DYNAMIC VEHICLE DISPATCHING: OPTIMAL HEAVY TRAFFIC PERFORMANCE AND PRACTICAL INSIGHTS

NOAH GANS

OPIM Department, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6366

GARRETT VAN RYZIN

Graduate School of Business, Columbia University, New York, New York 10027

(Received June 1996; revision received June 1997; accepted December 1997)

We analyze a general model of dynamic vehicle dispatching systems in which congestion is the primary measure of performance. In the model, a finite collection of tours are dynamically dispatched to deliver loads that arrive randomly over time. A load waits in queue until it is assigned to a tour. This representation, which is analogous to classical set-covering models, can be used to study a variety of dynamic routing and load consolidation problems. We characterize the optimal work in the system in heavy traffic using a lower bound from our earlier work (Gans and van Ryzin 1997) and an upper bound which is based on a simple batching policy. These results give considerable insight into how various parameters of the problem affect system congestion. In addition, our analysis suggests a practical heuristic which, in simulation experiments, significantly outperforms more conventional dispatching policies. The heuristic uses a few simple principles to control congestion, principles which can be easily incorporated within classical, static routing algorithms.

INTRODUCTION

In vehicle routing, loads of goods must be transported from a source location to a number of geographically dispersed destinations. Vehicles are assigned routes, and loads are assigned to vehicles in an attempt to optimize criteria that typically include measures of cost and of level of service. Such problems are usually modeled as static *design* problems. While there are many practical applications (such as school bus and garbage truck routing) that are undeniably *route design* problems, many applications involving routing and consolidation are, in reality, sequential dynamic decision problems. Loads arrive and vehicles are dispatched continuously over time. Examples include delivery of goods to retail stores, less-than-truckload (LTL) shipping networks, and parcel post delivery/pick-up, to name a few.

As Psaraftis (1988, 1995) notes, one important difference between dynamic and static routing environments is the possibility of congestion. With limited transportation capacity and variability in the mix and number of arriving loads over time, as well as variability in the times required to deliver loads, queueing delays are inevitable.

Such delays are more than a nuisance. For a carrier, they introduce inventories of loads waiting for delivery and directly drive the need for increased facility space. If delivery capacity is increased to eliminate congestion, the carrier risks severely underutilizing its transportation assets. For the shipper, long (and variable) throughput times directly increase pipeline inventories and indirectly drive the use of higher levels of safety stock. Indeed, based on a review of a several transportation industry surveys, Ballou (1985, p. 55) concludes that " . . . from a practical point of view, logistics customer service must focus on time-related elements."

With throughput time playing such an important role in both logistics cost and customer service, it is important for planners and managers to (1) understand what factors—such as mix of load types, constraints, level of variability—determine system congestion, (2) understand the precise tradeoff between delivery capacity and congestion, and (3) be able to design dispatching strategies that minimize congestion for a given delivery capacity.

In this paper, we propose and analyze a model of dynamic routing and consolidation that allows us to address these questions. In the model, a finite number of *load* types arrive randomly over time to a distribution facility and wait to be delivered. Deliveries from the distribution facility are made by a single vehicle, though it is not hard to extend the analysis to multiple vehicles. There is a finite collection of *routes* that the vehicle can use to deliver the loads. Each route is characterized by the number of each type of load it delivers, as well as by the time required to

Subject classifications: Transportation, freight/materials handling: queueing analysis of dynamic load, consolidation problems. Transportation, vehicle routing: queueing analysis of dynamic routing problems. Queues, applications: dynamic vehicle dispatching.
Area of review: TRANSPORTATION.

0030-364X/99/4705-0675 $05.00

complete the route. Whenever the vehicle becomes available, it can immediately be dispatched on another route or it can idle. The problem is to find a policy for dynamically selecting routes (*dispatching routes*) over time that minimizes congestion in the system. The measure of congestion we use, which we call *work*, is defined as the minimum time needed to deliver all waiting loads.

Using a sample path lower bound on system work (see Gans and van Ryzin 1997) and a novel analysis of an upper bound based on a simple batching heuristic, we find stability conditions and give a closed-form characterization of the optimal work in heavy traffic. The expression for optimal system work is a variant of the classical *GI*/*GI*/1 heavy traffic limit (see for example Daley et al. 1992, Kleinrock 1976). It combines first and second moment information on the arrival process with dual prices from an underlying set covering linear program derived from the collection of feasible routes. These dual prices have a natural interpretation as the work content (i.e., delivery time burden) that each load type imposes on the system. Together, the arrival statistics, dual prices and *GI*/*GI*/1 formula reveal the root sources of congestion and show directly how various model parameters—mix of incoming loads, collection of available routes, traffic intensity—affect optimal system congestion.

We also obtain several important insights on near-optimal dispatching policies. In particular, because the upper bound is constructive, our analysis provides an asymptotically optimal class of batching heuristics. While these heuristics do not appear to be practical in moderate traffic, they suggest some simple design principles for constructing more practical heuristics. Indeed, a prototypical heuristic incorporating these design principles, which we call CENTER, consistently outperforms other naive heuristics in a series of simulation tests. The rules used by the heuristic are simple and can be easily incorporated within classical, static routing algorithms.

Overview of Paper

The remainder of this paper is organized as follows. We begin in §1 with a literature review. Then §2 defines the model, and §3 presents two example problems that are used in the paper's numerical studies. In §4 we present our main analytical findings, and in §5 we offer the central argument for these results, providing technical proofs of intermediate results in an appendix. Readers who wish to concentrate on application, rather than proof, may proceed directly from §4 to §6, which presents the CENTER policy, as well as three competing heuristics. In §7 we provide results of numerical studies. Our conclusions are given in §8.

