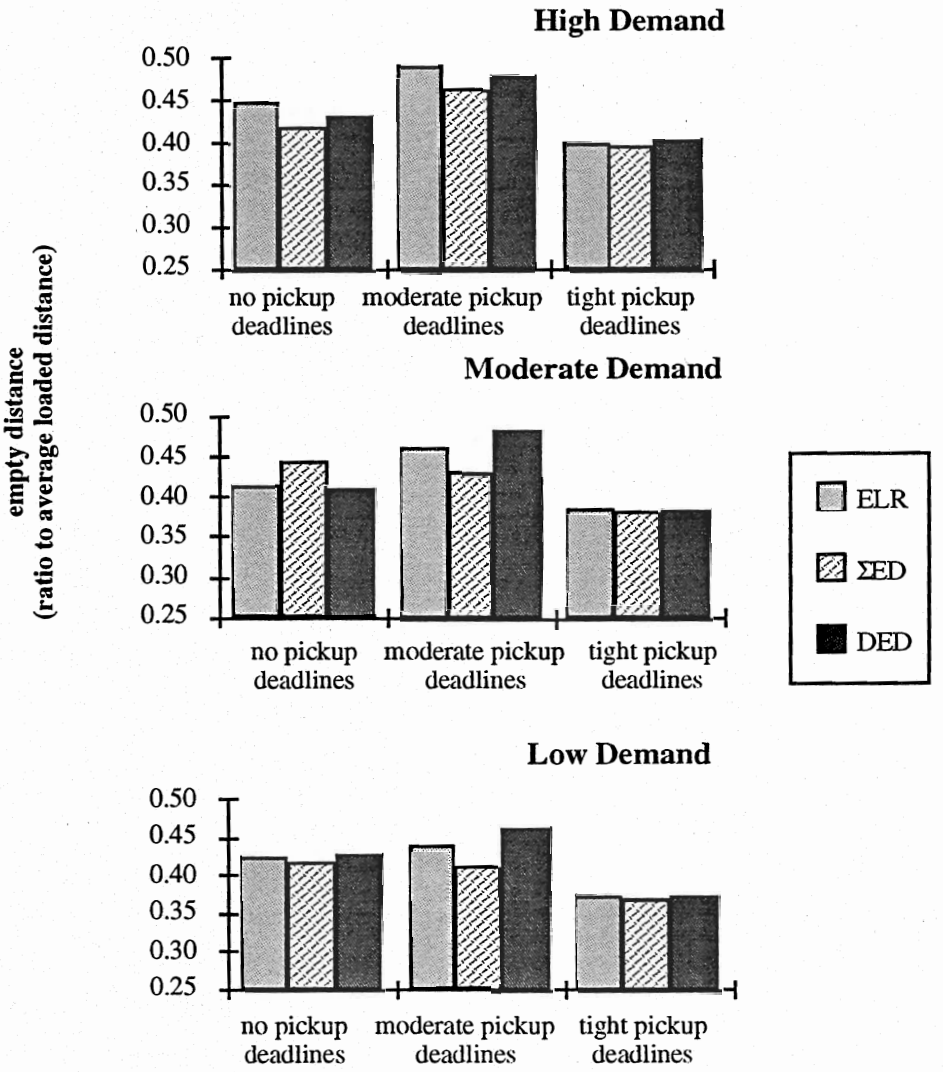


Figure 6.37 Comparison of average empty distance: three assignment rules under assignment strategy D^CR^C, no load acceptance thresholds applied, 10 vehicle fleet

Average wait time for service under ELR, ΣED and ΔED assignment rules

strategy D^CR^C, three demand levels, three distributions of pickup deadlines, 10 vehicle fleet. feasibility only load acceptance rule

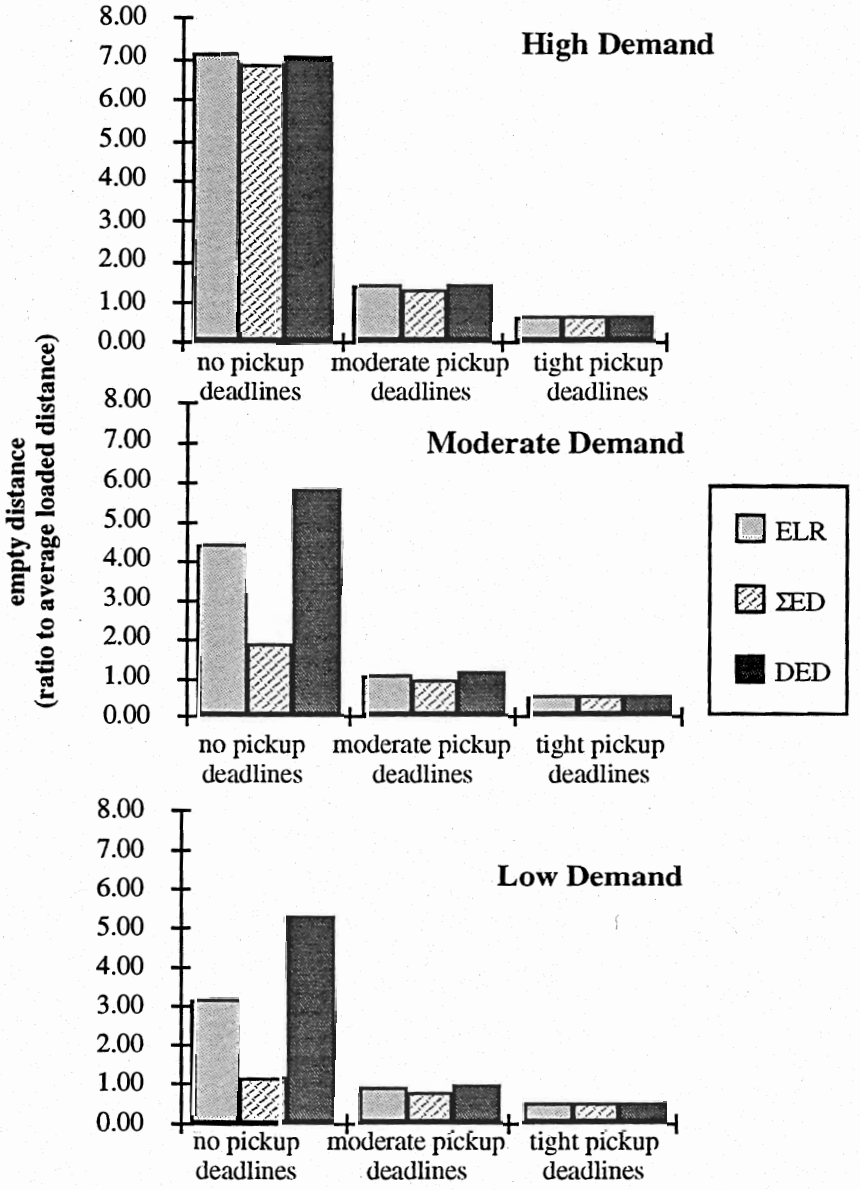


Figure 6.38 Comparison of wait time for service: three assignment rules under assignment strategy D^CR^C, no load acceptance thresholds applied,10 vehicle fleet

The observation that the relative performance of the assignment rules is due to changes in the congestion level of the system suggests that comparisons performed over different fleet sizes might yield different conclusions. In experiments performed the overall rate of service requests is proportional to fleet size so larger fleets have a higher geographic concentration of demands. These more heavily congested systems should favor SED assignment, while less congested systems should favor the assignment rules which build routes. This may be observed in figure 6.42, which illustrates the difference in the relative performance of the three rules across fleets of 2, 5, 10 and 20 vehicles. Figures 6.43 and 6.44 present simulation results over more finely discretized demand levels for fleets of five and twenty vehicles. While the general trends observed correspond to those in figure 6.39, for a fleet of 10 vehicles, there are marked differences. Because the system is less congested with a fleet of five vehicles, the region over which SED is dominated with respect to profitability and distance traveled is longer than in the ten vehicle scenario; in the twenty vehicle scenario SED dominates at all demand levels, but less so when demand is low.

Figure 6.45 displays the performance of SED assignment relative to r, for all three fleet sizes. The same general pattern may be observed, but with a progressive muting effect, due on increase in the overall congestion of the system corresponding to an increase in fleet size.

In most cases, observed differences in the performance of the assignment rules are not statistically significant (at a meaningful level) except with respect to the wait time for service. However, most of the scenarios examined in this section have been simulated over 1000 iterations and clear patterns of differences have emerged.

En-route Diversion, Re-assignment of Loads. Chapter 6 address the effects of allowing en-route diversion and re-assignment of loads. Of interest is the relative performance of the local assignment rules with and without the flexible assignment strategies, en-route diversion and re-assignment of loads. Tables 6.3 and 6.4 illustrate the difference in the performance of the three local decision rules, with and without pickup deadlines, for each of the four real-time assignment strategies, for three criteria: the average distance traveled empty, the average wait time for service and the operating profit generated. Noticeable about these tables is the same absence of consistency examined in the last section for the D^cR^c assignment strategy, namely, that in the absence of pickup deadlines no assignment rule dominates in all cases. As mentioned in the last section, when pickup deadlines are in place, SED, which loads vehicles more evenly, thereby opening up future opportunities to find assignments for deadline-constrained service requests, is the best performer with respect to all criteria examined. In some cases however, these

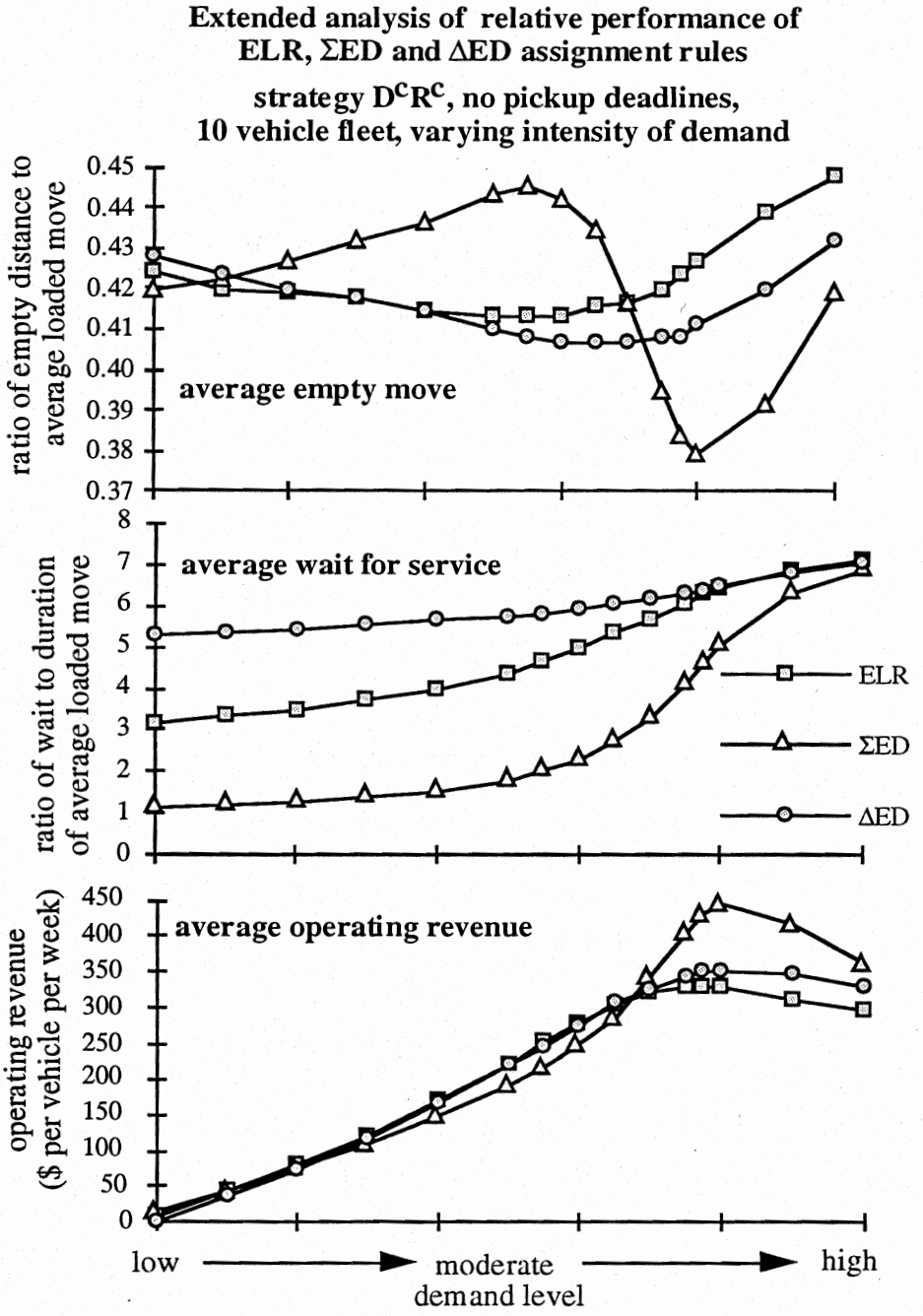


Figure 6.39 Relative performance of ELR, SED and DED across a set of finely discretized demand levels - 10 vehicle fleet, no pickup deadlines

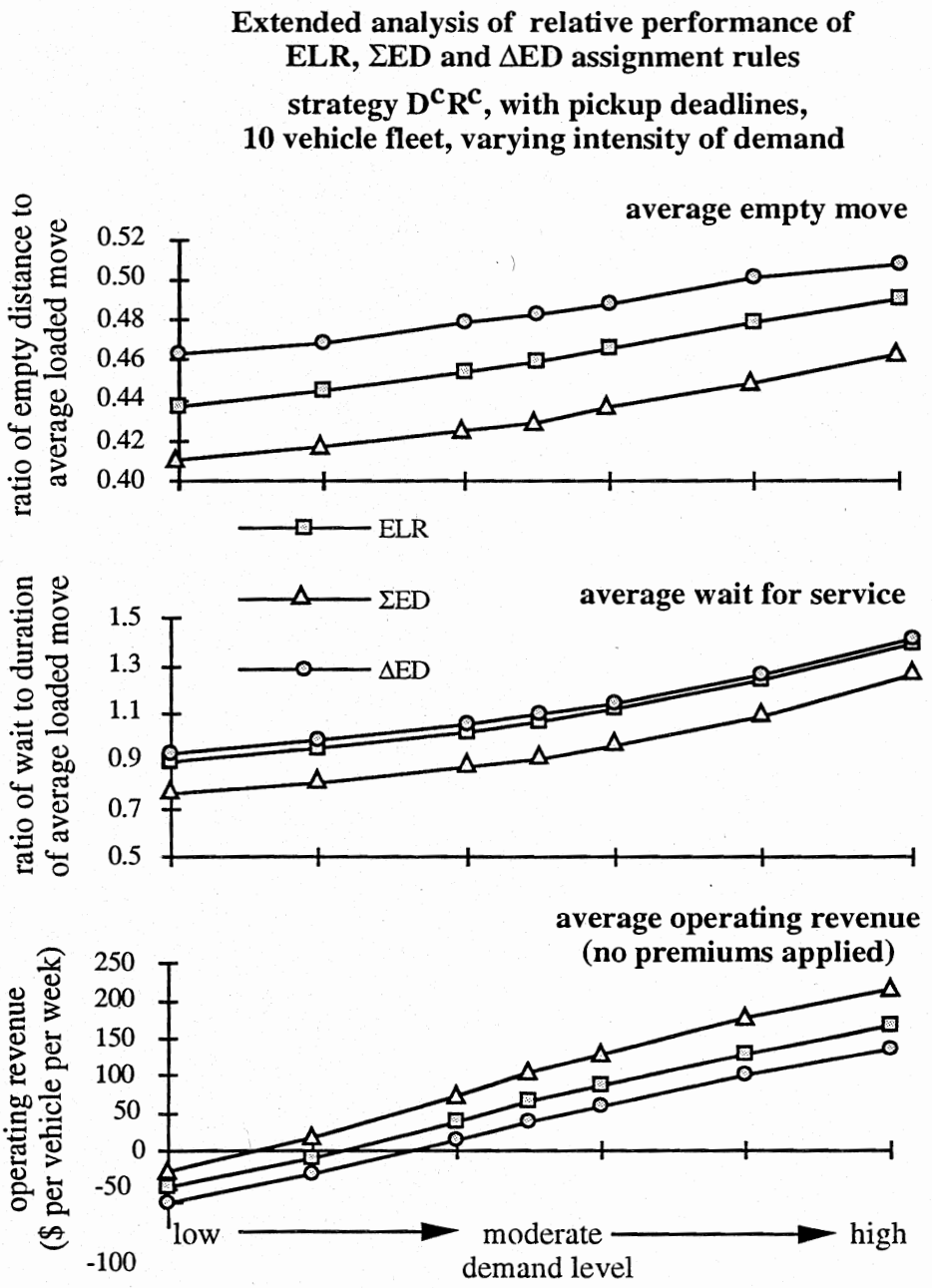


Figure 6.40 Relative performance of ELR, SED and DED across a set of finely discretized demand levels - 10 vehicle fleet, with pickup deadlines

differences may be quite small, and not statistically significant, even when one thousand realizations of the simulation program are evaluated. Under tight pickup deadlines there are few feasible solutions, hence all three decision rules tend to make the same assignments. Differences in performance may be barely noticeable. Performance is examined with respect to each of the criteria, average empty distance, average wait time for service and operating profit generated and is expressed as a percent of the value in the best case.

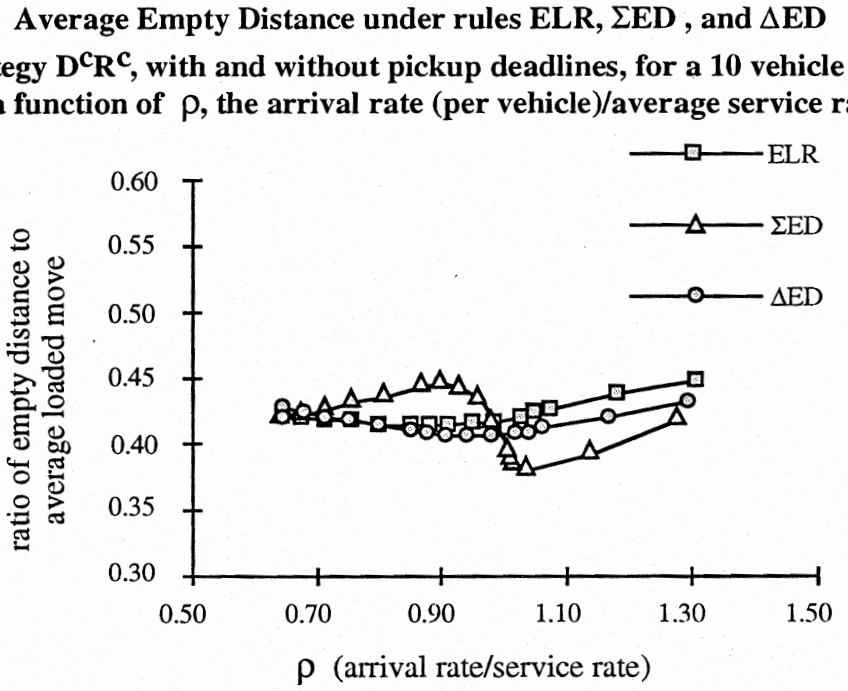


Figure 6.41 Average empty distance under rules ELR, SED and DED as a function of utilization level, (when r > 1, experienced r ≈ 1.0)

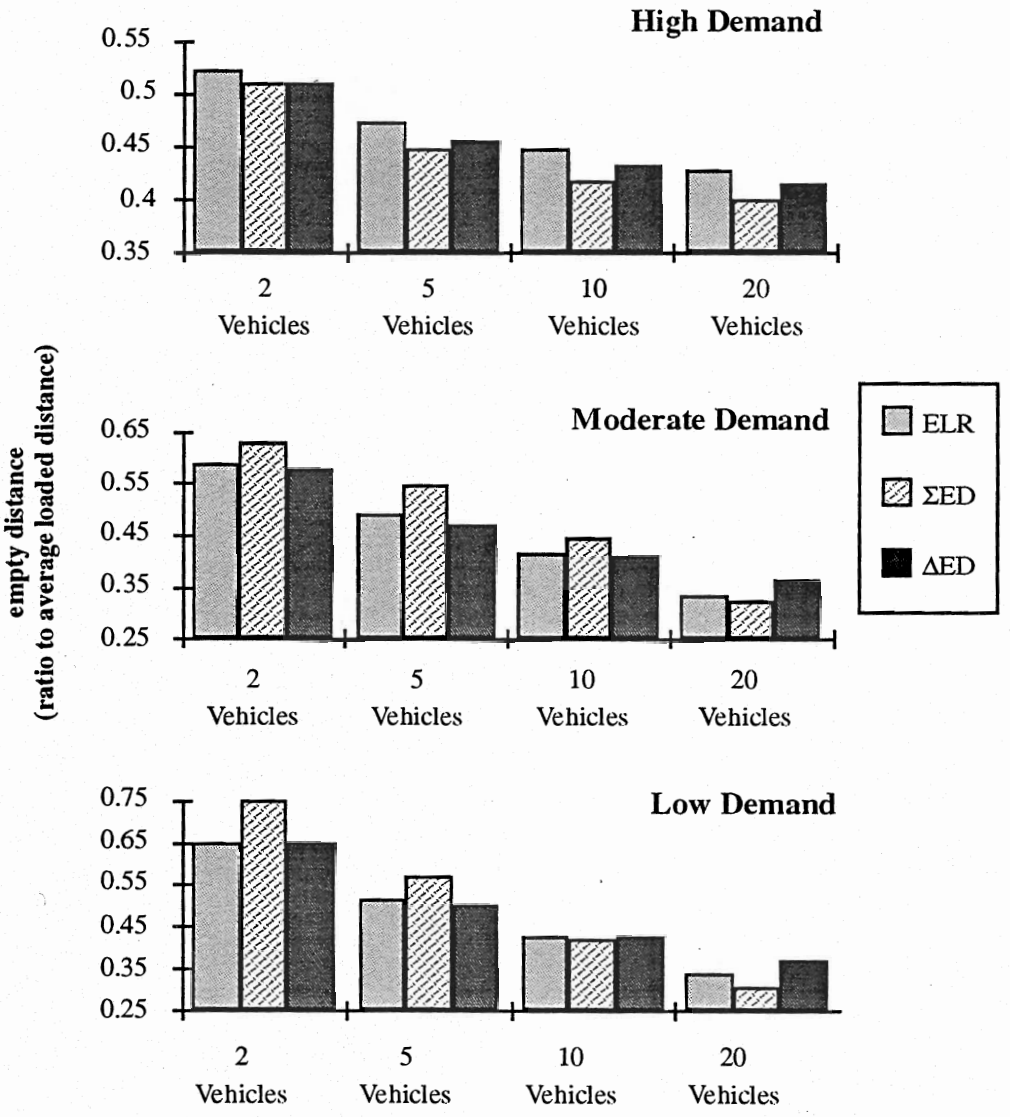


Figure 6.42 Comparison of average empty distance driven: three assignment rules, three demand levels, four fleet sizes, under assignment strategy D^CR^C (scale inconsistent across demand levels)

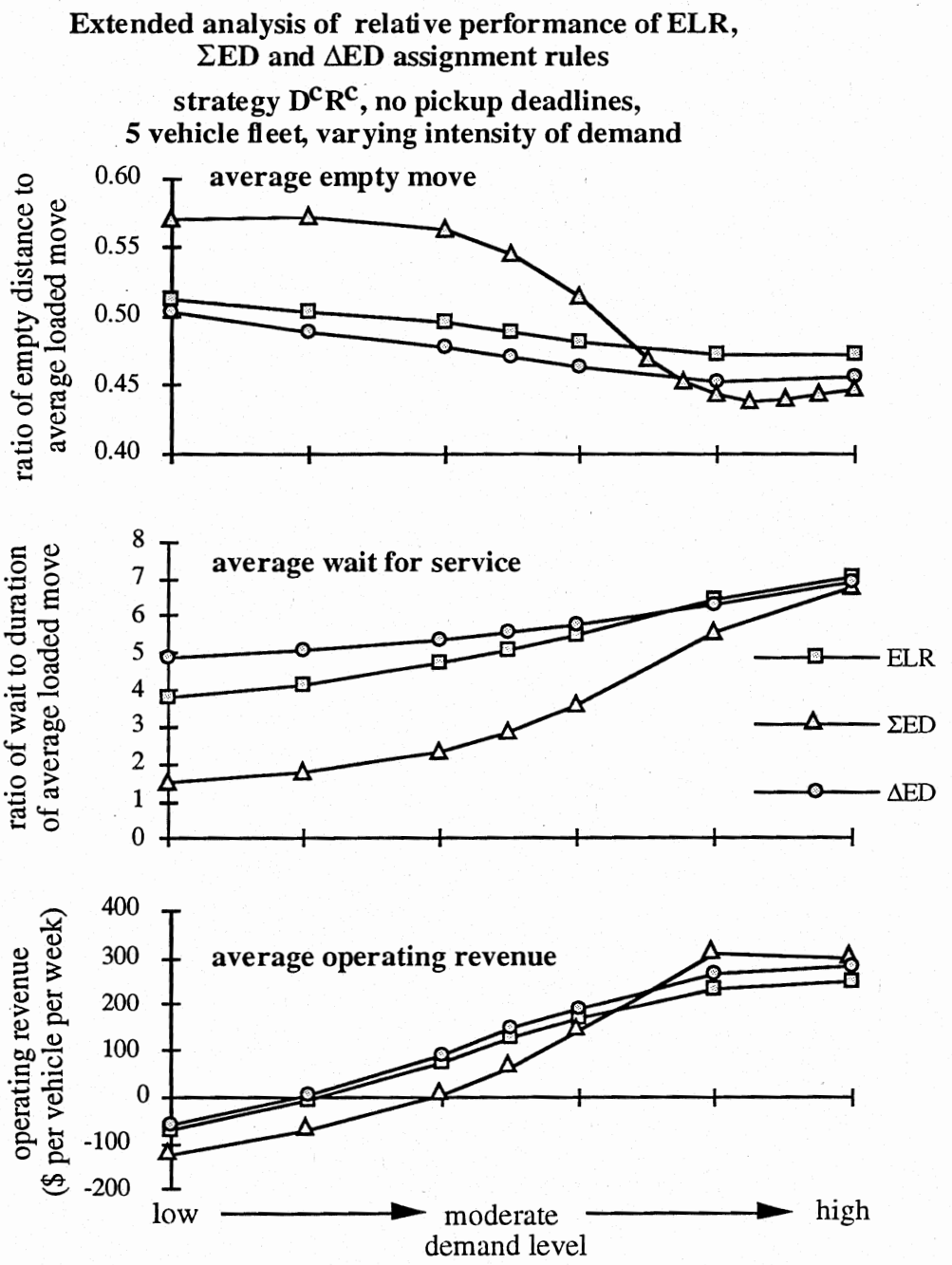


Figure 6.43 Relative performance of ELR, SED and DED across a set of finely discretized demand levels - 5 vehicle fleet, no pickup deadlines

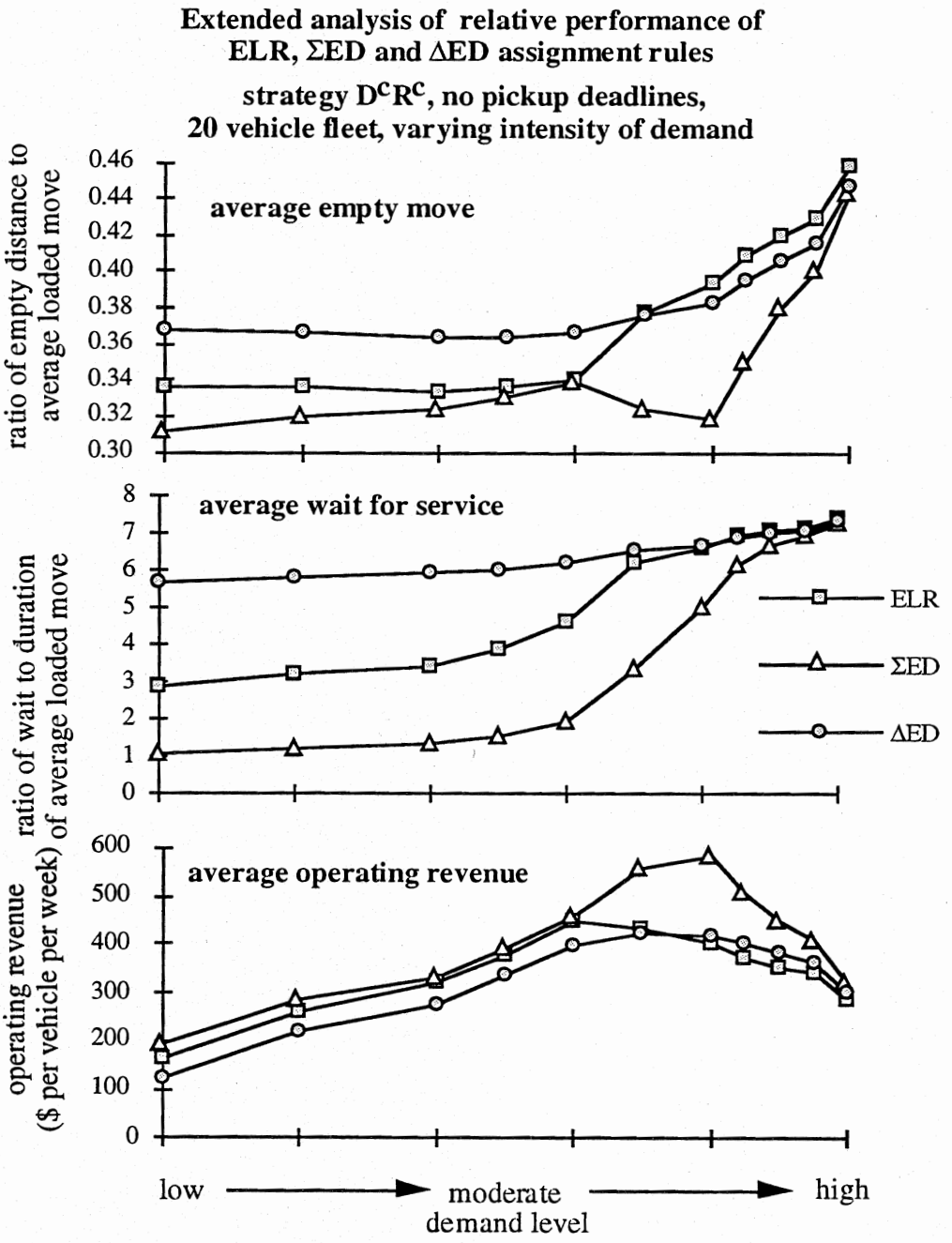


Figure 6.44 Relative performance of ELR, SED and DED across a set of finely discretized demand levels - 20 vehicle fleet, no pickup deadlines

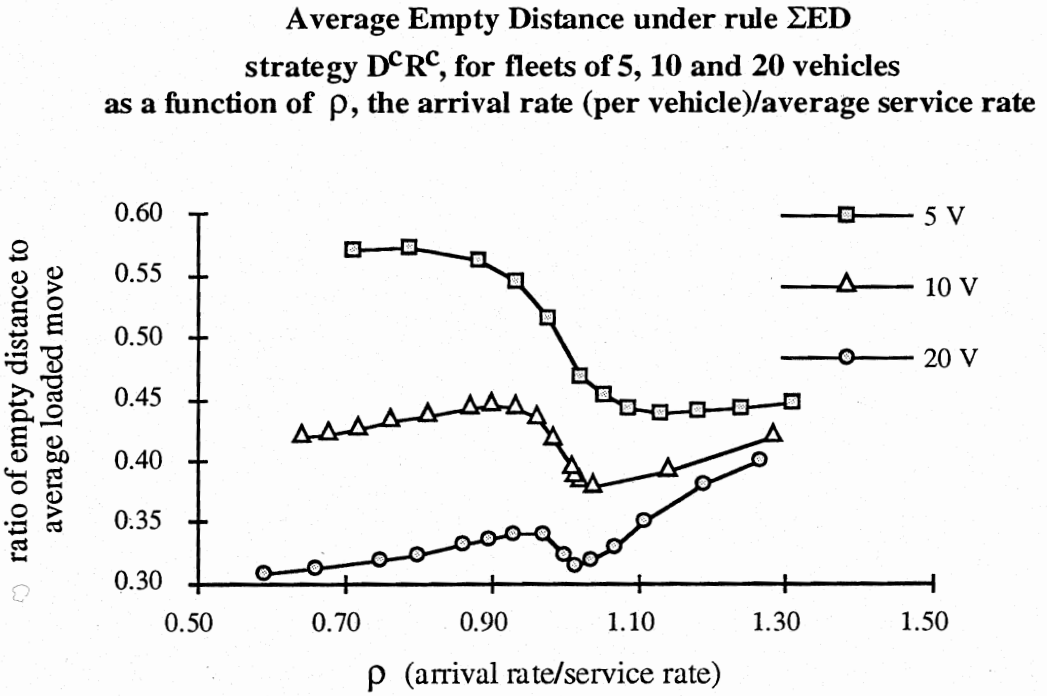


Figure 6.45 Average empty distance under rules ELR, SED and DED as a function of utilization level, (when $\rho > 1$, $\rho \approx 1.0$)

Effect of En-route Diversion

In this section, the effect of allowing the en-route diversion of vehicles is examined. The focus of this examination are assignment strategies D^CR^C and DR^C (cases 6 and 7 in figure 5.4). When en-route diversion is allowed, it is chosen a fraction of 0 to 0.20 times per load served in the cases examined. En-route diversion is chosen with much more frequency without pickup deadlines than with moderate or tight deadlines. When pickup deadlines are binding constraints the system lacks the flexibility to divert a driver en-route to an already assigned load to a newly arriving load. In fact, when pickup deadlines are tight en-route diversion takes place less than once for every one hundred loads served.

TABLE 6.4 RELATIVE PERFORMANCE OF THREE LOCAL DECISION RULES WHEN INCORPORATED WITH FOUR REAL-TIME ASSIGNMENT STRATEGIES. NO PICKUP DEADLINES,10 VEHICLE FLEET

high demand	D^cR^c			DR^c			D^cR			DR		
	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$
ELR	107%	104%	83%	109%	104%	81%	108%	105%	89%	111%	107%	86%
ΣED	-	102%	-	103%	-	96%	-	-	-	-	-	-
ΔED	103%	-	91%	-	101%	-	106%	103%	92%	105%	105%	94%

moderate demand	D^cR^c			DR^c			D^cR			DR		
	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$
ELR	101%	242%	95%	115%	354%	79%	101%	227%	97%	108%	218%	90%
ΣED	107%	-	83%	124%	-	68%	121%	-	69%	129%	-	66%
ΔED	-	317%	-	-	376%	-	-	232%	-	-	253%	-

low demand	D^cR^c			DR^c			D^cR			DR		
	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$
ELR	101%	278%	50%	114%	284%	30%	106%	251%	68%	110%	246%	64%
ΣED	-	-	-	114%	-	24%	114%	-	29%	119%	-	32%
ΔED	102%	461%	19%	-	316%	-	-	260%	-	-	315%	-

ELR = least empty to loaded ratio
ΣED = least overall empty distance
ΔED = least additional empty distance
E[E] = average empty distance traveled
E[W] = average wait time for service
$ = operating profit

D^cR^c = no en-route diversion , no re-assignment
DR^c = en-route diversion , no re-assignment
D^cR = no en-route diversion , re-assignment
DR = en-route diversion & re-assignment

TABLE 6.5 RELATIVE PERFORMANCE OF THREE LOCAL DECISION RULES WHEN INCORPORATED WITH FOUR REAL-TIME ASSIGNMENT STRATEGIES. MODERATE PICKUP DEADLINES, 10 VEHICLE FLEET

high demand	D^cR^c			DR^c			D^cR			DR		
	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$
ELR	106%	111%	88%	106%	111%	84%	106%	111%	88%	105%	111%	70%
ΣED	-	-	-	-	-	-	-	-	-	-	-	-
ΔED	103%	109%	93%	103%	109%	92%	104%	109%	93%	103%	109%	73%

moderate demand	D^cR^c			DR^c			D^cR			DR		
	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$
ELR	107%	117%	87%	107%	116%	94%	107%	117%	88%	106%	115%	88%
ΣED	-	-	-	-	-	-	-	-	-	-	-	-
ΔED	112%	120%	91%	104%	114%	96%	112%	120%	91%	104%	113%	92%

low demand	D^cR^c			DR^c			D^cR			DR		
	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$	E[E]	E[W]	$
ELR	106%	118%	85%	106%	118%	83%	106%	118%	85%	106%	117%	80%
ΣED	-	-	-	-	-	-	-	-	-	-	-	-
ΔED	103%	122%	74%	103%	114%	92%	103%	122%	89%	105%	114%	85%

ELR = least empty to loaded ratio
ΣED = least overall empty distance
ΔED = least additional empty distance
E[E] = average empty distance traveled
E[W] = average wait time for service
$ = operating profit

D^cR^c = no en-route diversion , no re-assignment
DR^c = en-route diversion , no re-assignment
D^cR = no en-route diversion , re-assignment
DR = en-route diversion & re-assignment

En-route diversion is also selected much more often under high demand levels. When demands are moderate or low there is often an idle and available driver to provide service. When demands are high it may be that the only driver(s) that can serve a load within its time constraints must be diverted to perform the service. Figure 6.46 shows how the rate of diversion varies across demand levels (rates of arrivals of requests) and assignment rules for a 10 vehicle fleet. The simulation results strongly suggest that allowing en-route diversion can improve the efficiency of an operation, both with respect to distance traveled empty, wait time for service and profitability. Figures 6.47 - 6.49 show the fairly dramatic increase in efficiency observed for simulation experiments conducted without pickup deadlines, figures 6.50 and 6.51, the corresponding minor increase in efficiency observed for systems in which loads have associated pickup deadlines.