1. LITERATURE REVIEW

Psaraftis (1988, 1995) provides a comprehensive discussion of dynamic vehicle routing and defines a network version of the problem. We refer the reader to this work for an overview of dynamic routing applications, a good critical comparison of the differences between static and dynamic routing problems, and a comprehensive survey of the relevant literature.

Powell (1995) provides a comprehensive overview of stochastic programming approaches to a class of dynamic assignment problems which includes the management of truck-load-trucking fleets. These models match trucks with requests for (whole) trucks over a finite horizon, and the focus of the work is on computationally tractable, numerical methods. In contrast, we look at dispatching problems in which the capacity of one truck may be split to fill multiple requests. We also analyze a stationary problem over an infinite horizon.

Minkoff (1993) uses a Markov decision process (MDP) model to analyze a dynamic dispatching problem and proposes a decomposition heuristic. For a shipment consolidation problem, Higginson and Bookbinder (1994a) also propose a MDP model and algorithm. Though MDP models nicely capture the sequential decision making process inherent in dynamic vehicle routing, the approach is limited to very small scale problems and does not provide a great deal of structural insight.

Heuristic methods for shipment consolidation have also been investigated. Higginson and Bookbinder (1994b) perform a simulation study of time and quantity policies. Powell and Humblet (1986) analyze a bulk service queueing model of consolidation and provide a numerical method for finding the Laplace transform of the queue length distribution under several simple control strategies. Both these works, however, consider only a scalar (weight/size) constraint. Our model and analysis, in contrast, allow for general packing constraints.

Bertsimas and van Ryzin (1991, 1993a, 1993b) analyze a Euclidean version of a dynamic vehicle routing problem, called the dynamic traveling repairman problem (DTRP). In the DTRP, customer requests arrive according to a renewal process and their locations are randomly distributed in a Euclidean service region according to a given probability density function. A vehicle traveling at constant velocity serves the customers. Bertsimas and van Ryzin obtain bounds on the waiting time under an optimal policy and propose several heuristics whose performance lies within a constant factor of optimality in heavy traffic.

In this Euclidean model, congestion is driven primarily by the geometry of the problem. Because customers take on a *continuum* of locations according to a probability density function, the set of locations becomes more "dense" as the backlog grows, and this density allows for more efficient travel. In our problem, however, there are a *fixed number* of delivery locations, so as congestion increases it is the number of loads destined for each fixed location that grows. As a result there are no "economies of density" in heavy traffic. The appropriateness of each model depends on the application, with the Euclidean model best suited to applications with very high levels of variety in delivery locations (e.g. delivering appliances to

homes) and the discrete model best suited to applications involving a moderate number of fixed locations (e.g., wholesale distribution).

Reiman et al. (1996) address a dynamic distribution system operating in a make-to-stock mode. A single product is stocked at m retailers, each of which has random demand. A vehicle replenishes the retailer inventories using either direct shipments (fixed DS), a tour of all m retailers (fixed TSP) or possibly a combination of both (dynamic routing). The objective is to minimize long-run average holding, backorder and transportation costs. The authors provide a heavy traffic analysis of the fixed TSP policy, adapting recent heavy traffic results of Coffman et al. (1995a, 1995b) for polling systems.

Reiman et al.'s work provides modeling detail and insights that are complementary to ours. Our model allows for a variety of load types (e.g., different product types or shipment sizes) and an essentially unlimited number of feasible routes, each of which can reflect complex bin-packing or other combinatorial constraints. Reiman et al., in contrast, consider only simple direct-ship and/or TSP routes delivering a single product, with no constraints on the quantity delivered to each location. Our model therefore better captures the complexities in load types and routing options found in many bulk-cargo shipping applications (e.g., delivering automobiles from assembly plants to dealer lots). However, to address this added routing complexity rigorously, we must settle for a courser measure of system performance (system work) relative to that of Reiman et al.'s more detailed modeling of holding, backorder, and transportation costs.

Finally, Bertsimas and Simchi-Levi (1996) provide a recent survey of the DTRP and a priori (2-stage) stochastic routing problems. Similarly, Powell et al. (1995) offer a recent survey of the dynamic assignment problem, the DTRP and a priori stochastic routing problems.

2. PROBLEM DEFINITION AND NOTATION

In this section, we formally define our notation and model.

2.1. Notation

When describing vectors, we use the following conventions. $\mathbf{R}^m$ is m-dimensional Euclidean space, and $\mathbf{R}^m_+$ is its nonnegative orthant. Similarly, $\mathbf{Z}^m$ is the m-dimensional lattice of integers and $\mathbf{Z}^m_+$ the nonnegative portion of the set. Boldface $\mathbf{0} \in \mathbf{R}^m$ represents a vector of zeroes, and boldface $\mathbf{1} \in \mathbf{R}^m$ a vector of ones. The vector $e^j \in \mathbf{Z}^m_+$ has a one in the jth element and zeroes elsewhere. For $x \in \mathbf{R}^m$ and real $\epsilon > 0$, $\mathcal{N}(x, \epsilon) \stackrel{\text{def}}{=} \{y \in \mathbf{R}^m : y_i \in (x_i - \epsilon, x_i + \epsilon);\ i = 1, \ldots, m\}$ is the L_∞ ϵ-neighborhood of x.

We use three different symbols to represent weak and strict forms of vector inequalities. For $a \in \mathbf{R}^m$ and $b \in \mathbf{R}^m$, we write "$a \leqslant b$" whenever $a_i \leqslant b_i$ for all m elements, i. When $a_i \leqslant b_i$ for all m elements and there exists at least one element, k, for which $a_k < b_k$ we write "$a < b$," and when $a_i < b_i$ for all i, we write "$a \ll b$."

Functions of vectors are performed on a component-wise basis. For example, for $a \in \mathbf{R}^m$ and $b \in \mathbf{R}^m$, "$\min\{a, b\}$" yields a vector whose ith component equals $\min\{a_i, b_i\}$. Similarly, "$\mathrm{mod}\{a, b\}$" produces a vector whose ith component equals $a_i \bmod b_i$.

We follow these conventions when describing probabilistic events: $\{\cdot\}$ represents an event and $\{\bar{\cdot}\}$ its complement; $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function of an event; and $P\{\cdot\}$ designates the probability of an event. The abbreviations *i.i.d.* and *a.s.*, respectively, stand for "independent and identically distributed" and "almost surely."

2.2. Arrival Process

Consider a stream of loads that arrive to a distribution facility and wait to be delivered. A load *type* denotes a particular set of attributes such as location, size, weight, etc. that uniquely define a load's delivery requirements. We assume there are m load types. At arrival epochs, $\{t_k : k = 1, 2, \ldots\}$, quantities of the m load types arrive into the system according to a renewal process. We set $t_0 \equiv 0$ and let $T_k \stackrel{\text{def}}{=} t_k - t_{k-1}$ denote the interarrival time between arrivals $k-1$ and k; $\{T_k \in \mathbf{R}_+ : k = 1, 2, \ldots\}$ is a sequence of *i.i.d.* positive random variables. Let $E[T] = 1/\lambda > 0$ and variance $\sigma_T^2 < \infty$ ($\sigma_T^2 < \infty$ implies that $E[T]$, too, is finite).

For each load type, i, and every arrival epoch k, we define $V_k^i \in \mathbf{Z}_+$ to be the total number of type-i loads entering the system. We assume that the random vectors, $\{V_k \in \mathbf{Z}^m_+ : k = 1, 2, \ldots\}$ are *i.i.d.* and that each V_k satisfies $\mathbf{0} < V_k \ll \mathbf{1}C_1$ *a.s.* for some fixed $C_1 < \infty$. Let $\gamma = E[V_k]$ and $\Gamma = \mathrm{var}(V_k)$, the variance-covariance matrix of V_k. Without loss of generality, we assume $\gamma \gg \mathbf{0}$ as well.

Note that the arrival process is quite general. At any arrival epoch more than one load and/or more than one *type* of load may arrive into the system. That is, while the sequence of vectors, $\{V_k\}$, is *i.i.d.*, we place no independence restrictions among the elements of each V_k. Indeed, Γ may represent a broad range of variance-covariance relationships among the arriving quantities.

2.3. The Distribution Facility

The distribution facility has a single vehicle that can use any one of n possible routes to deliver loads. Each route, j, requires τ_j units of time to execute and can simultaneously deliver up to a_{ij} type-i loads. The matrix $A \in \mathbf{Z}^{m \times n}_+$ defines the delivery capacities of all n possible routes.

We assume that A has rank m, which implies that for each type of load i there exists at least one route j, for which $a_{ij} > 0$. Thus, a simple upper bound on the time required to deliver one unit of an arbitrary type of load is $\mathbf{1}^\top \tau$. Together with the upper bound of C_1 on each element of V_k, this implies that an arbitrary arrival can *a.s.* be processed in $mC_1\mathbf{1}^\top\tau$ units of time.

The facility can be idle or using exactly one of its n routes. Accordingly, we define the n-vector U_t, where $U_t = \mathbf{0}$ if the facility is idle at time t, and $U_t = e^j$ if it is currently executing route j at time t. Once a route j is dispatched, it

may not be changed for τ_j units of time. We define $O_t \in \mathbf{R}_+$ to be the residual time remaining for the route in use at time t. Each time deliveries commence under route j, O_t is set equal to τ_j, and as the deliveries proceed under j, O_t decreases at rate one until it equals zero and the route has completed execution. The state of the vehicle is therefore defined by the pair of values (U_t, O_t).

2.4. Dispatching Policies and System Performance

The set of arrivals up to time t, $\{(t_k, V_k): t_k < t\}$, along with the sample path of routes and times used up to t, $\{(U_r, O_r): r < t\}$, is called the *history* of the process up to t, $\mathcal{H}_t$. We define a *dispatching policy* π to be a rule that, given $\mathcal{H}_t$, allows the distribution facility to determine which route to use at t. This mapping must be nonanticipating with respect to $\{\mathcal{H}_t: t \geq 0\}$.

The history of the process up to time t also determines the system backlog at t. Let $Q_t \in \mathbf{Z}_+^m$ represent the quantities of loads in the system that have not entered into delivery by time t. Note that at arrival epochs t_k, there is a discontinuity in Q_{t_k} because of the arrival of a vector of loads V_k. Similarly, at epochs at which a delivery commences, there is a discontinuity, as integral numbers of loads leave the pool of loads waiting to be delivered. Let $\{t_{k'}: k' = 1, 2, \ldots\}$ be the combined sequence of arrival and delivery commencement epochs. Then $\{Q_t: \in \mathbf{Z}_+^m: t \geq 0\}$ obeys the following recursion:

$$Q_0 \equiv 0,$$

$$Q_t = \begin{cases} Q_{t_{k'}}, & t_{k'} < t < t_{k'+1}, \\ Q_{t_{k'}} + V_k, & t = t_{k'+1} \text{ and } V_k \text{ arrives at } t_{k'+1}, \\ Q_{t_{k'}} - \min\{Q_{t_{k'}}, a_j\}, & t = t_{k'+1} \text{ and } a_j \text{ commences at } t_{k'+1}. \end{cases}$$

Our basic measure of performance is the total *work* in the system, denoted W_t, where

$$W_t = O_t + \min\left\{ \sum_{j=1}^{n} \tau_j x_j : Ax \geq Q_t, x \in \mathbf{Z}_+^n \right\}. \tag{1}$$

W_t is the minimum time required by the distribution facility to feasibly clear the system starting at time t, assuming that no additional loads arrive after t. In the optimal solution, x_j is the number of type-j routes used to clear the backlog, and we require x_j to be integral.