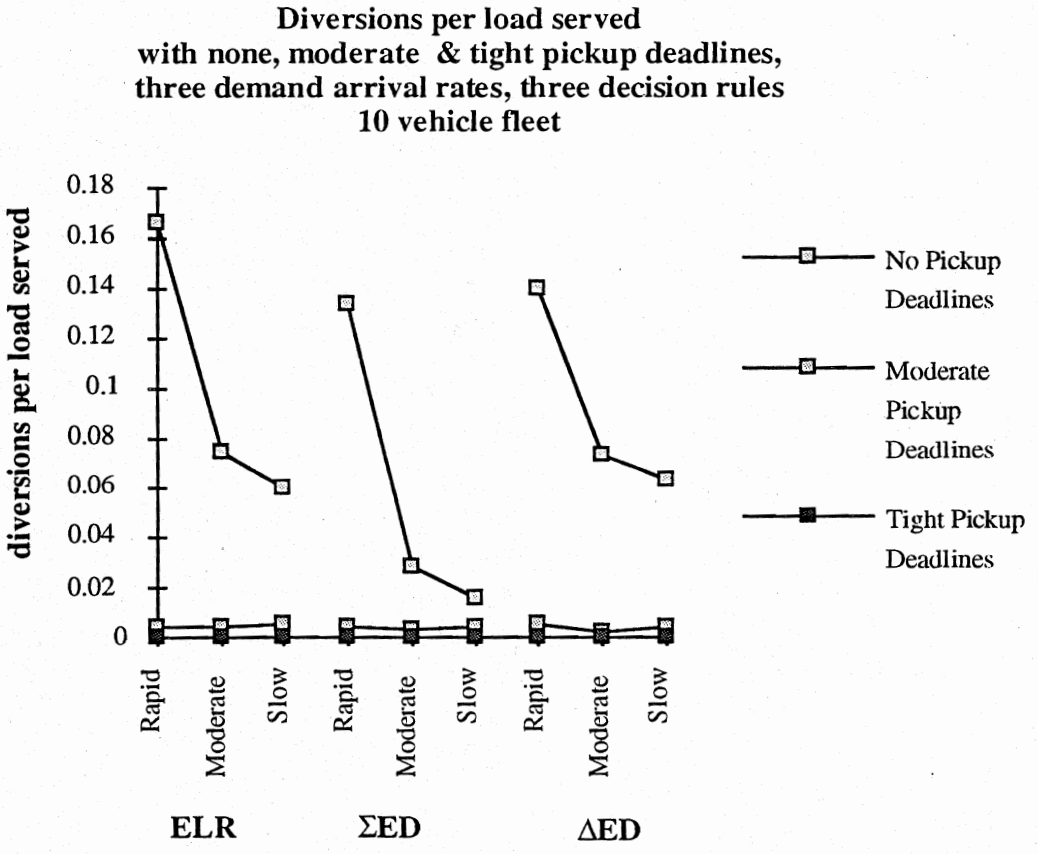


Figure 6.46 Diversions per load served

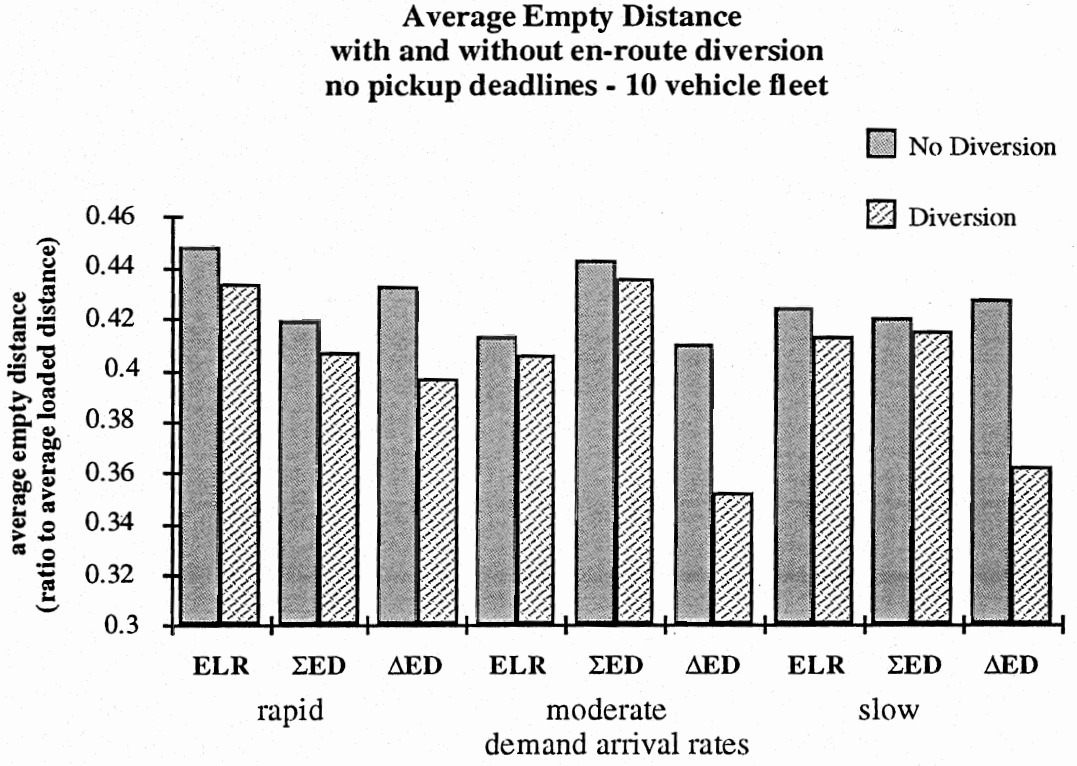


Figure 6.47 Empty distance traveled, with and without en-route diversion, no pickup deadlines

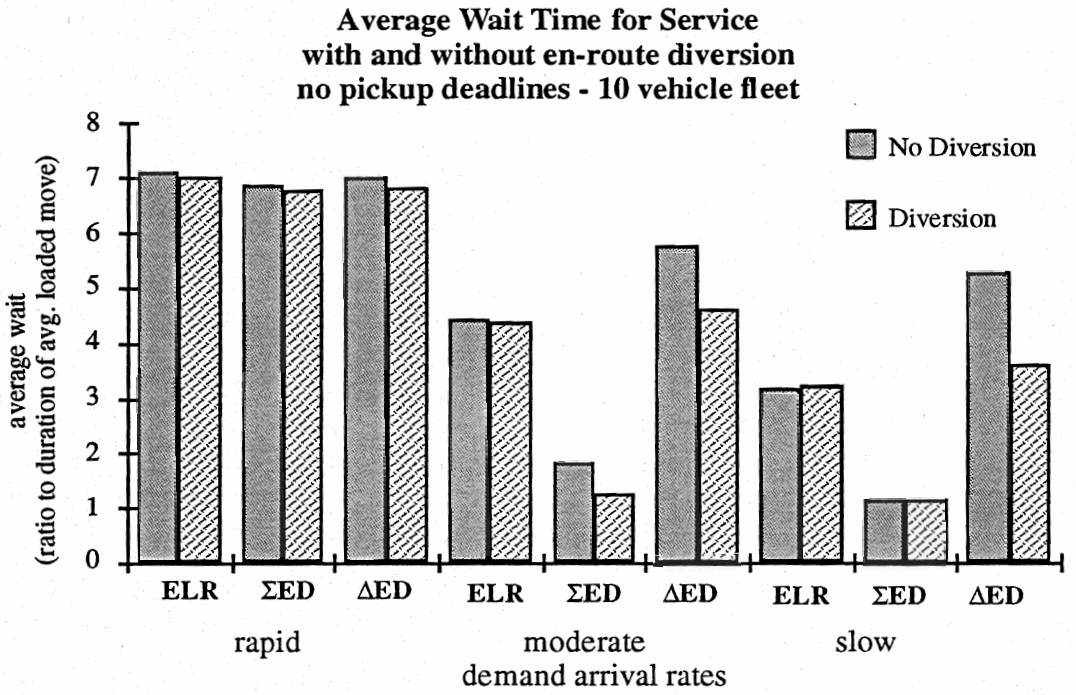


Figure 6.48 Average wait time for service, with and without en-route diversion, without pickup deadlines

Effect of Re-Assignment of Loads

An advantage of systems incorporating real-time communication is that loads assigned to one driver can be re-assigned to another, even if the first driver has been notified of the assignment. The simulation experiments discussed here explore a very simple re-assignment strategy. Under this strategy, after each newly arriving load is assigned, all vehicles are examined for load re-assignment. For each vehicle that has been assigned more than two loads, the last load is removed. This load is then a candidate for assignment to any vehicle. The "best" assignment is found, and if this assignment leads to a reduction in the overall empty distance traveled it is accepted. The best overall assignment is defined in the same way as it is for newly arriving loads. That is, once deadline feasibility is established, one of the three local decision rules is used to make the final assignment. The new assignment must pass one more test before the switch is final. The sum of the empty distances in the "routes" of the affected vehicles must be less after the switch than it was prior to the switch. Otherwise, the load is re-assigned to the vehicle from which it was removed. Loads can be assigned to a different vehicle up to five times (not a binding constraint in the systems examined), a vehicle may only have one load removed in each iteration of the rule.

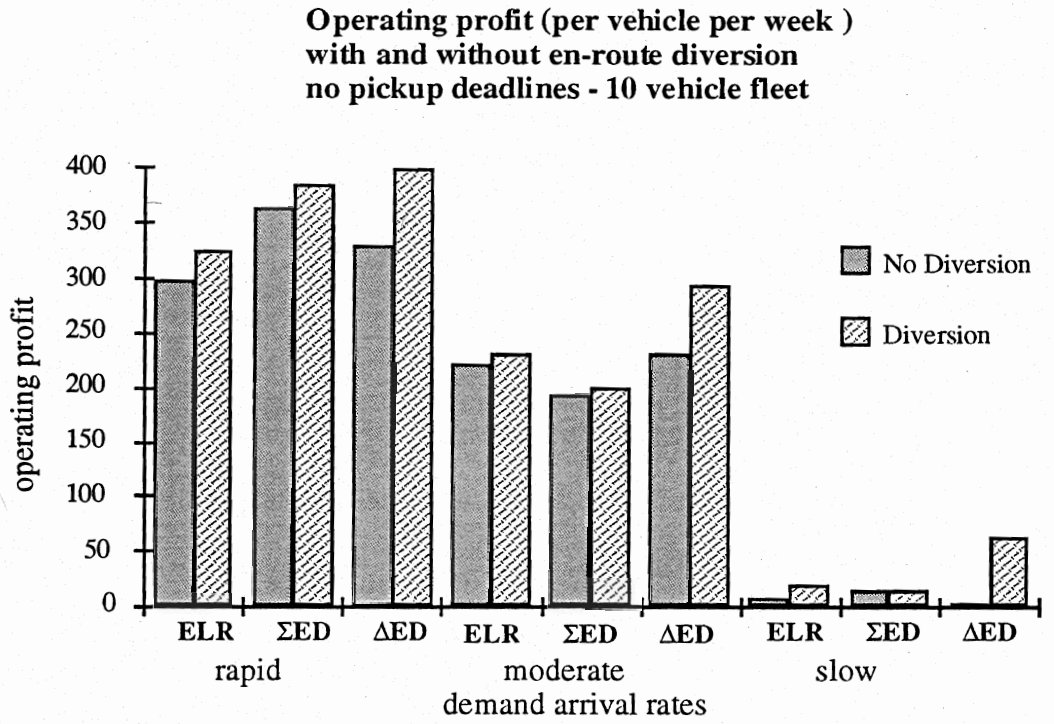


Figure 6.49 Operating profit, with and without en-route diversion, without pickup deadlines

This is clearly not an example of the best re-assignment rule. Intuition suggests that other local route improvement heuristics (2-opt or 3-opt for example) should be more beneficial. Improvements found using even this simple rule are significant. Reductions of more than 25% in the average empty distance traveled are observed in cases without deadline constraints, along with a corresponding reduction in the average wait time for service. Reductions are not uniform across demand levels or assignment rules however. The greatest improvements are found in the high demand case, which is more likely to have more candidates for reassignment. Improvements measured in the pickup deadline constrained cases examined were not statistically significant, although it appears that allowing re-assignment leads to a small improvement in system efficiency. Once a load is assigned in the deadline constrained cases, it is unlikely that another feasible assignment that is more efficient can be found. In the cases examined with tight pickup deadlines, no loads were re-assigned. Figure 6.52 displays the extent of the reduction in empty distances traveled to provide service when re-assignment is allowed when loads do not have associated pickup deadlines. The average empty distance moved under each of the three local decision rules under three levels of demand intensities is also shown with and without re-assignment of loads. It may be observed that the ELR and DED assignment rules benefit much

more from the re-assignment rule under moderate and low demand than the SED rule; in fact, SED, which dominated, with respect to the criterion of empty distance traveled without re-assignment, is no longer the most attractive option. Under SED, in a moderate and low demand environment, vehicles are loaded more evenly than other the two other rules. This even loading of vehicles precludes re-assignment under the re-assignment rule described. Pickup deadlines force the even loading of vehicles and as a result, the trading off of dominance between assignment rules was not observed in time-constrained systems examined.

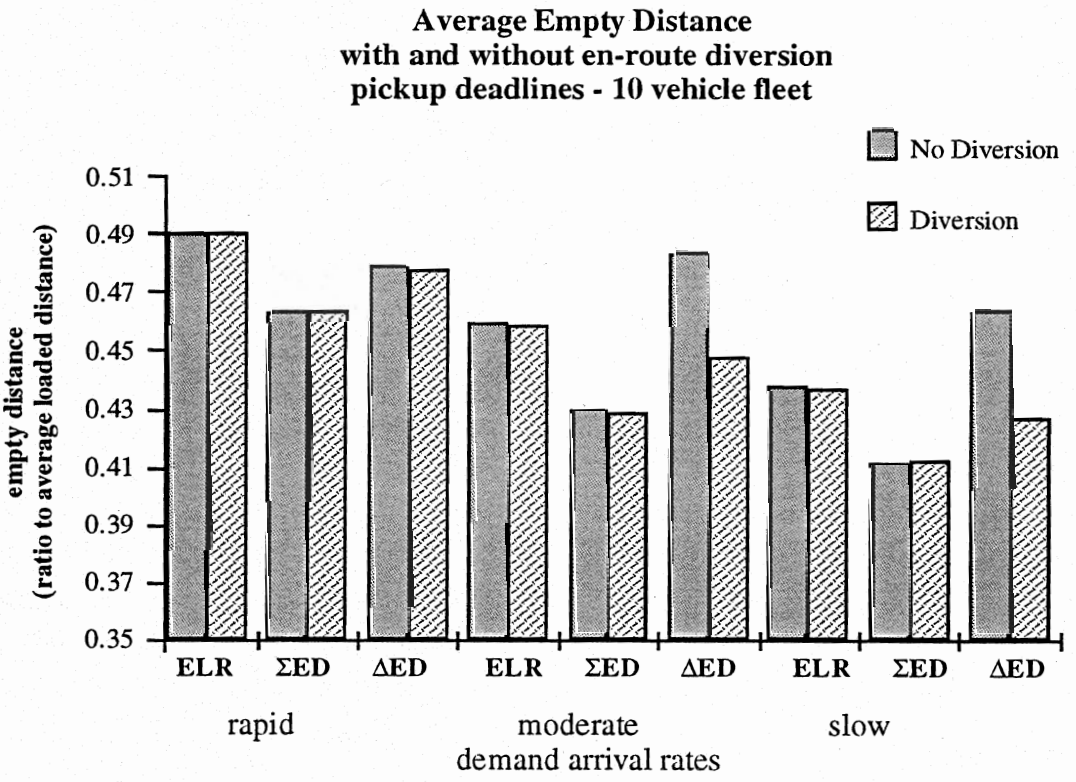


Figure 6.50 Average empty distance traveled, with and without en-route diversion, with moderate pickup deadlines

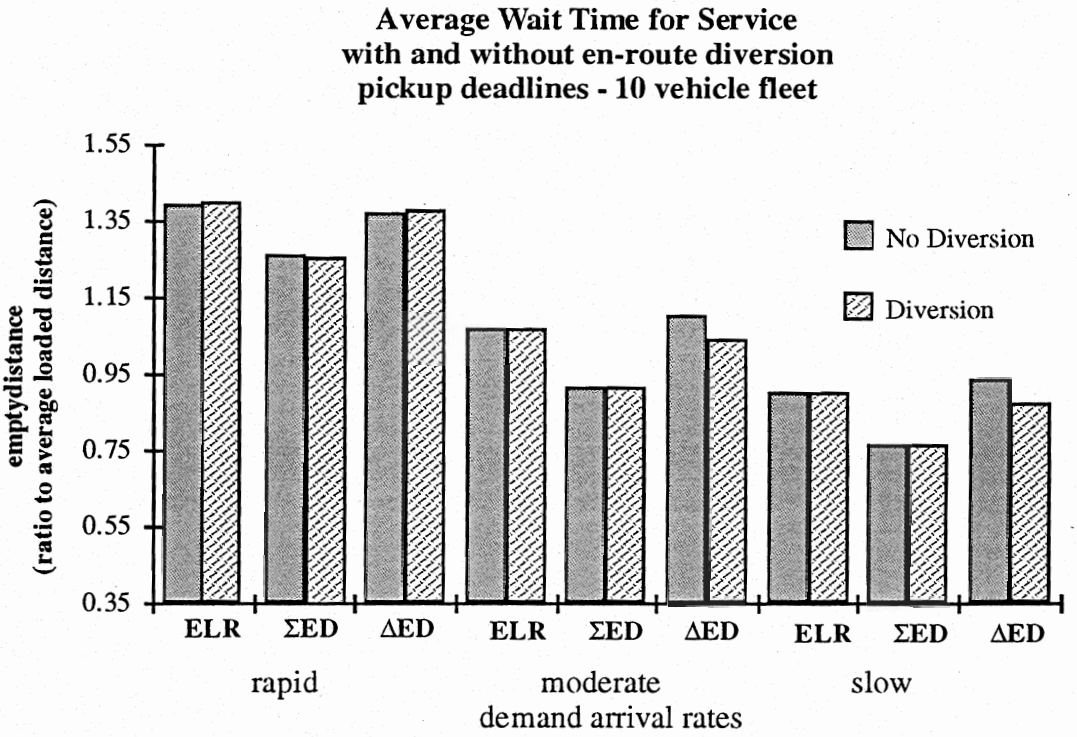


Figure 6.51 Average wait time for service, with and without en-route diversion, with moderate pickup deadlines

Combined Effect of En-Route Diversion and Re-Assignment of Loads

In this section, the combined effect of allowing en-route diversion and re-assignment of loads is examined. When both are allowed, these strategies are invoked in the following way: 1) when a request for service arrives, an attempt is made to assign the load to a pickup deadline feasible route, 2) the route is chosen based on whichever local assignment rule, ELR, SED, or DED, is in use, vehicles en-route to a pickup location are candidates for immediate assignment to the new load. 3) After an assignment is made, each vehicle is considered for load re-assignment, which, for the candidate vehicle amounts to load removal.

The combined effect of allowing en-route diversion and re-assignment of loads is significant- particularly when there are no pickup deadlines. The flexibility of the system in these cases allows the flexible assignment strategies to be evoked more often and with impressive results. With deadlines, the benefits are tangible in some cases, but damped because opportunities for re-assignment are scarce. Figure 6.53 shows the reduction in empty distance traveled when en-route diversion and load re-assignment are employed in the no-pickup deadline case, while figure 6.54 shows the corresponding increase in operating profits earned - in the high demand case an

increase of nearly 80%. Increases are observed in the low demand scenario as well. Under low demand the operating profits earned are very low (typically $10-50 per vehicle per week). In those cases a small increase can be a very large percentage increase and expressing it in this form exaggerates the actual increase.