This definition of work is the same as that of *completion time* in Bambos and Walrand (1993) and is a generalization of the definition of *work in system* for simple, single server queues in Wolff (1989). While somewhat course and aggregate, this definition of work is a quite natural measure of congestion. For example, it differentiates a low backlog—one that can be delivered in a few hours—from a high backlog—one that may take several weeks to completely clear. Furthermore, in most cases, high levels of work go hand-in-hand with high numbers of loads in queue, long throughput times, and other deleterious effects of congestion, and the minimization of work tends also to mitigate these negative effects of congestion. One may therefore view work as measuring undesirable system behavior rather than as a detailed service cost, similar in spirit to the squared error function of classical linear-quadratic control. As discussed in Gans and van Ryzin (1997), work also appears to be fundamental in determining other measures of system performance, playing a role analogous to that of the *workload process* in classical heavy traffic analysis.

Observe that the sample paths of $\{Q_t \in \mathbf{Z}_+^m: t \geq 0\}$, $\{O_t \in \mathbf{R}_+: t \geq 0\}$ and $\{W_t \in \mathbf{R}_+: t \geq 0\}$ depend on the dispatching policy π, as well as on the sample sequences of interarrival times $\{T_k\}$ and arrival quantities $\{V_k\}$. When we wish to emphasize the dependence on π, we will write Q_t^π, O_t^π, and W_t^π.

We say that a policy π is *stable* if

$$E_\pi[W] \stackrel{\text{def}}{=} \lim_{t\to\infty} \frac{1}{t} \int_0^t W_s^\pi \, ds$$

exists *a.s.* and is finite. All bounds and policies we analyze are in fact asymptotically stationary, so that $\lim_{t\to\infty} E[W_t^\pi] = \lim_{t\to\infty} 1/t \int_0^t W_s^\pi \, ds$, as our notation suggests. Let Π denote the class of all policies that are nonanticipating, stable and asymptotically stationary. We shall henceforth treat $E_\pi[W]$ as an expectation and restrict our attention only to policies $\pi \in \Pi$.

We call a policy, $\pi^* \in \Pi$, optimal if

$$E_{\pi^*}[W] = \inf_{\pi\in\Pi} E_\pi[W].$$

Let ρ be a measure of system utilization, which we define below (see Equation (6)). Then we call policy π° *asymptotically optimal* in heavy traffic if

$$\lim_{\rho\to 1} \frac{E_{\pi^\circ}[W]}{E_{\pi^*}[W]} = 1.$$

We address the definition of stability, the asymptotic characterization of optimal work and dispatching policies in Section 4; however, to be concrete, we next give two brief examples of "prototypical" transportation problems that can be modeled using the set-covering formulation: a load consolidation problem, and a one-warehouse, multiple-retailer problem. In Section 7 we use instances of these example problems as the basis for simulation experiments.

3. TWO EXAMPLE PROBLEMS

3.1. A Load Consolidation Problem

In the load consolidation problem, one truck ships m types loads from a source to a destination location. The m types of loads arrive at the source location at random time intervals and wait in queue to be delivered to the destination.

The truck can transport many loads in one delivery, and for any single delivery there are n ways in which the vehicle may be packed. Each packing must satisfy feasibility constraints: Typical restrictions are physical, based on the

volume and weight capacities of the vehicle. Each feasible packing constitutes a different route j, and a_{ij} is the number of type-i loads that are delivered by the truck under route j. For the load consolidation problem, the truck always moves between the same source and destination pair on each route. Therefore, all delivery times τ_j are equal, and without loss of generality we may set them to one. Our objective is to dynamically dispatch routes (i.e., packing configurations) over time so as to minimize the long-run expected backlog of loads waiting to be shipped to the destination.

3.2. A One-Warehouse, Multiple-Retailer Problem

In the one-warehouse, multiple-retailer example, a distribution facility replenishes a number of geographically dispersed retail outlets, each of which experiences random demands for several types of products. The warehouse has a single truck of limited capacity (maximum weight) that replenishes products to each of a number of retailers. In this case, each of the m load types designates a particular product destined for a particular retail location (i.e., a product-location pair).

The truck may deliver to more than one retail location on each route, and each product type may have a different weight. Based on the total weight constraints, we can enumerate each feasible delivery route j. Here, a_{ij} is the number of type i loads delivered under route j, and τ_j is the time required to complete route j (e.g., the shortest TSP tour of the sites the route visits). Other restrictions may be placed on the definition of a feasible route, including limits on distances traveled, numbers of sites visited, etc. Again, the objective is to dynamically dispatch routes over time so as to minimizes the long-run expected backlog of loads waiting to be delivered from the warehouse to a retailer.

4. MAIN RESULTS

Building on our earlier work in Gans and van Ryzin (1997), we develop a class of dispatching policies, $\{\pi_{q_N}\}$, that we demonstrate is asymptotically optimal in heavy traffic. The analysis in Gans and van Ryzin (1997) considers a relaxation of the distribution problem in which fractional routes may be used and fractional loads may be delivered. (This model has applications in analyzing flexible production and service systems.) Below, we extend these results to the distribution systems described in Section 3, for which the vehicle can only dispatch complete (nonfractional) routes.

4.1. The Relaxed System

Suppose the dispatching restrictions described in Section 2.3 are relaxed so that the distribution facility can use fractional quantities of the n routes. Furthermore, suppose the facility can use routes to process fractional numbers of loads.