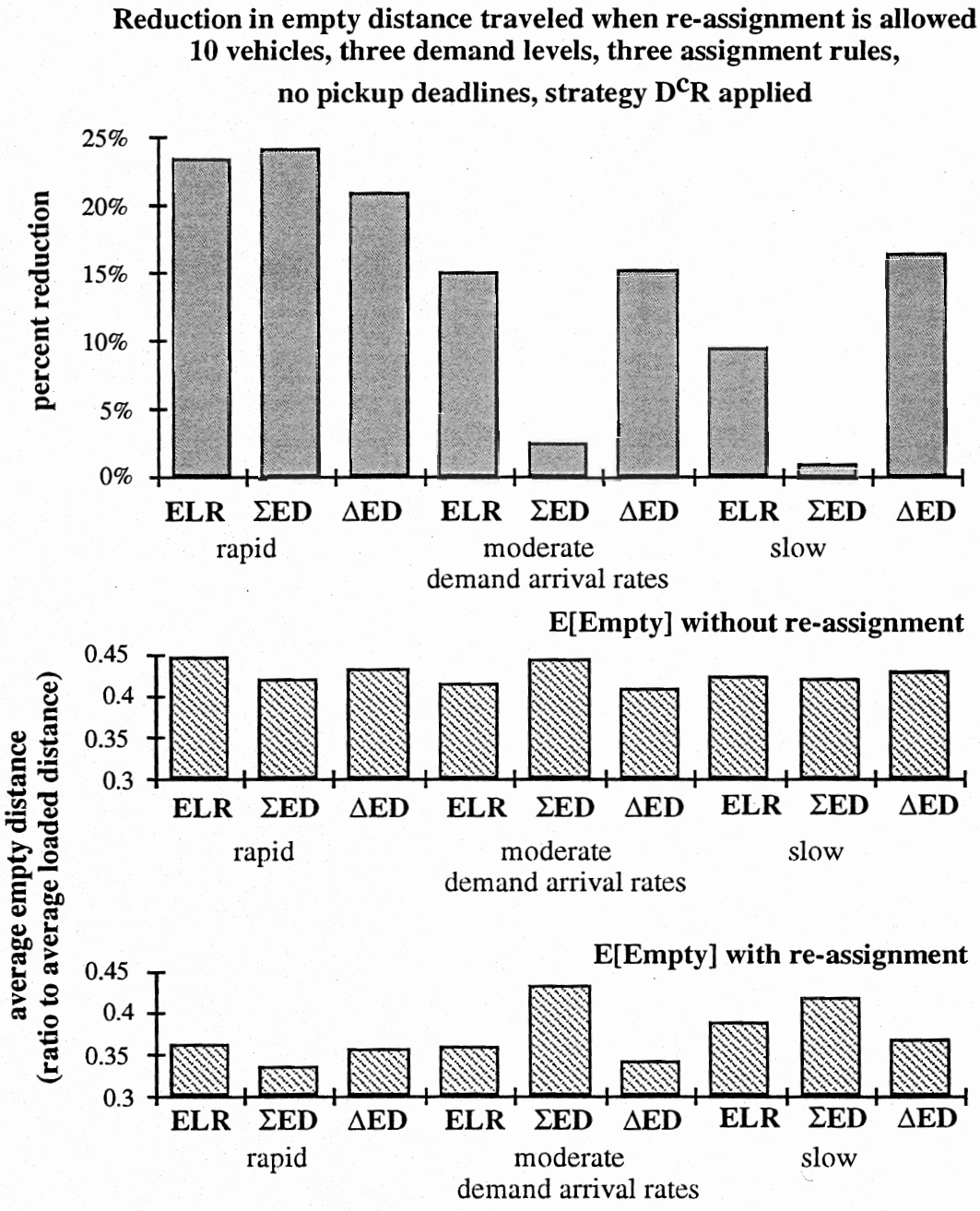


Figure 6.52 Percent reduction in empty distance traveled when re-assignment is allowed and corresponding average empty distance

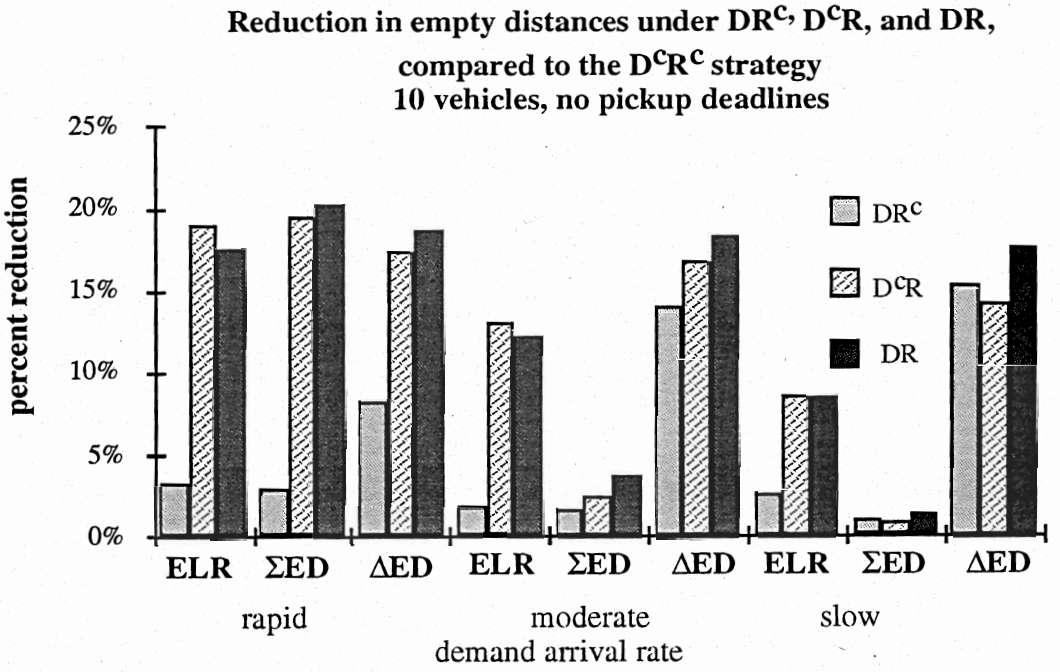


Figure 6.53 Reduction in empty distance traveled under flexible assignment rules compared to the scenario without - no pickup deadlines.

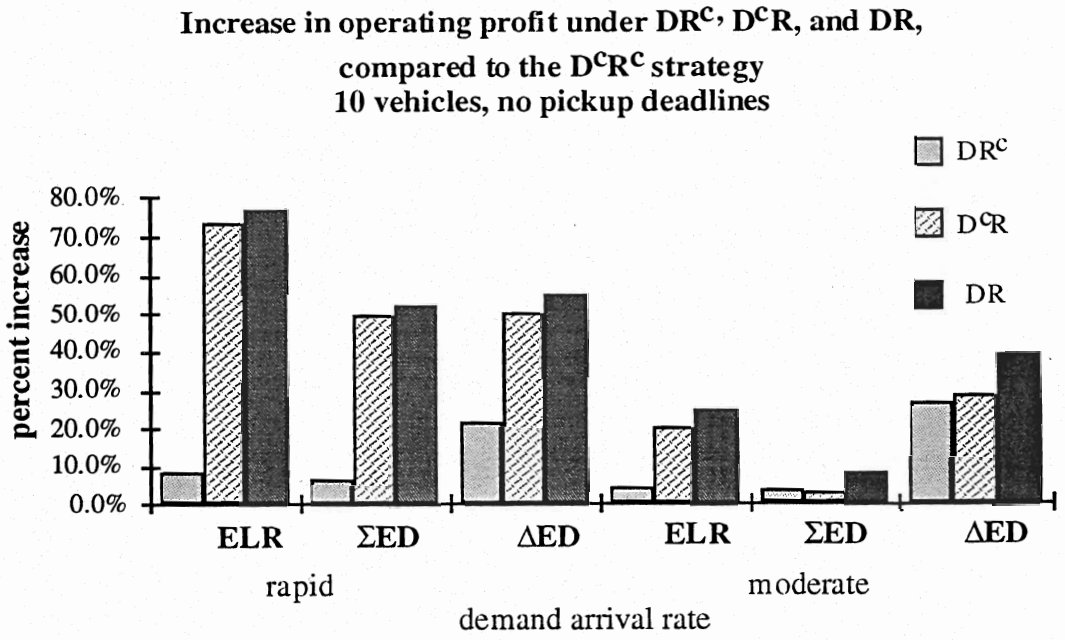


Figure 6.54 Increase in operating profits under flexible assignment rules compared to the scenario without - no pickup deadlines.

Effect of Profit Based Load Acceptance Decisions

When demand is high a fraction of requested loads must be refused service. Rejecting loads on the basis of their expected profitability can lead to significant improvements, both in overall throughput and profitability. Described in chapter 5, the "best" assignment for the load is found using one of the three assignment rules: least empty to loaded ratio assignment, least overall empty distance assignment and least additional empty distance assignment. After this assignment is chosen, the ratio of the additional empty distance attributable to the new load to the loaded distance associated with the load is calculated. If this ratio exceeds a pre-specified threshold the load is rejected. The calculation is performed as soon as the request is received and takes seconds or a fraction of a second to perform. A waiting customer is given a decision immediately. The overall ratio of empty to loaded distances traveled in the cases evaluated varies from 0.08 in highly efficient systems to 0.50 in less efficient systems. In the systems examined, thresholds between 0.5 and 1.2 produce the best results, with lower thresholds rejecting too many loads and higher thresholds leading to the rejection of too few. The choice of the best threshold value varies with the assignment rule applied the demand intensity and differs across systems allow en-route diversion and those that do not.

The empirical analysis presented in this section suggests that the thresholds applied perform better when en-route diversion is allowed, and that the threshold value should be set lower in those cases than when en-route diversion is not allowed. The reason that more restrictive values should be used with en-route diversion and lower values without is that, on average, the number of loads sequenced with a candidate load will be higher under the diversion strategy than under a strategy that does not allow diversion. Under diversion, for empty vehicles, the current first load in the queue is a candidate for re-sequencing. without diversion, it is not. The difference in the average E/L ratio when one more load is sequenced is significant, especially when few loads are sequenced (Regan, Mahmassani & Jaillet [1996] provide a diagram illustrating the difference). The average E/L ratio for longer routes will be lower than the E/L ratio of even slightly shorter routes. Increases in profitability of up to 80% may be attained. Figures 6.55 and 6.56 show this dramatic effect and illustrate the fact that the appropriate choice of such thresholds varies from system to system.

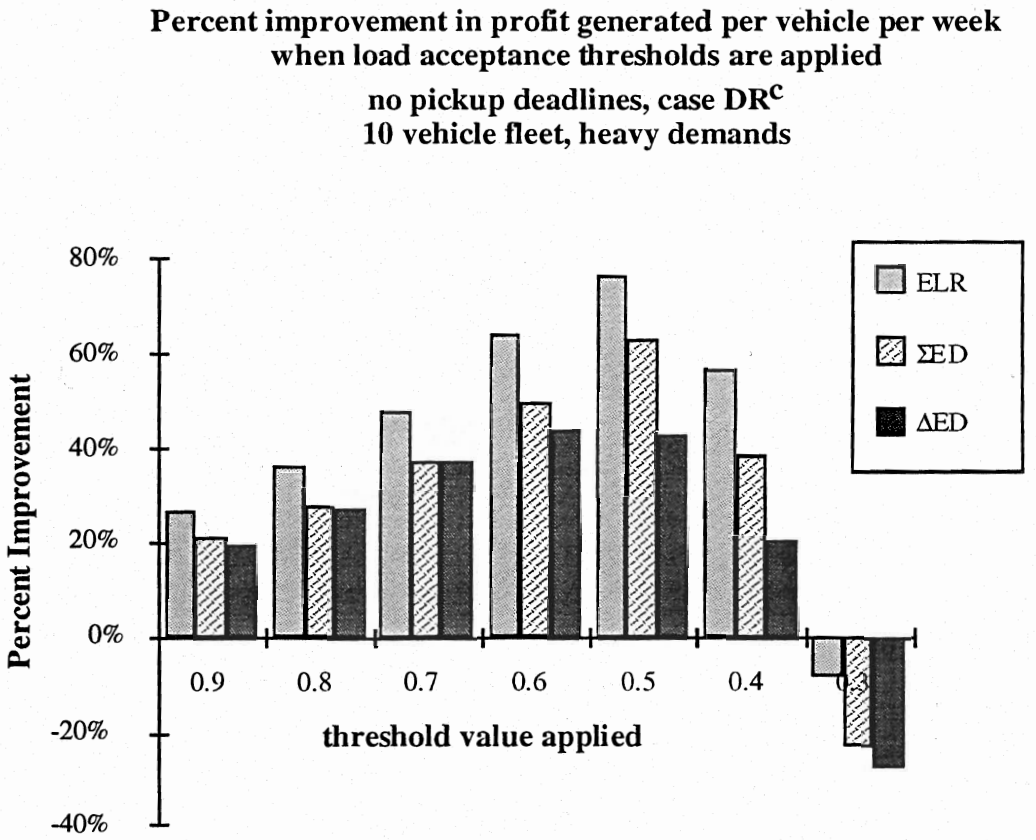


Figure 6.55 Effect of acceptance thresholds on operating profits. En-route diversion allowed.

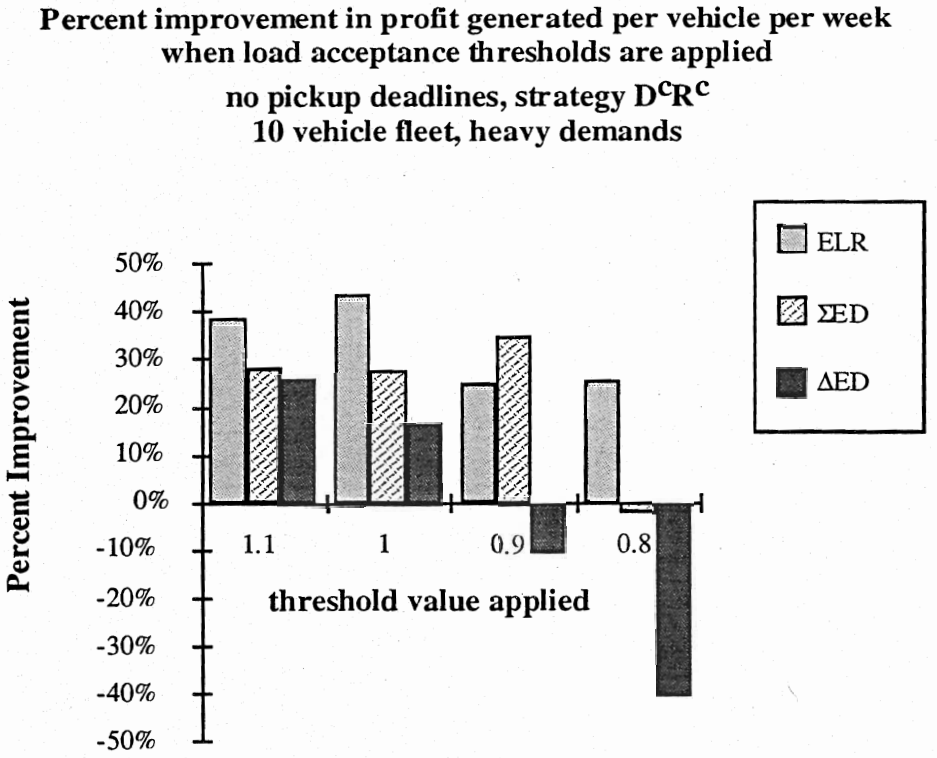


Figure 6.56 Effect of acceptance thresholds on operating profits. No en-route diversion allowed.

Pickup deadline constrained scenarios also benefit from the use of carefully chosen load acceptance thresholds. Under heavy demands, systems under time constraints can benefit significantly from turning away inconvenient (and hence, unprofitable) loads. Figures 6.57 and 6.58 compare the operating profit and empty to loaded ratio for systems of 10 and 20 vehicles in which demands have pickup deadlines drawn from the moderate distribution, and both en-route diversion and re-sequencing of loads are allowed under feasibility only and profit based load acceptance. Under profit based load acceptance the thresholds applied were 0.8, 1.0 and 1.2 for heavy, moderate and low demand levels (rapid, moderate and slow rates of requests for service). As expected, the greatest reduction in E/L and increase in profits are observed in the heavy demand case. Under moderate and low demand applying the threshold rule improves the E/L ratio but does not significantly improve profitability. When premiums are applied for serving loads with deadline constraints, under the threshold values applied, feasibility based load acceptance can in fact perform better. The reason is simple-applying a premium means a fixed charge is earned for serving loads. Some of the loads refused by the profit based acceptance rule, would in fact be profitable. Including the premium in the load acceptance decision would eliminate many refusals of profitable loads.

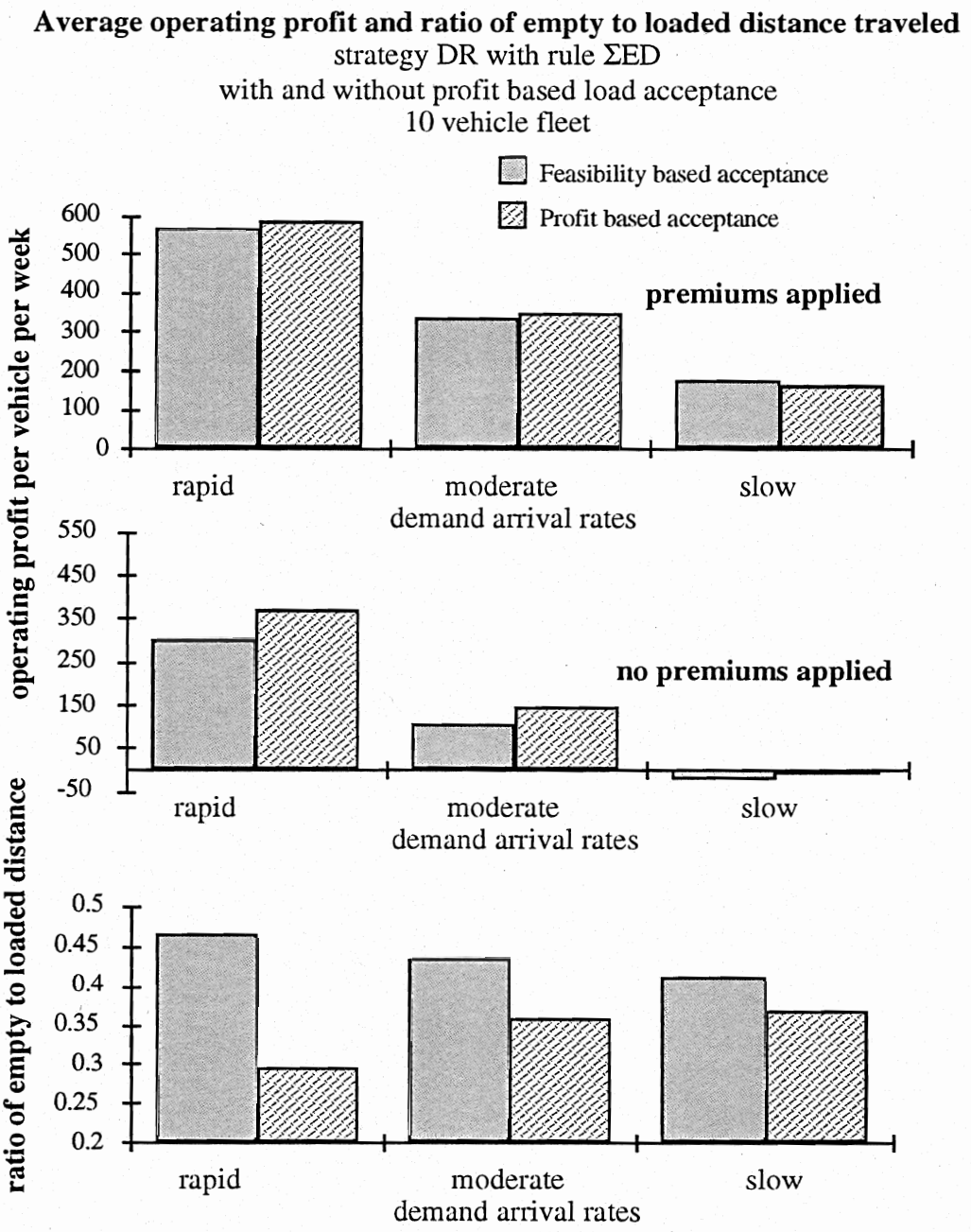


Figure 6.57 Comparison of average operating profit and E/L ratio with and without profit based load acceptance - DR applied with SED, moderate deadlines, 10 vehicles

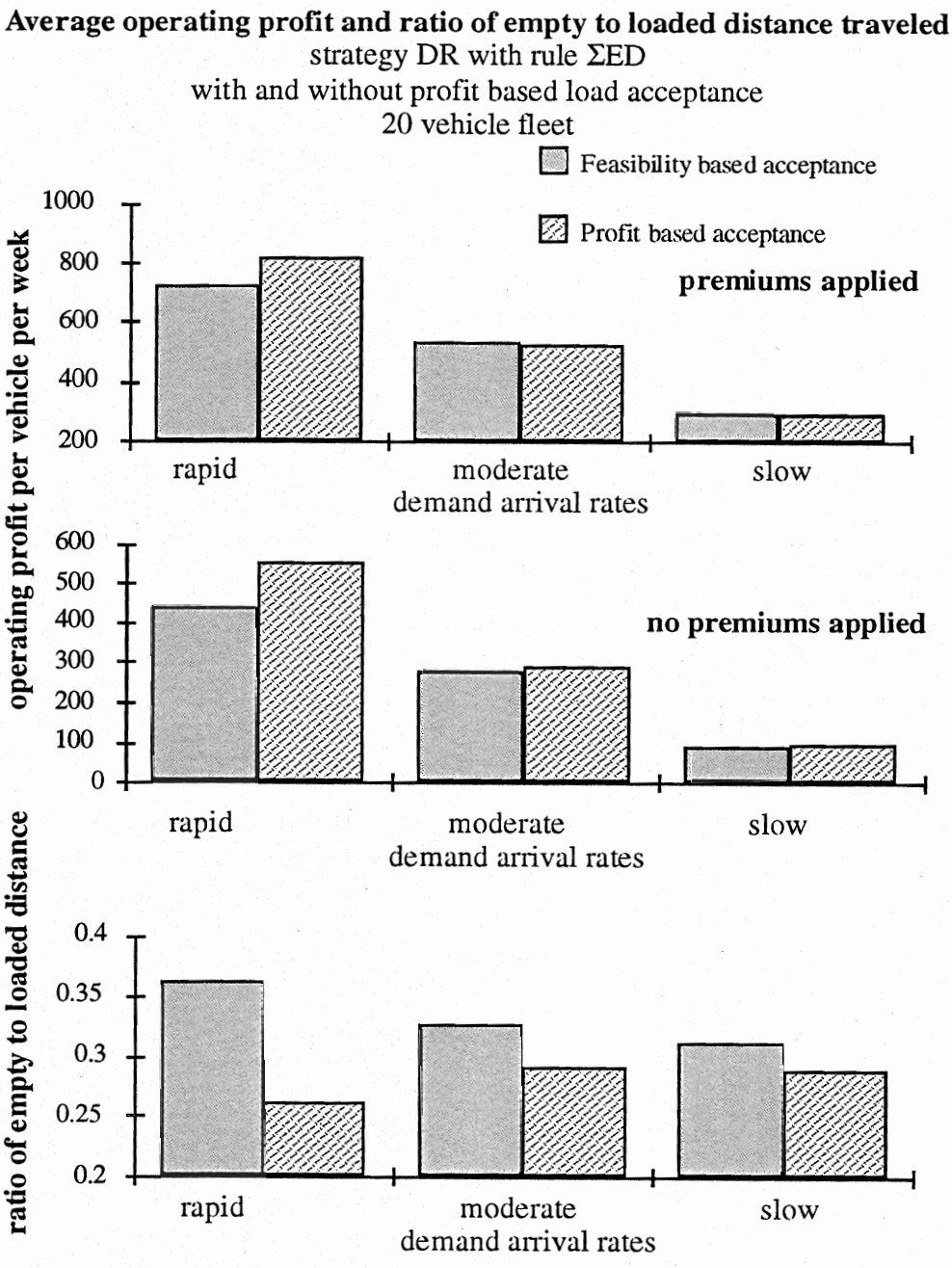


Figure 6.58 Comparison of average operating profit and E/L ratio with and without profit based load acceptance - DR applied with SED, moderate deadlines, 10 vehicles

In addition, with pickup deadlines in place, applying a profit based load acceptance rule can lead to an increase in the fraction of loads with tighter deadlines accepted. In simulation experiments in which deadlines associated with loads were chosen randomly from a distribution of three values, the fraction accepted that were from the tight or moderate categories rose, while the fraction accepted in the loose category fell. This was despite the fact that the profit measure used did not take the pickup deadlines (and resulting profitability under the cost model implemented) into account. The deadlines were not a factor in the acceptance process, the distribution changed because as a result of rejecting less attractive loads the system became more efficient and the tight deadline loads were more likely to be feasible. Figure 6.59 illustrates the difference in the distribution of pickup deadlines associated with accepted loads with profit based load acceptance and without. The distribution of requests is

(tight, medium, loose; $\frac{1}{3},\frac{1}{3},\frac{1}{3}$), corresponding to (4, 8, 12 hours; $\frac{1}{3},\frac{1}{3},\frac{1}{3}$) in the moderate pickup deadline case. It should be noted that this result does not hold in all cases. The choice of the load acceptance threshold makes the difference; the threshold value must be neither to high or too low. Choosing the best threshold requires significant sensitivity analysis.

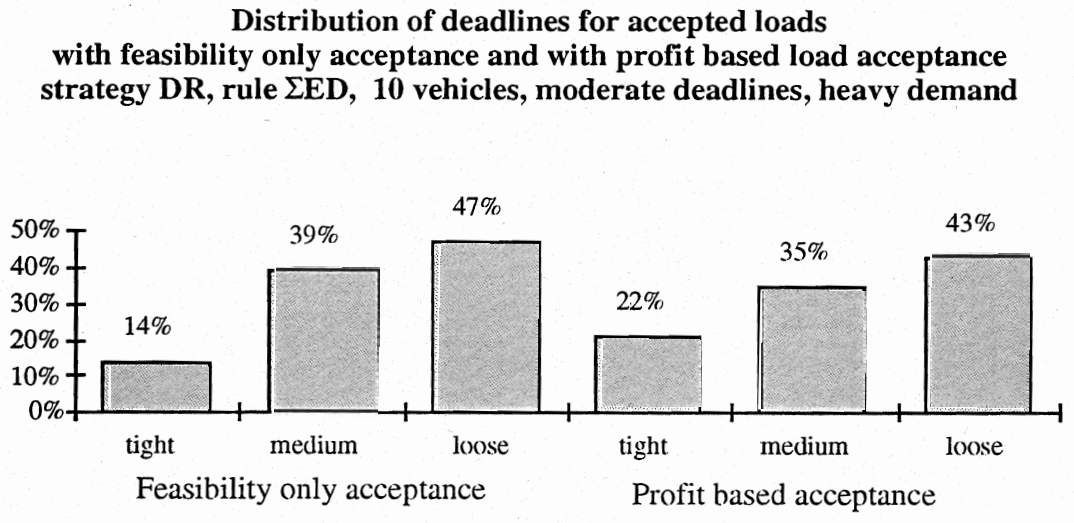


Figure 6.59 Increase in tight deadline loads accepted with profit based load acceptance - 10 vehicle fleet, DR applied with SED, moderate deadlines, heavy demand

Ability to Respond to Pickup Deadlines

Perhaps the most important feature of the real-time operational strategies examined is the ability to take pickup deadlines into account, both in load acceptance and assignment decisions. With pickup deadlines in place, loads that cannot be served within their associated constraints must be turned away. Figure 6.60 compares the fraction of service requests accepted with and without pickup deadlines while figure 6.61 illustrates the effect that the (effective) reduction in demands has on profitability. When premiums are applied in the pickup constrained cases the systems can be more profitable than those without pickup constraints.

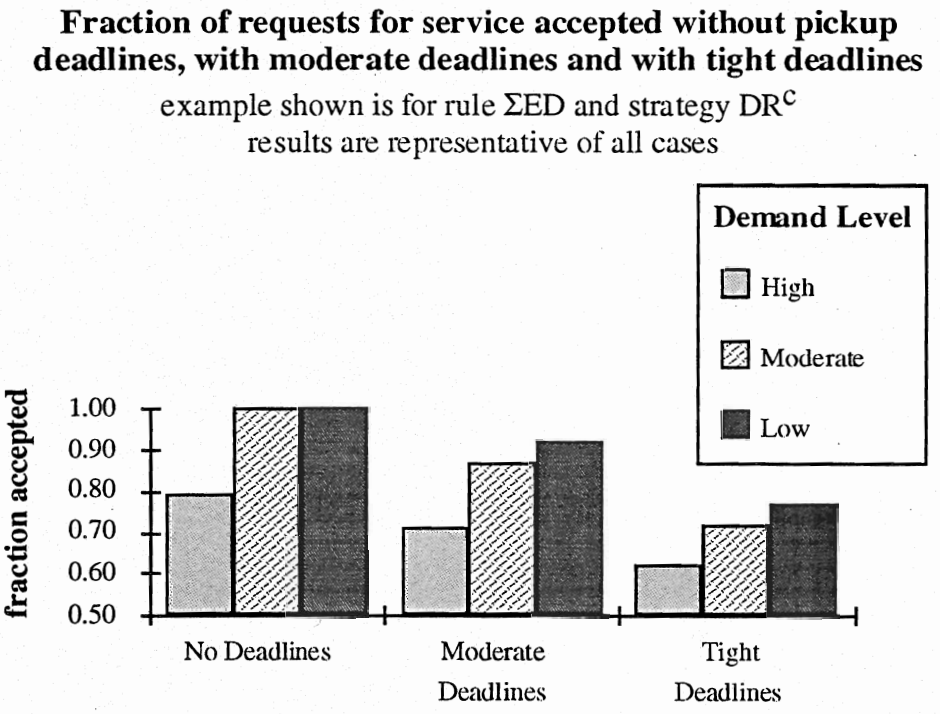


Figure 6.60 Fraction of service requests accepted with and without pickup deadlines

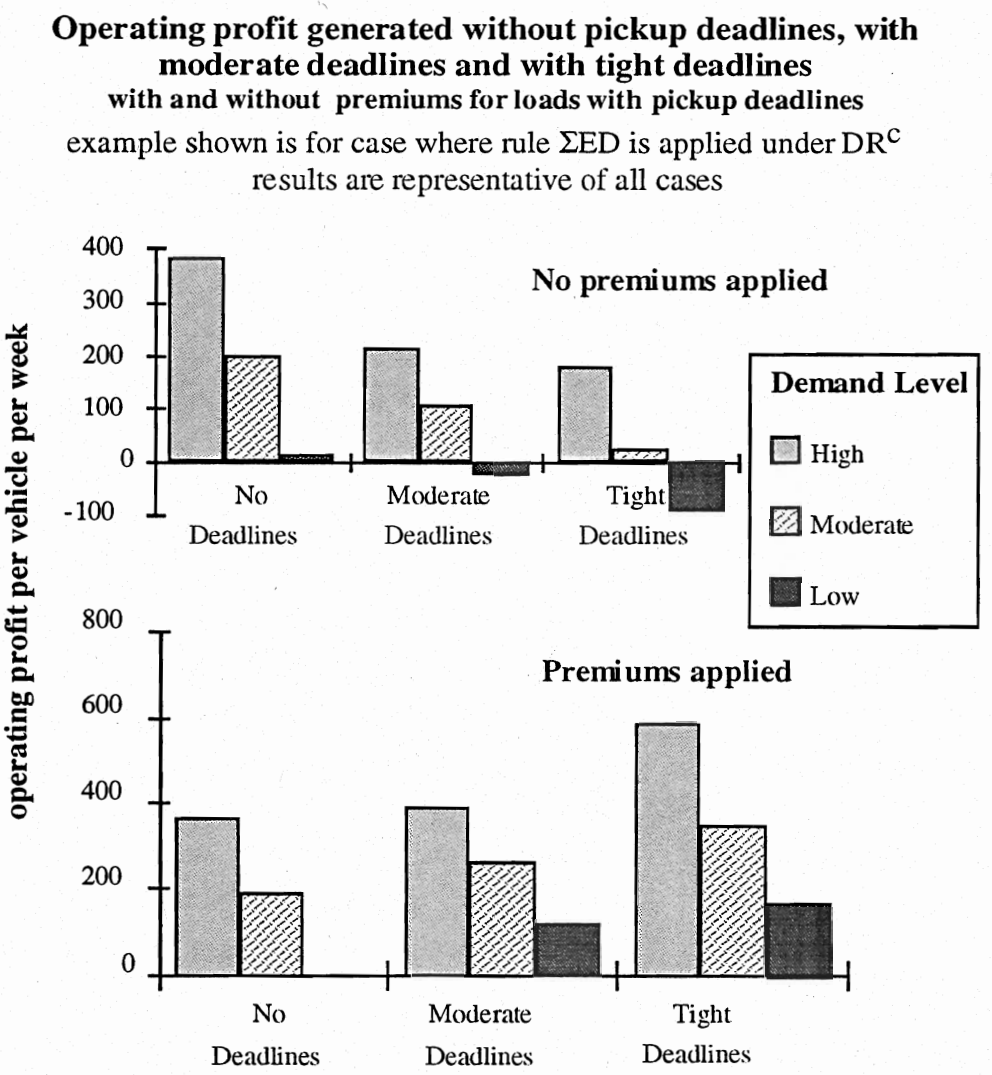


Figure 6.61 Operating profit generated with and without pickup deadlines, with and without premiums earned for responding to deadlines

Performance with Respect to Even Assignments of Loads to Vehicles

A performance measure specified in chapter 5 is the mean, and standard deviation of loaded and empty distances traveled for vehicles across the fleet. The coefficient of variation of empty and loaded distances traveled are examined, and, while the real-time operational strategies examined here do not achieve the nearly perfectly even loading of loads to vehicles that all three of the base cases do, the differences are so small as to be insignificant. Over a 26 week horizon, no matter what the fleet size and with or without pickup deadlines, the maximum coefficient of variation observed for either empty or loaded distances traveled across the fleet is under ten

percent; in most cases it was closer to two-three percent. Without taking that objective explicitly into account, it is satisfied by these assignment rules.

Summary of Real-Time Cases

Three interesting points can be made about the relative performance of the real-time information scenarios examined. The first is that deadline constrained cases do not benefit from flexible assignment strategies as much as might be expected and that in unconstrained scenarios flexible assignment strategies seem to be the key to generating cost effective, customer responsive and profitable assignments. The second point is that although in deadline constrained cases the choice of the assignment rule is clear (SED - least overall empty distance assignment dominates the others), in the unconstrained case the "best" rule varies with the demand level and whether flexible assignments are allowed. Even within a specific assignment strategy, for example one in which en-route diversion and re-assignment of loads is allowed, SED is the preferred choice under high demand but is dominated by DED, least addition empty distance assignment under moderate or low demand when fleet sizes are small. A third observation is that when the criterion for evaluation is the ratio of empty to loaded distances traveled, the assignment rules which minimize the overall empty distance traveled and the additional empty distances traveled perform better, in most every case than the rule which seeks to assign loads to the vehicle with the minimum empty to loaded ratio.

The relative lack of benefits observed in the deadline constrained cases notwithstanding, en-route diversion and re-assignment of loads offer opportunities for improving efficiency. Operations where some loads are deadline constrained and some are not are likely to offer opportunities to benefit from these flexible assignment strategies and at little cost to the system operators and drivers. Since en-route diversion is used little, there is little evidence to support the fear that without additional constraints that a system allowing en-route diversion might divert the same driver over and over again, forcing that driver to incur many more empty moves that loaded ones. The significant benefits observed from the implementation of a very unsophisticated re-assignment rule beg the development of methods which are even more effective. Such methods abound in both the vehicle routing and scheduling literature and are in use in many automated (but generally static) dispatching systems. Finally, particularly in high demand cases, the implementation of simple cost or profit based load acceptance rules lead to significant improvements in system efficiency. The successful implementation of such rules depends upon careful tuning of the rule for the congestion levels observed (and in some cases, desired).

COMPARISON C - COMPARISON OF BASE CASES TO REAL-TIME INFORMATION CASES

Rather than compare all of the base cases with each of the real-time information cases, the real-time operational strategies with the best performance are compared to the best performing base cases. Two cases, one, without pickup deadlines in which strategy D^CR^C is applied with the SED and DED rules, and another, with moderate pickup deadlines, in which DR is applied with SED and profit based load acceptance, are compared to the base cases NO, BAT(a), BAT(a - with half look ahead) and BAS(b). Results vary over fleet sizes so the comparison is presented for the ten, twenty and fifty vehicle fleets.

No Pickup Deadlines

At high demand, with no pickup deadlines, the nearest origin assignment method significantly outperforms any of the "real-time" assignment rules examined. The quasi real-time base case strategies, BAT(a - with look ahead) and BAS(b) also perform well. Figure 6.62 provides diagram showing the relative performance of these assignment strategies with respect to empty distance traveled, average and standard deviation of wait time and operating profitability. Results are shown for ten and twenty vehicle fleets. Results for fleets of fifty vehicles are not shown because the computation complexity of the fifty vehicle, high demand scenario proved too high to simulate over a sufficiently large number of iterations. Demands so high that the system is always working at capacity are unusual. Under moderate and low demands the D^CR^C strategy, the least intelligent of the real-time assignment strategies, is competitive with the base cases, including the quasi real-time base cases. Figures 6.63 and 6.64 display corresponding results under moderate and low demand.

Relative Performance with Pickup Constraints

When pickup deadlines are respected, under the real-time strategies a significant fraction of requests must be turned away. As a result, under high demand, the real-time strategies generate less operating profits, with or without the selected premiums meeting pickup deadlines. Under moderate or low demand these strategies generate higher profits than the base cases - even the quasi real-time assignment cases.

In addition, as defined, the base case assignment strategies are incapable of satisfying pickup deadlines. To construct a case for comparison. Loads are assigned pickup deadlines, considered soft constraints, from the moderate pickup deadline distribution examined in the real-time assignment strategies. In these cases, roughly half of the loads are not served within their

deadlines and on average the deadlines are missed by more than twice the duration of the average loaded move.