For a given backlog Q_t, the time required to clear such a *relaxed* system will provide a lower bound on the time required to clear the system defined in Section 2. In particular, by removing the integrality restriction from (1) we create a linear program (LP) lower bound on system work at time t:

$$\underline{W}_t = O_t + \min\left\{ \sum_{j=1}^{n} \tau_j x_j : Ax \geqslant Q_t,\ x \geqslant \mathbf{0} \right\}. \tag{2}$$

REMARK. In Gans and van Ryzin (1997) each column of the matrix A is normalized by dividing by τ_j and represents *rates* of delivery for the m classes of loads, rather than quantities delivered. Similarly, because the facility processes arbitrarily small fractional units of backlog quantities, in Gans and van Ryzin (1997) Q_t is continuously reduced as the facility delivers the backlog, and $O_t = 0$ for all $t \geqslant 0$. We also note that what (2) defines to be $\underline{W}_t$ was labeled W_t in Gans and van Ryzin. In turn, Gans and van Ryzin defined $\underline{W}_t$ to be a further lower bound on (2).

4.2. A Lower Bound on System Work

Recall that $\gamma = E[V_k]$, and let $y^* \in \mathbf{R}^m$ be the optimal solution to the LP

$$\max\{\gamma^\top y : y^\top A \leqslant \tau,\ y \geqslant \mathbf{0}\}. \tag{3}$$

In Gans and van Ryzin (1997) it was shown that the lower bound

$$y^{*\top} Q_t \leqslant \underline{W}_t \tag{4}$$

holds for any backlog, Q_t. Observe that (3) is the dual of

$$\min\left\{ \sum_{j=1}^{n} \tau_j x_j : Ax \geqslant \gamma,\ x \geqslant \mathbf{0} \right\}. \tag{5}$$

For this relaxed system, we can interpret y^* as an allocation of work content, or processing time, to the different classes of loads. Similarly, the optimal solution to (5) identifies a *basis*, $B \in \mathbf{Z}_+^{m \times m}$, of efficient routes. That is, for any point Q_t in the cone of B, $B^{-1}Q_t$ is a feasible solution to (2). Furthermore, from (4) we see that the associated processing time, $\mathbf{1}^\top B^{-1} Q_t = y^{*\top} Q_t$, equals the optimal solution to (2).

In Gans and van Ryzin (1997) we show that the system work process in the $GI/GI/1$ queue with interarrival times $\{T_k : k = 1, 2, \dots\}$ and service times $\{y^{*\top} V_k : k = 1, 2, \dots\}$ provides a sample-path lower bound for work in the relaxed system under any policy. Then defining

$$\rho \stackrel{\text{def}}{=} \lambda y^{*\top} \gamma, \tag{6}$$

as we would for system utilization in the $GI/GI/1$ queue, we use this lower-bound process to derive stability results and a heavy-traffic lower bound for expected work in the relaxed system.

THEOREM 1 (Gans and van Ryzin 1997). (*i*) *If* $\rho > 1$ *then* $E_\pi[\underline{W}] = \infty$ *for any policy* $\pi \in \Pi$ *and the system is unstable*; (*ii*) *if we scale* T *so that* $\rho \to 1$ *from below as* $\lambda \to 1/y^{*\top}\gamma$, *then*

$$\lim_{\rho\to 1}(1-\rho)E_{\pi}[\underline{W}] \geq \frac{\lambda(\sigma_T^2 + y^{*\top}\Gamma y^*)}{2}.$$

REMARK. Theorem 1 uses the limit "$\rho \to 1$." Technically, however, this limit is shorthand for a description of a sequence of stable systems, (indexed by n) for which V_n and T_n both converge in distribution and $\rho_n \to 1$ from below (for example see Wolff 1989, p. 518). To simplify the exposition, we will henceforth consider the distribution of the arrival sequence $\{V_k; k \geq 1\}$ to be *fixed*, and interpret $\rho \to 1$ to mean an increase in the *rate* at which these job quantities arrive.

Recall that $\underline{W}_t$ is a lower bound for W_t. Therefore, the bounds provided by Theorem 1 may be used to define instability and a heavy-traffic lower bound on expected system work in the restricted system as well.

4.3. Upper Bounds for the Relaxed System Based on Batching Policies

We demonstrate in Gans and van Ryzin (1997) that a class of policies, $\{\pi_{B_N}: N = 1, 2, \dots\}$, is asymptotically optimal as $\rho \to 1$. In the policies $\{\pi_{B_N}\}$, the distribution facility acts as a *bulk-service queue* in which the individual arrival vectors $\{V_k\}$ are served in batches of N. While somewhat clumsy from a practical standpoint, these batching policies have the advantage of being analytically tractable, providing constructive, closed-form upper bounds on the optimal work.

We may think of bulk service as operating in two stages. In the first stage, an *accumulator* collects batches of N arrivals. In the second, a *batch server* processes these batches of N. For the policy π_{B_N} in particular, a batch is formed every Nth arrival, where N depends on ρ. We call every Nth arrival epoch a *batching epoch*, because these are times at which the accumulator passes batches to the batch server.

In the policies $\{\pi_{B_N}\}$, the batch server processes incoming batches on a first-come, first-served basis. For each batch, l, it substitutes the quantities associated with the lth batch ($\Sigma_{k=1}^{N} V_{N(l-1)+k}$) for Q_t in the right-hand side of (2) and uses the LP's optimal solution to determine the times for which the various routes will run. Thus, the batch server behaves as a $GI/GI/1$ queue with interarrival times that are the sums of N system interarrival times T_k and service times that are determined by solving the LP (2) for each successive batch.

The intuition behind using the class $\{\pi_{B_N}\}$ is to pick a large N so that quantities to be processed in each batch are likely to fall in the cone of B. Then the processing time of each batch l is likely to achieve the lower bound, $y^{*\top}\Sigma_{k=1}^{N} V_{N(l-1)+k}$. If the solution to (5) is not degenerate, then γ lies in the interior of the cone of B, so as N grows large, the probability that $\Sigma_{k=1}^{N} V_{N(l-1)+k}$ falls outside of the cone decreases rapidly.