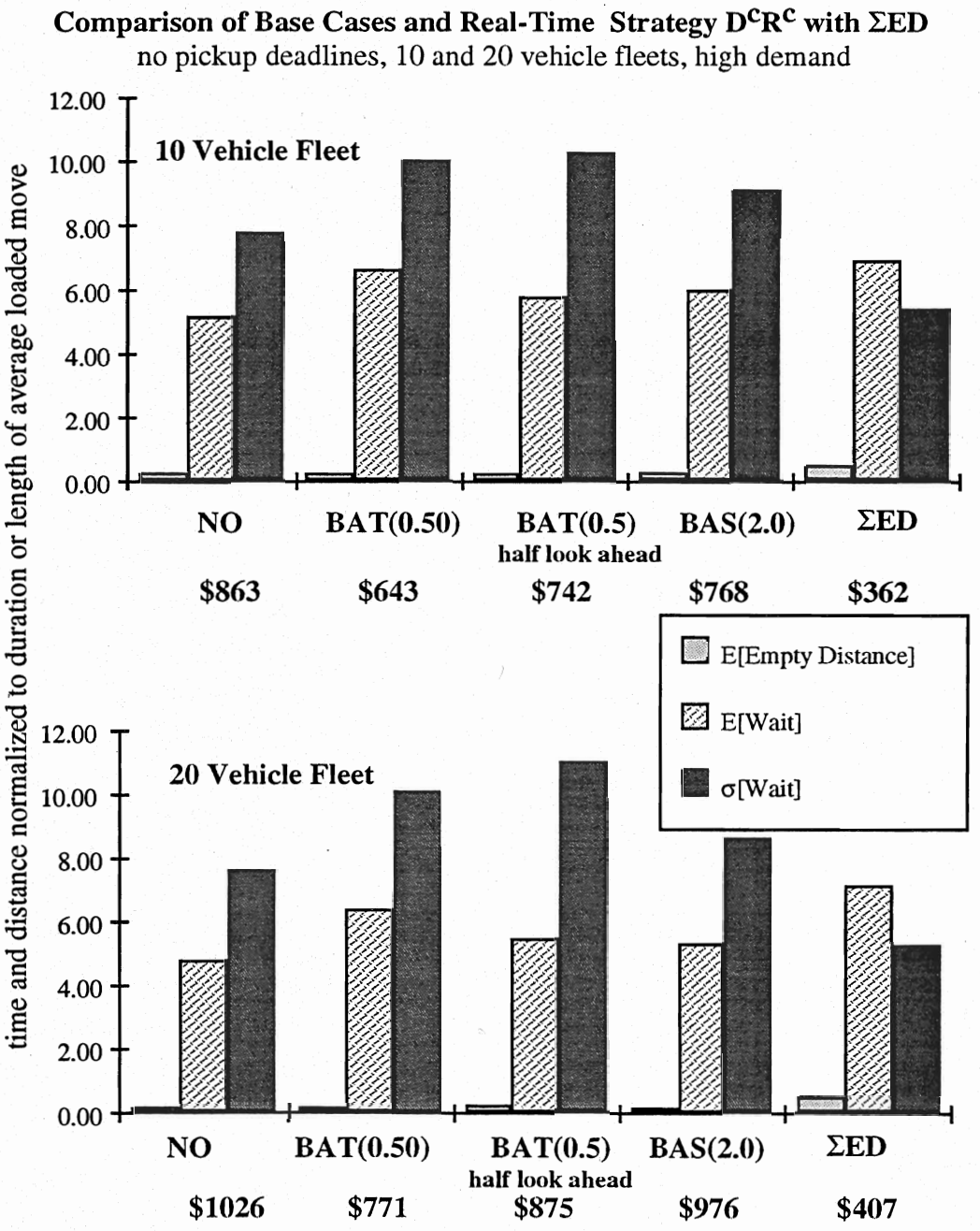


Figure 6.62 Relative performance of the base cases and D^CR^C with SED under high demand

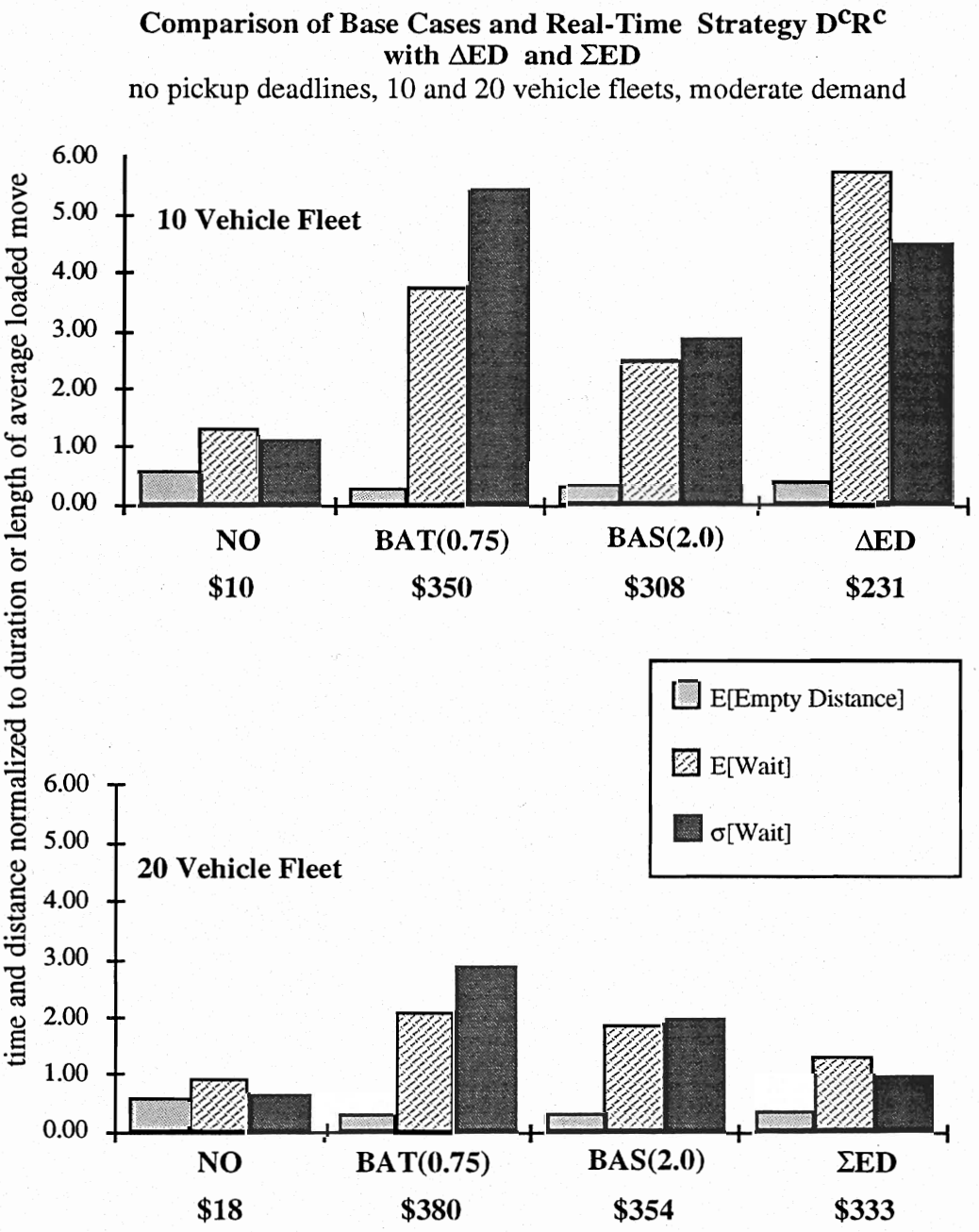


Figure 6.63 Relative performance of the base cases and D^CR^C with SED under moderate demand

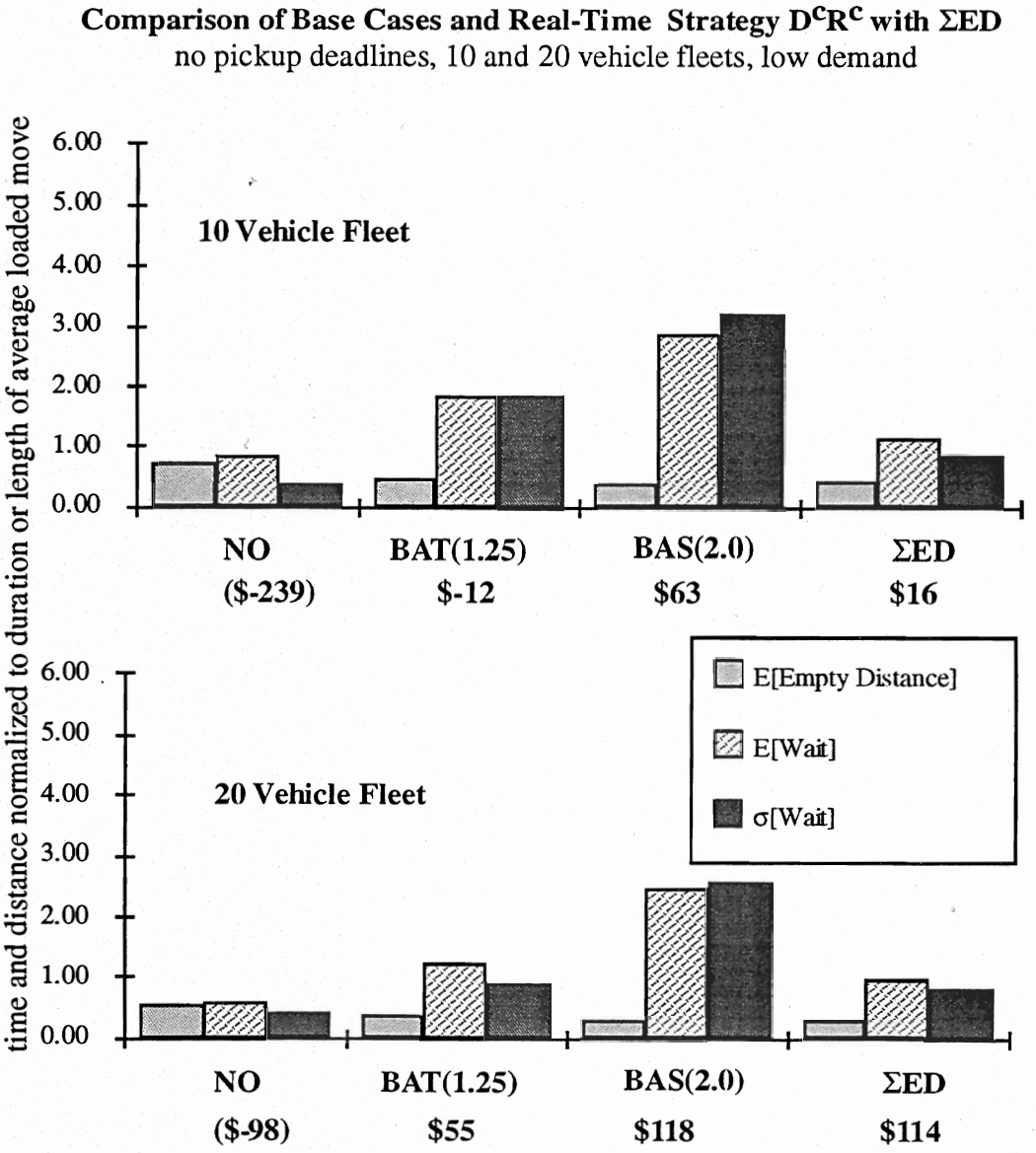


Figure 6.64 Relative performance of the base cases and D^CR^C with SED under moderate demand

Figure 6.65 illustrates the relative performance of the NO, BAT(0.5), BAT(0.5) with half look ahead, BAS(2.0), and DR assignment strategies under high demands. While the operating profit in the DR assignment strategy is not competitive with the others, the wait time and variability of wait time for service is nearly an order of magnitude lower. Operating profits are shown with and without premiums applied for deadline compliance. The differences in profitability under high demands are less for larger fleets, as are the magnitude of the differences in wait time. Under

moderate and low demands (figures 6.66 and 6.67), and for larger fleet sizes, the real-time assignment strategy (which includes profit-based load acceptance) is competitive, even without premiums applied. For an operation designed to provide service to time sensitive customers, the real-time assignment strategy is preferable to the base cases.

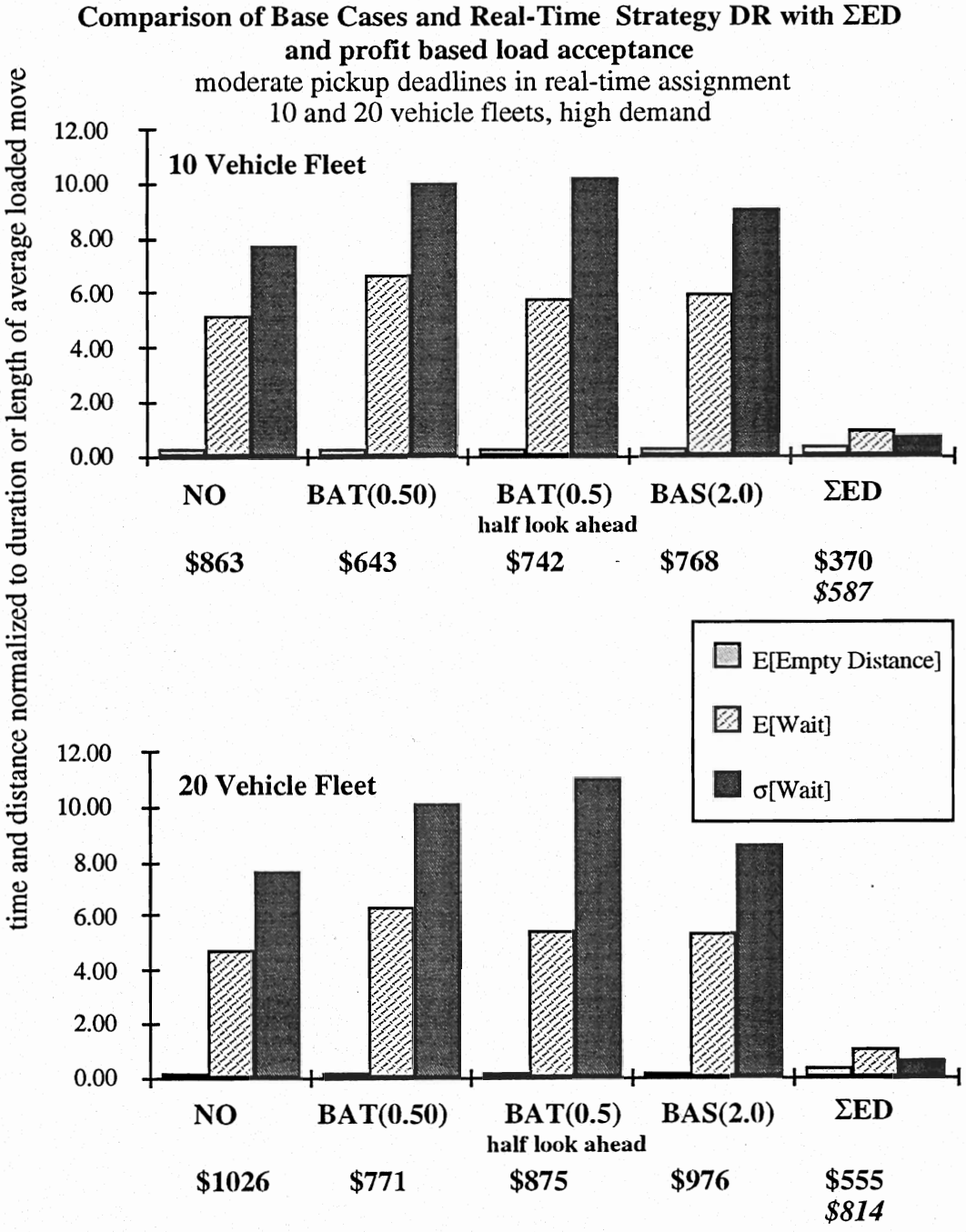


Figure 6.65 Relative performance of the base cases and DR with SED under high demand

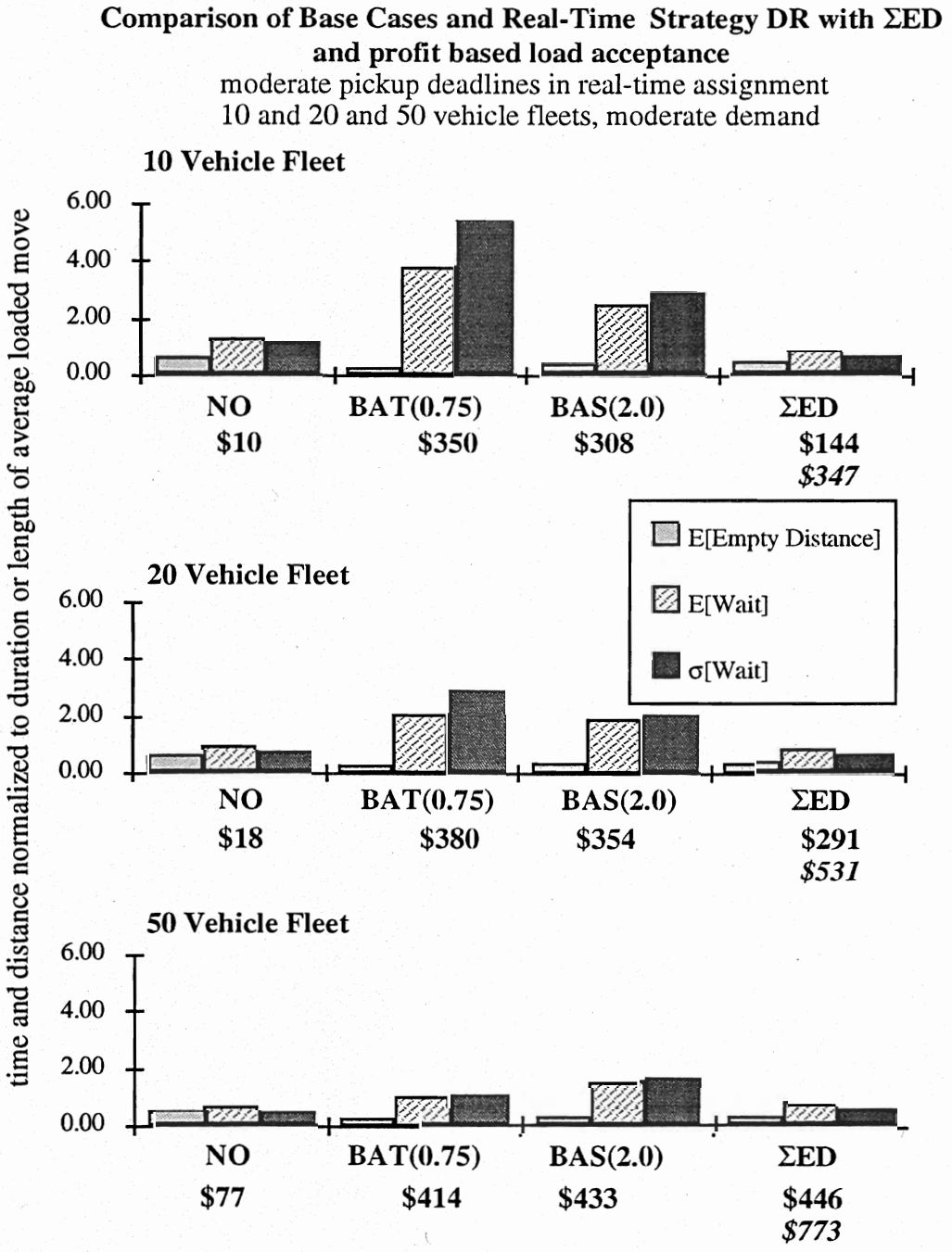


Figure 6.66 Relative performance of the base cases and DR with SED under moderate demand

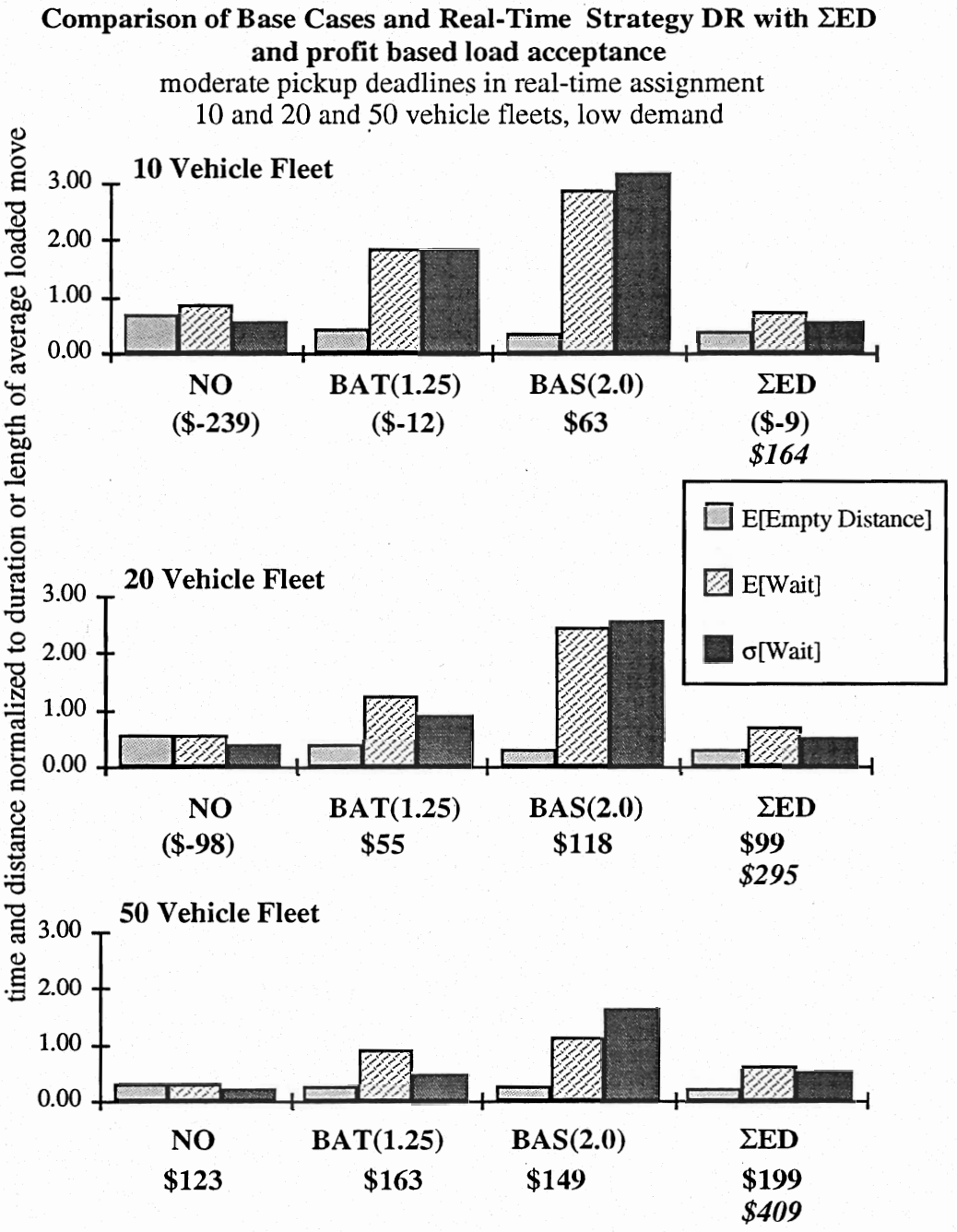


Figure 6.67 Relative performance of the base cases and DR with SED under low demand

COMPARISON D - COMPARISON OF REAL-TIME ASSIGNMENT STRATEGY TO SOLUTIONS CORRESPONDING TO ASYMMETRIC TSP PROBLEMS

In this case the measure of interest is the average empty distance driven to provide service. The average number of loads served per week is 8.08 in the real-time information case. With only one vehicle, the local assignment rules do not have an effect on the solution. With no pickup deadlines, loads are accepted into the system if there are less than five loads already awaiting service. This constraint is rarely binding under moderate arrivals. The ratio of the average empty distance traveled to provide service in the real-time solution to the ATSP solution in which eight loads are ordered and serviced each week is approximately 0.69 to 0.43. The real-time solution is 60% less efficient than the ATSP solution.

SUMMARY

The real-time assignment methods examined offer opportunities to provide service to time sensitive customers. Pickup deadlines can be met within the framework of a profitable operation. For larger fleets, the performance of the best real-time assignment strategy results in higher operating profits than even the most profitable "base" case -- and meets customer service needs.

If customers can be persuaded to pay a premium for fast and reliable service then opportunities for improving profitability are even more significant. Even without extracting premium prices, companies offering guaranteed pickups within requested deadlines might attract more customers and improve their profitability by improving utilization of the fleet. The simulation results presented show that using these flexible assignment methods, a company can provide much better service to customers and at the same time remain nearly as profitable in most cases examined and more profitable in some cases as those providing significantly less responsive service.

The flexible assignment strategies, en-route diversion and re-assignment of loads have demonstrated promise when applied in no-deadline, mixed deadline or loose deadline environments. The re-assignment method implemented in this study is clearly sub-optimal. Its impressive performance encourages the development of more intelligent re-assignment methods and those suited to an environment in which pickup deadlines are binding. A simple load swapping rule which identifies loads with similar deadlines and looks for opportunities to re-assign them to different vehicles would be the obvious choice.

In unconstrained scenarios, the performance of the local assignment rules applied under the real-time operational strategies are highly sensitive to the congestion level of the system and individual vehicle queue limits imposed. Results suggest that a hybrid system, which would

choose the assignment rule based on the current congestion level of the system might result in increased efficiency, when compared to the application of any one of the assignment rules.

The final chapter of this dissertation presents a summary of findings, and makes recommendations for future research.

Chapter 7 Conclusions

SUMMARY OF FINDINGS

The primary objectives and hypotheses examined in this research were outlined in chapter one. Each of these is reviewed and related conclusions and findings from the examination is discussed.

The first two objectives were to state, formulate and analyze the driver assignment (or dynamic vehicle allocation and routing) problem in a way that explicitly takes real-time information on vehicle locations and demands into account, to develop operations research methodologies to assist with dispatching, load acceptance, and dynamic pricing strategies and to test these methodologies under the assumption of the availability of real-time information on vehicle locations and demands.

A set of four real-time assignment methods were formulated and examined, along with three lower level decision rules. These four are all intended for use in operations requiring demand responsive service to customers with explicit service deadlines.

Specified in the third stated objective of this research, a simulation framework has been developed to analyze carrier fleet operations under real-time information and to evaluate the effectiveness of strategies developed. Results of the evaluation of these real-time assignment heuristics demonstrate the benefits associated with flexible dispatching strategies which require continuous updates on the current status of customers and vehicles and which benefit from continuous communication between a central authority, the dispatch center, and the same. Dispatching strategies which explicitly honor customer pickup deadlines are shown to be competitive with base case assignment strategies which cannot honor such deadlines and in which wait times for service can be unreasonable. The real-time assignment strategies are competitive, with respect to the criterion of operating profits earned, with a quasi real-time, state-based bipartite assignment method, the most effective of the less information intensive strategies examined. In addition, when premiums are awarded in pickup constrained cases the real-time assignment strategies significantly out-perform the base cases; as defined, the base case strategies can offer no such guarantee that deadlines will be met.

Meeting objective four, to provide quantitative estimates of the benefits of real-time information for vehicle assignment and routing decisions for trucking operations poses significant difficulties. A recent study conducted by the American Trucking Association Foundation [1996, p. 22] estimates the benefit/cost ratio of mobile communication systems alone at between 1.5:1 to

5.0:1. The study does not quantify the benefits of computer aided dispatching and routing, but mentions that 14, 46 and 74 percent of small, medium and large trucking companies, respectively, have invested in such systems and that the benefits appear to far outweigh the costs. The kinds of dispatching heuristics developed in this research should lead to even greater benefits.

Analysis shows that the performance of the real-time information assignment heuristics varies significantly over demand levels and the distribution of pickup deadlines. What is clear is that the wait time for service can be reduced by more than 2-5 times when loads are assigned deadlines for pickup over the most efficient corresponding cases which do not rely on real-time information. This is a significant benefit to both customers and companies, who benefit from improved customer relations. While under extremely high demand the simplest of the base cases performs better than the real-time operational strategies, under more reasonable demand levels the real-time strategies can be as efficient and in some cases slightly more efficient than the base cases, with respect to the cost to provide service.

The first hypothesis of this research, namely that real-time information on vehicle locations and demands can increase the efficiency of carrier fleet operations with respect to measures of trucking company profitability and responsiveness to customer requests or desires has certainly been demonstrated, particularly with respect to responsiveness to customer requests or desires. The second, that real-time assignment rules perform well, with respect to those requiring less real-time information, under certain conditions with respect to fleet size, level of demand and pickup deadlines has also been demonstrated. The conditions under which they perform well have been identified as: fairly large fleet sizes, moderate to low demands and time-constrained pickups.

RECOMMENDATIONS FOR FUTURE RESEARCH

The key issues and recommendations for further research are presented here. The first two have broad applicability to fleet operations management, while the rest are direct at extensions of the current work.

•Extension of observations and dispatching strategies to related real-time fleet management systems

Less than truckload operations and truckload operations providing service to intermodal terminals (including, rail, maritime and air terminals), are of particular interest. Assignment strategies developed may also be directly applicable to other dynamic fleet management systems (for example, taxi fleets, dial-a-ride and local courier operations).

•Development of approximations for performance measures in distributed queueing systems

Such approximations, coupled with tighter bounds on key performance measures (average wait time for service, for example) would be very useful to both long term, and day-to-day fleet operations management.

•Identification of efficient re-assignment (load swapping) heuristics

Many of the route improvement heuristics developed to solve multiple traveling salesperson or vehicle routing problems could be applied directly. Some obvious choices are 2-opt, 3-opt or k-opt load swapping in which likely candidates for swapping are identified by the similarity of their associated time windows.

•Development and implementation of hybrid assignment strategies, combining the strengths of the best performing strategies identified so far

The assignment strategy formulated in chapter 3 in which costs in the bipartite assignment step represent the cost associated with the optimal solution of a TSPWT sub-problem has significant promise.

•Further examination of the extent to which congestion of demand and vehicle locations determines the effectiveness of the assignment strategies described in this research

Of particular interest is the robustness of local assignment strategies under more varied conditions with respect to the dispersion of locations of customers and demands. Performance when pickups and deliveries are clustered, both geographically and temporally should be examined.

•Tests of identified dispatching heuristics over a more natural geographic region and across a set of known demands or demands generated from a model which includes realistic origin-destination locations and demand levels

Implementing these models in a geographic information system is a natural extension. Regional or national freight demand data could be used to build a model to a more realistic demand arrival pattern.

•Extension of pickup deadlines to time constraints which include time windows for pickup and delivery

•Examination of the effects of queue and pool limits under heavy demand

In particular, the extent to which limits chosen effect the relative performance of assignment strategies and local assignment rules should be investigated.

•The development of cost models that represent industry practice

Customer willingness to pay a premium for timely service should be investigated.

Appendix I Simulation Details

In this section the details of the simulation framework and of the implementation of assignment strategies in a simulation framework are discussed.

Three separate simulation programs are described. These include: the simulation of the nearest origin assignment strategy, of which first called first served is a special case; simulation of the bipartite assignment strategy; and, simulation of the real-time assignment strategies.

The simulation is deterministic, with stochastic inputs. It is event based. The simulation was developed in C. Experiments were conducted on a DEC Alpha workstation. Three separate simulation programs are used to conduct the experiments discussed in this chapter. Pseudo code is included for each of these programs. The full code for the Hungarian Assignment Algorithm is included (Supplied by Kishore Sarathy, former Ph.D. Candidate, University of Texas, Department of Operations Research).

A.1.1 COMMON FEATURES OF SIMULATION PROGRAMS

1) All simulations begin at rest, with all vehicles at the center of the circular work area.
2) Vehicles travel around circular work area at a speed of one unit of distance per unit of time.
3) Demands arrive according to a Poisson Distribution. The time between arrivals is a parameter input to the simulation.
4) The number of vehicles is a parameter to the system.

Recipes in C Example book, Flannery, Teukolsky and Vettering[1988])

A.1.2 SIMULATION OF NEAREST ORIGIN ASSIGNMENT ALGORITHM (INCLUDES FCFS ASSIGNMENT)

```
if FCFS maxToSearch = 1;   /* loads are kept in the pool in the order of arrival to the system */
else maxToSearch = maxPL;

/************ Begin Simulation ******************/

    iteration = 0;
    while (iteration < MaxIterations and ConvergenceTest1 == FALSE)
    {
        iteration = iteration+1;
        function intialize variables();

        while (nextEventTime < EndOfHorizon)
        {

            for (all vehicles)
            {
                while (loadsInPool > 0)
                {
                    if (currentVehicle is Idle)
                    {
                        function selectClosestLoadinFirst
                            MaxToSearchSlotsInPool();
                        assign LoadSelected to currentVehicle;
                        remove LoadSelected from Pool;
                    }
                }
            }

            function getNextEvent();

             while(nextEventType == PICKUP and nextEventTime < EndOfHorizon)
             {
                elapsedTime = nextEventTime;
              function UpdateVehicleInvolvedInNextEvent();

            function getNextEvent();
                }

            if (nextEventType == DELIVERY && nextEventTime < EndOfHorizon)
                {
                function selectClosestLoadinFirst
                        MaxToSearchSlotsInPool();
                assign LoadSelected to currentVehicle;
                remove LoadSelected from Pool;
              }
```

```
        if (nextEventType == ServiceRequests && nextEventTime < EndOfHorizon)
        {
                elapsedTime = nextEventTime;
                for all vehicles
                {
                        function undateVehiclePosition();

                        if (ConvergenceTest2 == TRUE)
                        {
                                check and record number of customers in the Pool
                                and in the system;
                        }

            if (loadsInPool < maximumAllowableLoads)
            {
                function sendLoadToPool();

                 if (loadsInPool == 1 && idleVehicles)
                /* load just sent to empty pool  - assign to nearest idle vehicle */
                {
                if (NO)
                {
                        function selectClosestIdleVehicle();
                }
                elseif (FCFS)
                {
                        function selectLongestIdleVehicle();
                }
                assign to Vehicle Selected;
                remove Load From Pool;
                }
             }
               else rejectRequestForService;
               function getNextEvent();
        }

    if (nextEventTime > EndOfHorizon)
            {
                    elapsedTime == EndOfHorizon;
                    {
                    for all vehicles in fleet
                            function undateVehiclePosition();
                    }
             }
      }              /*** END OF ITERATION ***/
                      function sumResultsOfIteration();



}       function printResultsOfSimulation();
        /*** END OF SIMULATION ***/
```