4.4. Modified Policies to Serve Restricted Systems

Our goal is to apply this batching heuristic to the dynamic dispatching problem. However, the integrality restrictions imposed in (1) introduce a potentially significant source of idleness that does not exist in the relaxed version of the system, and it is not clear a priori that Theorem 1 provides a tight lower bound on expected system work in heavy traffic. For example, one can show that rounding the LP relaxation in the batch policy will not produce an asymptotically optimal policy.

Nevertheless, we are able to modify the batching policies so that they satisfy the integrality restrictions of (1) and converge to the lower bound of Theorem 1 (ii). The modification uses results on totally dual integral (TDI) systems of linear inequalities. More specifically, we construct batches so that the backlogs of all m types of loads in each batch are multiples of some large integer, q. This, in turn, ensures that for every batch, the solution to the LP relaxation of (1) is TDI and produces an optimal feasible solution for the original integer program. The new class of policies is called $\{\pi_{q_N}: N = 1, 2, \dots\}$. Again, we point out that this class of policies is clearly not practical. However, the policies have the correct asymptotic behavior and therefore allow us to precisely characterize optimal work in heavy traffic.

THEOREM 2. *(i) If $\rho < 1$ and y^* is the unique solution to* (3), *then there exist a $C_2 > 0$, a $\theta > 0$ and an integer, N^*, such that all members of the class $\{\pi_{q_N}: N = 1, 2, \dots\} \in \Pi$ for which $N \geq N^*$ are stable and satisfy*

$$E_{\pi_{q_N}}[\underline{W}] \leq \frac{\lambda(y^{*\top}\Gamma y^* + \sigma_T^2)}{2(1-\rho-\lambda O(e^{-\theta N}))} + \frac{C_2}{N(1-\rho-\lambda O(e^{-\theta N}))} + O(N).$$

(ii) If the conditions of Theorem 1 (*ii*) *and Theorem* 2 (*i*) *hold, and we define $N \stackrel{\text{def}}{=} \lceil (1-\rho)^{-b} \rceil$ for an arbitrary, fixed $b \in (0, 1)$, then*

$$\lim_{\rho\to 1}(1-\rho)E_{\pi_{q_N}}[W] \leq \frac{\lambda(\sigma_T^2 + y^{*\top}\Gamma y^*)}{2},$$

so that the policies in the class $\{\pi_{q_N}\}$ are asymptotically optimal as $\rho \to 1$. Moreover, together with Theorem 1 (*ii*) *this implies*

$$\lim_{\rho\to 1}(1-\rho)E_{\pi^*}[W] = \frac{\lambda(\sigma_T^2 + y^{*\top}\Gamma y^*)}{2}.$$

Theorem 2 shows that as the traffic intensity grows, the integrality restrictions of the distribution problem do not fundamentally affect the optimal level of work relative to the relaxed system. In turn, the dual prices to the LP relaxation of (1), y^* accurately describe the time required to ship each of the m types of loads in heavy traffic.

In the following section we develop these results, and in Section 6 we present an effective, practical heuristic that uses insights from the analysis. The policy is dynamic,

choosing a new route each time the previous one completes execution. In selecting the next route to be used, the policy uses the dual prices, y^*. Readers who wish to concentrate on the practical application of these results may proceed directly to Section 6.

5. PROOF OF THE MAIN THEOREM

Our proposed policy creates batches every Nth arrival epoch as before. However, at batching epochs quantities of the m types of loads in the accumulator are *rounded down* to the greatest multiple of q to form a batch. Any leftover loads after the rounding become part of the pool of loads in the accumulator from which the following batch is formed.

The presence of these leftover loads introduces dependencies among the sequence of batched quantities. However, by viewing the sequence of quantities arriving at the batch server as a discrete time Markov chain, we show that the covariance between any pair of quantities vanishes exponentially quickly and that the expected work is bounded above by the analogous $GI/GI/1$ upper bound plus a constant.

5.1. Totally Dual Integral Polyhedra

The modified class of policies makes use of a well-known result in polyhedral combinatorics concerning total dual integrality. To motivate the new class of policies, we first present this result.

A rational linear system of inequalities, $\{y^\top A \leqslant \tau^\top, y \geqslant \mathbf{0}\}$, is *totally dual integral* if, for any integer objective Q for which $\mathbf{max}\{y^\top Q: y^\top A \leqslant \tau^\top, y \geqslant \mathbf{0}\}$ has an optimal solution, the corresponding dual program $\mathbf{min}\{\tau^\top x: Ax \geqslant Q, x \geqslant \mathbf{0}\}$ has an integral optimal solution. Giles and Pulleyblank (1979) proved the following concerning TDI polyhedra.

THEOREM 3 (Giles and Pulleyblank 1979). *For any rational linear system of inequalities, $\{y^\top A \leqslant \tau^\top, y \geqslant \mathbf{0}\}$, there exists a rational $\alpha > 0$ such that $\{y^\top(\alpha A) \leqslant (\alpha\tau^\top), y \geqslant \mathbf{0}\}$ is TDI. When A is integral, α^{-1} is integer valued as well.*

We can apply this result to our system to show there exists a rational $\alpha > 0$ such that the solution to $\mathbf{min}\{\sum_{j=1}^n \alpha\tau_j x_j: \alpha Ax \geqslant Q_t, x \geqslant \mathbf{0}\}$ is integral for all integral Q_t for which there exist an optimum. Since we assume that $A \in \mathbf{Z}_+^{m\times n}$ is of full rank, there always exists a finite optimal solution for any finite backlog vector, $Q_t \in \mathbf{Z}_+^m$. Furthermore, since A is integral, we can define $q \stackrel{\text{def}}{=} \alpha^{-1}$ and rescale the linear program by q to show that $\mathbf{min}\{\sum_{j=1}^n \tau_j x_j: Ax \geqslant qQ_t, x \geqslant \mathbf{0}\}$ has an integral optimal solution for all integral Q_t.

Therefore, if we can ensure that the backlogs of all m types of goods are multiples of q units, then the solution to the LP relaxation in (2) will produce an optimal feasible solution for the original integer program in (1).

5.2. The Class of Modified Batching Policies

For the class $\{\pi_{q_N}\}$ the system acts as a bulk service queue. At batching epochs, the accumulator forms a batch by rounding down the backlogs of each of the m types to the nearest multiple of q. It then sends the batch to the batch server and retains the remaining loads. These remaining loads become a part of the total backlog in the accumulator at the next batching epoch.

To facilitate our analysis, we define three random sequences $\{\hat{T}_l \in \mathbf{R}_+: l = 0, 1, \dots\}$, $\{\hat{V}_l \in \mathbf{Z}_+^m: l = 0, 1, \dots\}$, and $\{\hat{R}_l \in \mathbf{Z}_+^m: l = 0, 1, \dots\}$ as follows:

$$\hat{T}_l \stackrel{\text{def}}{=} \sum_{k=1}^{N} T_{N(l-1)+k},$$

$$\hat{V}_l \stackrel{\text{def}}{=} \hat{R}_{l-1} + \sum_{k=1}^{N} V_{N(l-1)+k} - \hat{R}_l,$$

$$\hat{R}_l \stackrel{\text{def}}{=} \text{mod}\left(\hat{R}_{l-1} + \sum_{k=1}^{N} V_{N(l-1)+k}, q\mathbf{1}\right). \tag{7}$$

$\hat{T}_l$ is the interarrival time between the $(l-1)$st and lth batches, $\hat{V}_l$ is the vector of loads making up the lth batch, and $\hat{R}_l$ is the set of loads remaining in the accumulator after the lth batching epoch. For all batches, $l \geqslant 1$, each element of $\hat{V}_l$ is an integral multiple of q and each element of $\hat{R}_l$ is an integer between 0 and $q-1$. Note that $\hat{R}_0$ specifies an initial system backlog at time zero.

In turn, we define the sequence of random variables, $\{\hat{S}_l \in \mathbf{R}_+: l = 0, 1, \dots\}$, to be the processing times derived by substituting each $\hat{V}_l$ in the right-hand side of (2) and solving the LP. Thus, the batch server behaves as a $GI/G/1$ queue with interarrival times $\{\hat{T}_l\}$ and service times $\{\hat{S}_l\}$.

5.3. An Upper Bound on the Expected Backlog

To develop an upper bound on $E_{\pi_{q_N}}[W]$ we will separately bound the time averages of the work found in the accumulator and in the batch server. The sum of these two bounds provides a crude upper bound on total average system work.

We begin with the accumulator. Under policy π_{q_N} the quantity of loads in the accumulator starts at $\hat{R}_{l-1}$ at the beginning of the lth batching cycle, increases throughout the cycle, and reaches its peak, $\hat{V}_l + \hat{R}_l$, as the Nth arrival of the batch occurs and the batch is dispatched to the batch server. Therefore, the processing time of the backlog in the accumulator peaks at the end of any cycle.

Recall from Section 2.2 that the time required to process an arbitrary arrival V_k is bounded above by $mC_1\mathbf{1}^\top\tau$. Similarly, recall that $\mathbf{1}^\top\tau$ is an upper bound on the time required to process a unit of the backlog of arbitrary type and that each element of $\hat{R}_{l-1}$ is bounded above by q. Because the backlog in the accumulator at the lth batching epoch is $\hat{V}_l + \hat{R}_l = \hat{R}_{l-1} + \sum_{k=1}^{N} V_{N(l-1)+k}$, we can define

$$C_3 \stackrel{\text{def}}{=} (mC_1 + q)\mathbf{1}^\top\tau, \tag{8}$$

so that NC_3 is an upper bound on the time required to process the backlog in the accumulator at any time during an arbitrary cycle.

Our next task is to find an appropriate upper bound on the average time required to process the backlog at the batch server. Under policy π_{q_N} the batch server processes incoming batches on a first in, first out (FIFO) basis. While the batch server behaves as a single-server queue with *i.i.d.* interarrival times $\{\hat{T}_l\}$, service times $\{\hat{S}_l\}$ are *not* independent of each other. In particular, $\hat{V}_l$ depends on the size of the previous batch through the remainder term, $\hat{R}_{l-1}$. Note that if $\{\hat{R}_l\}$ is asymptotically stationary, the sequence of distribution times $\{\hat{S}_l\}$ is stationary as well.

Suppose $\{\hat{R}_l\}$ is stationary and let $\hat{D}$ be the delay that a batch finds upon arrival to the batch server. Then the following lemma provides an upper bound on $E[\hat{D}]$.

LEMMA 1 (found in Daley et al. 1992, p. 187). *Let D be the delay found upon arrival in a GI/G/1 queue with* i.i.d. *interarrival times T, stationary service times S, and* $\{S\}$ *independent of* $\{T\}$. *If there exist dependencies among service times, then whenever* $E[S] < E[T]$ *and* $E[D^2] < \infty$ *the following bound holds*: $E[D] \leq E[(S - T)^2]/2(E[T] - E[S])$.

In the case of the batch server, with service times $\hat{S}_l$ and interarrival times $\hat{T}_l$, Lemma 1 is equivalent to

$$E[\hat{D}] \leq \frac{\text{var}(\hat{S}_l) + \text{var}(\hat{T}_l)}{2(E[\hat{T}_l] - E[\hat{S}_l])} + \frac{E[\hat{T}_l] - E[\hat{S}_l]}{2}.$$

In turn, the following well-known relationship between the time average of system work and expected delay upon arrival will allow us to characterize an upper bound on time-average work waiting at the batch server.