A.1.3 SIMULATION OF CLASSICAL ASSIGNMENT ALGORITHM

```
/*** BEGIN SIMULATION OF SYSTEM WITH ASSIGNMENT ALGORITHM ***/
iteration = 0;
{
While (iteration < MaxIterations and ConvergenceTest1 == False)
{
        iteration = iteration +1;
        function initialize variables();
        While (nextEventTime < EndOfHorizon)
        {
                function getNextEvent();
        /**** next event is a change in vehicle status ****/
While (nextEventType != ServiceRequest && TimeToPerformNextAssignment < elapsedTime &&
nextEventTime < EndOfHorizon)
                {
                        if (ConvergenceTest2 == TRUE)
                        {
                                check and record number of customers in queue and in system;
                        }

                        elapsedTime = nextEventTime;

                        function updateVehicleInvolvedInNextEvent();
                                /*** update position  and status of vehicle involved in the next
                                        event (a change of status) ***/
                        if (NextEvent is change from loaded to idle state)
                        {
                        if (STATE_BASED ASSIGNMENT AND idle_vehicles < bPL)
                        TimeToPerformNextAssignment = elapsedTime;
                        }
                        fuction getNextEvent();
                }
                        elapsedTime = nextEventTime;

                        /**** NEXT EVENT IS A REQUEST FOR SERVICE ****/

                if (PL > maxPL) /* loads in pool exceeds allowable number */
                {
                        rejectLoad;
                }

                elseif (elapsedTime < EndOfHorizon)
                {
                        function sendLoadToPool();
                        if (STATE_BASED ASSIGNMENT AND PL > b(idle_vehicles))
                                TimeToPerformNextAssignment = elapsedTime;
                        }
```

```
                for all vehicles in fleet
                {
                        function updateVehiclePosition();
                  }

                if (elapsedTime > TimeToPerformNextAssignment)
                {
                if (idle or nearIdle vehicles > 0 && PL > 0) /* if vehicles and loads to assign */
                {
                        function assignLoads();
                        update TimeToPerformNextAssignment;
                }
                if (nextEventTime > EndOfHorizon)
                {
                        elapsedTime = EndOfHorizon;
                        {
                        for all vehicles in fleet
                                function undateVehiclePosition();
                        }
                }

                function sumResultsOfIteration();
                iteration = iteration + 1;

            }/*** END OF ITERATION ***/
function printResultsofSimulation();
}/*** END OF SIMULATION ***/
/*************************************************************************

function assignLoads() -- calls bipartite assignment code
*************************************************************************/
        function assignLoads()
        {
                n = MAX(numberOfLoadsInPool, NumberOfIdleVehicles)

                Allocate (CostMatrix,VehicleArray, LoadsArray, Uvector,Vvector,Xvector);

                Populate CostMatrix;    /*** distances from vehicle location for idle vehicles,
                                                next idle location for busy vehicles ***/

                Populate VehicleArray, LoadsArray; /*** for vehicles,
                                                        loads underconsideration ***/

                function ass();  /*** Perform Hungarian Assignment ***/

                for (i = 0; i < n; i++)  /*** If loads considered exceed vehicles, no load can be
                                                assigned to vehicle[i] ***/
                {
                        if (vehicle[i] has been assigned a load)
                        {
                                currentLoad = loadsArray[x[i]];
```

```
                                    function AssignLoad();  /*** Assign CurrentLoad to
                                                            Vehicle i ***/
                                    remove currentLoadFromPool;
                                }
                            }
                        Free allocation space;
                        }
/*************************************************************************/

/*************************************************************************
File        : assign.c
Author      : Kishore Sarathy
Project     : Vehicle Routing Problem
Topic       : Assignment problem

Description : This file contains a program to solve an assignment problem
              using the Hungarian method. This implementation is for dense
              graphs and expects the costs or distances to be specified as
              a matrix. The function assign is to be called to obtain
              solution for the assignment problem.

Input       : Distance matrix, addresses of arrays for dual variables.
              NOTE: Allocation of space for these vectors is done inside
              this function.

Output      : Reduced cost matrix, (the cost matrix is modified directly)
              vectors of dual variables.

WARNING     : Cost matrix is modified inside this routine. Send a copy if
              want it preserved.
*************************************************************************/

#include <stdio.h>
#include "myStdefs.h"   /* include the standard definitions */
#define  ASSFILE        /* define to include the definitions of variables */
                        /* from header file */
#include "ass.h"        /* include header file for ass.c */
#include "errmsg.h"
#include "nrutil.h"

/*-------------------------------------------------------------------
  init -  initializes variables and arrays for assignment algorithm
-------------------------------------------------------------------*/
int   init (int **c, int *U, int *V, int n)
{
  int   i, j,
        mincol,
        minrow;

  Optimal = FALSE;

  cLbls = ivector(0,n);
  rLbls = ivector(0,n);
  rMatch = ivector(0,n);
  cMatch = ivector(0,n);
```

```
  for (i = 0; i < n; i++){
    cLbls[i] = rLbls[i] = UNLABLD;
    rMatch[i] = cMatch[i] = UNMATCHD;
    U[i] = V[i] = INF;
  }

  for (i = 0; i < n; i++) {
    for (j = 0; j < n; j++)
      if (U[i] > c[i][j]) {
        U[i] = c[i][j];
        mincol = j;
      }
    if ((rMatch[i] == UNMATCHD) && (cMatch[mincol] == UNMATCHD)) {
      rMatch[i] = mincol;
      cMatch[mincol] = i;
    }
    for (j =0; j < n; j++)
      if (c[i][j] != INF)
        c[i][j] -= U[i];
  }

  for (j = 0; j < n; j++) {
    for (i = 0; i < n; i++)
      if (V[j] > c[i][j]) {
        V[j] = c[i][j];
        minrow = i;
      }
    if ((cMatch[j] == UNMATCHD) && (rMatch[minrow] == UNMATCHD)) {
      cMatch[j] = minrow;
      rMatch[minrow] = j;
    }
    for (i = 0; i < n; i++)
      if (c[i][j] != INF)
        c[i][j] -= V[j];
  }
}

/*-------------------------------------------------------------------
  improve - improves the current assignment
-------------------------------------------------------------------*/
int   improve(int **c, int *u, int *v, int n)
{
  int i, j, root, FoundPth = FALSE, tnode;

  for (i = 0; i < n; i++)
    rLbls[i] = cLbls[i] = UNLABLD;

  for (root = 0; (rMatch[root] != UNMATCHD)&&(root < n); root++)
    ;
  if (root == n)
    Optimal = TRUE;
  else {
    augPath(c, n, root, &tnode, &FoundPth);
    if (FoundPth)
      swapMatchng(root, tnode);
```

```
      else {
        update(c, u, v, n);
          }

   }

}

/*-------------------------------------------------------------------
  augPath -  Finds an augmenting path
-------------------------------------------------------------------*/
int   augPath(int **c, int n, int start, int *tail, int *found)
{
  int i;

  for (i = 0; (i < n)&&(!(*found)); i++)
    if ( (!c[start][i]) && (cLbls[i] == UNLABLD) )
      if (cMatch[i] == UNMATCHD) {
        rLbls[start] = i;
        *tail = i;
        *found = TRUE;  /* An augmenting path has been found */
      }
      else {
        rLbls[start] = i;
        cLbls[i] = cMatch[i];
        augPath(c, n, cMatch[i], tail, found);
      }

  if ((i == n) && (!(*found))){
    rLbls[start] = INF;
  /*   *tail = start; */
  }
}

/*-------------------------------------------------------------------
  swapMatchng -  Swap the current matching when an augmenting path
                 is found
-------------------------------------------------------------------*/
int   swapMatchng (int root, int tip)

{
  int   done = FALSE, i = root;

  while (!done) {
    if (rLbls[i] == tip)
      done = TRUE;
    rMatch[i] = rLbls[i];
    cMatch[rLbls[i]] = i;
    i = cLbls[rLbls[i]];
  }
}

/*-------------------------------------------------------------------
  update -  update dual variables and cost matrix after swapping
-------------------------------------------------------------------*/
```

```
int update(int **c, int *u, int *v, int n)

{
  int i, j, delta = INF;

  for (i = 0; i < n; i++)
    for (j = 0; j < n; j++)
      if ( (c[i][j]) && (rLbls[i] != UNLABLD) && (cLbls[j] == UNLABLD) )
        delta = MIN(delta, c[i][j]);

  for (i = 0; i < n; i++) {
    if (rLbls[i] != UNLABLD)
      u[i] += delta;
    if (cLbls[i] != UNLABLD)
      v[i] -= delta;
  }

  for (i = 0; i < n; i++)
    for (j = 0; j < n; j++)
      if ( (rLbls[i] != UNLABLD) && (cLbls[j] == UNLABLD)
        && (c[i][j] != INF) )
        c[i][j] -= delta;
      else if ( (rLbls[i] == UNLABLD) && (cLbls[j] != UNLABLD)
            && (c[i][j] != INF) )
        c[i][j] += delta;
}

/*---------------------------------------------------------------------
  assign - find an optimal assignmnet
---------------------------------------------------------------------*/
int   assign(int **C, int N, int *U, int *V, int *X, int *Z)

{
  int   i, j;
    /* printf("INITIALIZATION BEGINNING \n"); fflush(stdout); */
  init(C, U, V, N);
        /* printf("INITILIZATION COMPLETE\n"); */
  while (!Optimal)
    {
/*        printf("CALLING improve\n");*/
        improve(C, U, V, N);
    }
  for ( i = 0, *Z = 0; i < N; i++)
    {
      *Z += (U)[i] + (V)[i];
      X[i] = rMatch[i];
    }

  free_ivector(rMatch,0,N);
  free_ivector(rLbls,0,N);
  free_ivector(cLbls,0,N);
  free_ivector(cMatch,0,N);

}
```

A.1.4 SIMULATION OF ASSIGNMENT UNDER REAL-TIME INFORMATION

```
/*** BEGIN SIMULATION OF SYSTEM WITH REAL-TIME ASSIGNMENT  ***/
iteration = 0;
{
While (iteration < MaxIterations and ConvergenceTest1 == False)
{
        iteration = iteration+1;
         function initialize variables();

          while (nextEventTime < EndOfHorizon)
          {

            function getNextEvent();

              while(nextEventType != REQUEST && nextEventTime < HORIZON)
             {

                    if (ConvergenceTest2 == TRUE)
                    {
                            check and record number of customers in all queues and in
                            system;
                    }

                    elapsedTime = nextEventTime;

                    function updateVehicleInvolvedInNextEvent();
                            /*** update position  and status of vehicle involved in the next
                                    event (a change of status) ***/
                    fuction getNextEvent();
             }

            elapsedTime = nextEventTime;

           if (nextEventType == REQUEST nextEventTime < EndOfHorizon)
            {

            for all vehicles  /* update location */
                    {
                            function undateVehiclePosition();
                    }

            /* Load is Accepted -- at least temporarily -- now find a feasible assignment */

            minCost = BIGCOST;
            cost = 0;
            SED = 0;
            minCostVehicle = 0;
            feasible = FALSE;
```

```
                for all vehicles  /* This is where en-route diversion comes in */
                {
                        calculate SED associated with best feasible ordering of currently
                        assignmed loads and candidate load
                }
                /* if ordering is both pickup deadline feasible and ELRatio attributable to
                        candidate load is less than the THRESHOLD  for acceptance */
                if ((SED < BIGCOST) &&
                ((SED - previous SED)/(Loaded Distance of Candidate Load) < THRESHOLD))
                {
                        feasible = TRUE;
                        switch(MEASURE)
                        {
                                case ELRATIO:
                                cost = SED/SLD
                                break;

                                case SED:

                                cost = SED;
                                break;

                                case DED:

                                cost = SED - previous SED
                        }
                }

                         if (cost < minCost)
                          {
                              minCostVehicle = currentVehicle;
                              minCost = cost;
                            }
                }
                if (feasible == TRUE)     /* assign and resequence loads */
                {
                function assignToQueueofminCostVehicle;
                /* if re-assignment of loads is allowed -
                        check each load for re-assignment now */
                }
                else loadisRejected;
            }
        }

                function sumResultsOfIteration();
                iteration = iteration + 1;
          }/*** END OF ITERATION ***/
function printResultsofSimulation();
}/*** END OF SIMULATION ***/
```

References

Ahuja, R.K., T.L. Magnanti and J.B. Orlin (1993), Network Flows, Prentice Hall, Englewood Cliffs.

American Trucking Associations (1978), Financial Analysis of the Motor Carrier Industry, ATA.

American Trucking Associations (1987), Financial Analysis of the Motor Carrier Industry, ATA.

American Trucking Associations (Statistics Department) (1996), American Trucking Trends, ATA.

American Trucking Associations Foundation (1996), Assessment of Intelligent Transportation Systems/Commercial Vehicle Operations User Services: ITS/CVO Qualitative Benefit/Cost Analysis, ATA Foundation.

Association of American Railroads (1992), 1992 Truckload Firm Profiles, AAR Policy and Special Projects Report.

Baker, E.K. (1983), An exact algorithm for the time constrained traveling salesman problem, *Operations Research*, 31, pp. 938-945.

Bertsimas, D.J., P. Jaillet, and A.R. Odoni (1990), A priori optimization, *Operations Research*, 36, pp. 1019-1033.

Bertsimas, D.J. (1992), A vehicle routing problem with stochastic demand, *Operations Research*, 40, pp. 574-585.

Bodin, L.D., B.L Golden, A.A. Assad, and M. Ball (1983), Routing and scheduling of vehicles and crews, the state of the art, *Computers and Operations Research*, 10, pp. 69-211.

Bookbinder, J.H. and S.P. Sethi (1980), The dynamic transportation problem: a survey, *Naval Research Logistics Quarterly*, 179, pp. 65-87.

Brown, A. (1992), A low cost vehicle location and tracking system, *Proceedings, IEEE Position Location and Navigation Symposium*, pp. 516-523.

Cheung, R.-M. and W.B. Powell (1995), An algorithm for multi-stage dynamic networks with random arc capacities, with an application to dynamic fleet management, *Operations Research*.

Christofides, N. and S. Eilon (1969), An algorithm for the vehicle dispatching problem, *Operational Research Quarterly*, 20, pp. 309-318.

Christofides, N. and S. Eilon (1969), Expected distances in distribution problems, *Operational Research Quarterly*, 20, pp. 437-443.

Christofides, N., A. Mingozzi and P. Toth (1981), State-space relaxation of bounds to routing problems, *Networks*, 11, pp. 145-164.

Christofides, N. (1985), Vehicle Routing, in Lawler, E.L., J.K Lenstra, A.H.G. Rinnooy Kan, and D. Shmoyes (eds), The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization, John Wiley and Sons, New York.

Clarke, G. and J.W. Wright (1964), Scheduling of vehicles from a central depot to a number of delivery points, *Operations Research*, 12, pp. 568-581.

Corsi, T.M. (1993), Motor carrier industry structure and operations, proceedings, International Symposium on Motor Carrier Transportation.

Crainic, T.G. , M. Gendreau, and P. Dejax (1992), Dynamic and stochastic models for the allocation of empty containers, *Operations Research* , 41, pp. 102-126.

Dantzig G. B. and J. H. Ramser (1959), The truck dispatching problem, *Management Science*, 6, pp. 80-91.

Desrosiers, J., Y. Dumas, M.M. Solomon and F. Soumis (1995), Time Constrained Vehicle Routing, in Ball, M.O., T.L. Magnanti, C.L. Monma and G.L. Nemhauser (eds), *Handbooks in Operations Research and Management Science, Vol 8., Network Routing*, Elsevier (North-Holland), Amsterdam, pp. 35-140.

Dejax, P.J. and T.G. Crainic (1987), A review of empty flows and fleet management models in freight transportation, *Transportation Science*, 21, 227-247.

Dumas, Y., J. Desrosiers and F. Soumis (1991), The pickup and delivery problem with time windows, *European Journal of Operational Research*, 54, pp. 7-22.

Eilon, S., C.D.T. Watson-Gandy and N. Christofides (1971), Distribution Management: Mathematical Modeling and Practical Analysis, Hafner, New York.

EnRoute Technology (1992), 1, 13, Waters Information Services.

Fisher, M.L. and R. Jaikumar (1981), A generalized assignment heuristic for vehicle routing, *Networks*, 11, pp. 109-124.

Fisher, M.L. (1995), Vehicle Routing, in Ball, M.O., T.L. Magnanti, C.L. Monma and G.L. Nemhauser (eds), *Handbooks in Operations Research and Management Science, Vol 8, Network Routing*, Elsevier, Amsterdam, pp. 1-33.

Frantzeskakis L.F. and W.B. Powell (1989), A successive linear approximation procedure for stochastic, dynamic vehicle allocation problems, *Transportation Science* , 24, pp. 40-57.

Gans, N. and G. van Ryzin (1996a), Optimal control of a multi-class, flexible queueing system, *Operations Research* (forthcoming).

Gans, N. and G. van Ryzin (1996b), Dynamic vehicle dispatching: optimal heavy traffic performance and practical policies, *Operations Research* (under review).

Glaskowsky, N. A. (1986), Effects of Deregulation on Motor Carriers, Eno Foundation for Transportation, Inc., Wesport.

Golden, B.L. and A.A. Assad (1986), Perspectives in vehicle routing: exciting new developments, Operations *Research*, 34, pp. 803-810.

Golden, B.L., and A.A. Assad (1988) (eds), Vehicle Routing: Methods and Studies, Elsevier (North-Holland), Amsterdam.

ITS America (1994), National ITS Program Plan.

Jacobs, I.M. and A. Salmasi (1991), The application of a novel two-way mobile satellite communications and vehicle tracking system to the transportation industry, *IEEE Transactions on Vehicular Technology,* 40, pp. 57-63.

Jaillet, P. (1988), A priori solution of a traveling salesman problem in which a random subset of the customers are visited, *Operations Research,* 36, pp. 929-936.

Jaillet, P. and A.R. Odoni (1988), The Probabilistic Vehicle Routing Problem, in Golden, B.L. and A.A. Assad (eds), Vehicle Routing: Methods and Studies, Elsevier (North-Holland), Amsterdam, pp. 293-318.

Knight, K.W. and J.P. Hofer (1968), Vehicle scheduling with timed and connected calls: a case study, *Operational Research Quarterly* , 19, pp. 299-310.

Kolen, A.W.J., A.H.G. Rinnooy Kan and H.W.J.M. Trienekens (1987), Vehicle routing with time windows, *Operations Research,* 35, pp. 266-273.

Koskosidis, Y.A., W.B. Powell and M.M. Solomon (1992), An optimization- based heuristic for vehicle routing and scheduling with soft time-window constraints, *Transportation Science*, 26, pp. 69-85.

Magnanti, T.L. (1981), Combinatorial optimization and vehicle fleet planning: perspectives and prospects, *Networks,* 1, pp. 179-213.

Mobility 2000 (1990), Report on the working group on commercial vehicle operations.

Nozaki, S.A. and S.M. Ross (1978), Approximations in finite-capacity multi-server queues with Poisson arrivals, *Journal of Applied Probability*, 15, pp. 826-834.

Powell, W.B. (1986), A stochastic formulation of the dynamic vehicle allocation problem, *Transportation Science*, 20, pp. 117-129.

Powell, W.B. (1987), An operational planning model for the dynamic vehicle allocation problem with uncertain demands, *Transportation Research,* 21-B, pp. 217-232.

Powell, W.B. (1988), A Comparative Review of Alternative Algorithms for the Dynamic Vehicle Allocation Problem, in Golden, B.L. and A.A. Assad,(eds), Vehicle Routing: Methods and Studies, Elsevier (North-Holland) Amsterdam, pp. 249-291.

Powell, W.B. and L. Frantzeskakis (1994), Restricted recourse strategies for dynamic networks with random arc capacities, *Transportation Science*, 28, pp. 3-23.

Powell, W.B., P. Jaillet and A. Odoni (1995), Stochastic and Dynamic Networks and Routing, in Ball, M.O., T.L. Magnanti, C.L. Monma and G.L. Nemhauser (Eds) *Handbooks in Operations Research and Management Science, Vol 8, Network Routing*, Elsevier, Amsterdam, pp. 141-296.

Powell, W.B. (1996), A stochastic formulation of the dynamic assignment problem, with an application to truckload motor carriers, *Transportation Science*, 30, pp. 195-219.

Psaraftis, H. N. (1988), Dynamic vehicle routing problems, in Golden, B.L. and A.A. Assad (eds), Vehicle Routing: Methods and Studies, Elsevier (North-Holland) Amsterdam.

Tijms, H. C. (1986), Stochastic Models and Analysis: an Algorithmic Approach, John Wiley and Sons, New York.

Thomas, J.H. (1971), Trucking: History and Legend, Ph.D. dissertation, Oklahoma State University.

Tillman, F. (1969), The multiple terminal delivery problem with probabilistic demands, *Transportation Science*, 3, pp. 192-204.

Regan, A.C., H.S. Mahmassani and P. Jaillet (1994), Improving the efficiency of commercial vehicle operations using real-time information, *proceedings of the First World Conference on Applications of Transport Telematics and Intelligent Vehicle-Highway Systems*, pp. 1547-1554.

Regan, A.C., H.S. Mahmassani and P. Jaillet (1995), Improving the efficiency of commercial vehicle operations using real-time information: potential uses and assignment strategies, *Transportation Research Record*, 1493, pp. 188-198.

Regan, A.C., H.S. Mahmassani and P. Jaillet (1996), Dynamic decision making for commercial fleet operations using real-time information, *Transportation Research Record*, 1537, pp 91-97.

Ross, S.M. (1989), Introduction to Probability Models, Academic Press, Boston.

Rothblatt, M. (1992), The first GPS satellite radio optimized for automatic vehicle location, *Proceedings, IEEE Position Location and Navigation Symposium,* pp. 524-527.

Rothenberg, L.S. (1994), Regulation, Organizations and Politics: Motor Freight Policy at the Interstate Commerce Commission, University of Michigan, Ann Arbor.

Sampson, R.J., M.T. Farris and D.L. Shrock (1985), Domestic Transportation: Practice, Theory and Policy, Houghton Mifflin, Boston.

Schmitt, R.R. and T.D. Feinberg (1994), (eds), Transportation Statistics Annual Report 1994, Washington, U.S. Department of Transportation.

Solomon, M.M. (1987), Algorithms for the vehicle routing and scheduling problem with time window constraints, *Operations Research*, 35, pp. 254-265.

Solomon, M.M. and J., Desrosiers (1988), Time window constrained routing and scheduling problems, *Transportation Science,* 22, pp. 1-13.

Van Hoorn, M.H. (1984), Algorithms and implementations for queuing systems, CWI Tracts, Amsterdam.

Wolff, R.W. (1989), Stochastic Modeling and the Theory of Queues, Prentice Hall, Englewood Cliffs.

Yee, J.R. and B. L. Golden (1980), A note on determining operating strategies for probabilistic vehicle routing, *Naval Research Logistics Quarterly,* 27, pp. 159-163.