LEMMA 2 (found in Wolff 1989, p. 279). *Let* $E[D]$ *be the expected delay found upon arrival and* $E[W]$ *be the time average of system work in a stable, work-conserving G/G/1 queue with stationary interarrival times T and stationary service times S. Then*

$$E[W] = \frac{E[SD]}{E[T]} + \frac{E[S^2]}{2E[T]}.$$

Unfortunately, we cannot immediately use Lemmas 1 and 2 to develop a simple upper bound on $E_{\pi_{q^N}}[W]$. While interarrival times to the batch server are *i.i.d.* with

$$E[\hat{T}_l] = N/\lambda \quad \text{and} \quad \text{var}(\hat{T}_l) = N\sigma_T^2, \tag{9}$$

an analysis of the sequence of service times $\{\hat{S}_l\}$ is much more difficult. In particular, each $\hat{S}_l$ is the solution to an LP. Furthermore, dependencies among service times make the moments of $\hat{S}_l$, as well as $E[\hat{S}_l\hat{D}_l]$, difficult to calculate.

Therefore, rather than attempting to calculate moments directly, we take the approach of Gans and van Ryzin (1997) and define a sequence of service times $\{\bar{S}_l: l = 1, 2, \dots\}$, which is easier to analyze and provides a sample-path upper bound on the sequence $\{\hat{S}_l\}$. The expected delay upon arrival to the batch server under policy π_{q_N} can then be bounded by

$$E[\hat{D}] \leq \frac{\text{var}(\bar{S}_l) + \text{var}(\hat{T}_l)}{2(E[\hat{T}_l] - E[\bar{S}_l])} + \frac{E[\hat{T}_l] - E[\bar{S}_l]}{2}. \tag{10}$$

Similarly, the time average of work in a system with batch service times $\{\bar{S}_l\}$ will provide an upper bound on the average system work under π_{q_n}.

We begin the construction of $\bar{S}_l$ by noting that for some batches $\hat{S}_l = y^{*\top}\hat{V}_l$ and the processing time achieves the lower bound. This happens when $\hat{V}_l$ lies within the cone of the optimal basis B of (5). More formally, if the solution to (3) is unique, then (5) is nondegenerate, and there exists an $\epsilon > 0$ such that y^* remains the vector of dual prices for all backlogs $Q \in \mathcal{N}(\gamma, \epsilon)$ that are substituted in the right-hand side of (2) (see Bazaraa et al. 1990, p. 260). Since LPs are homogeneous of degree one in their right-hand sides, y^* is the optimal vector of dual prices for all $\hat{V}_l \in \mathcal{N}(N\gamma, N\epsilon)$. Thus, $\hat{S}_l = y^{*\top}\hat{V}_l$ for any batch, l for which $\hat{V}_l \in \mathcal{N}(N\gamma, N\epsilon)$.

For other batches, $\hat{V}_l \notin \mathcal{N}(N\gamma, N\epsilon)$ and the dual prices y^* may not apply. NC_3 provides a uniform upper bound on the processing time of these batches. Let $\{E_l\} \stackrel{\text{def}}{=} \{\hat{V}_l \notin \mathcal{N}(N\gamma, N\epsilon)\}$. The following proposition shows that the probability that $\{E_l\}$ occurs is exponentially decreasing in N.

PROPOSITION 1. *If* y^* *is the unique solution to* (3), *then there exists a* $\theta > 0$ *such that* $P\{E_l\} = O(e^{-\theta N})$.

Using the bounds on the processing times of $\hat{V}_l$ and the definition of $\{E_l\}$ we then define the upper bound on $\hat{S}_l$:

$$\bar{S}_l \stackrel{\text{def}}{=} y^{*\top}\hat{V}_l + NC_3\mathbf{1}_{\{E_l\}}. \tag{11}$$

In turn, we use (11) to derive the following upper bound on time-average system work under π_{q_n}:

PROPOSITION 2. *If* y^* *is the unique solution to* (3), *then for any* $\rho < 1$ *there exists an integer* $N_1^* < \infty$ *such that for all* $N \geq N_1^*$, $E_{\pi_{q^N}}[W] \leq E[\hat{D}] + O(N)$.

The proofs of Propositions 1 and 2 may be found in the Appendix.

Before we can demonstrate the asymptotic optimality of the policies $\{\pi_{q_N}\}$, we must solve two problems. First, (10) holds only for stationary queues for which $E[\hat{D}^2] < \infty$. We must show that, without loss of generality, we may analyze a stationary version of the batch server and must bound the second moment of its delay. Second, to use (10) to prove asymptotic optimality we must find sufficiently tight upper bounds on the first two moments of $\bar{S}_l$. To do this we must address the dependence among elements of the sequence $\{\bar{S}_l\}$. We accomplish both tasks by viewing the sequence of service times at the batch server as a function of a discrete-time Markov chain.

5.4. The Batch Service Queue as a Mixing Process

We recall that the service time of the lth batch is linked to that of previous batches through the remainder term, $\hat{R}_{l-1}$. Not only does this remainder term have a direct effect on the size of the following batch, it also has an indirect effect on subsequent batches, through later remainder terms. Still, we might imagine that the farther out in the sequence of $\hat{V}_l$s we look, the less effect the realization of $\hat{R}_{l-1}$ has on batch sizes, particularly if the number of arrivals included in a batch N dwarfs $q - 1$, the maximum value that $\hat{R}_{l-1}$ may obtain.

For a stationary sequence of random variables, this asymptotic independence among terms is made precise using the notion of a *mixing process*. Specifically, suppose $\{S_1, S_2, \dots\}$ is a stationary sequence of random variables. For $a \leqslant b$ define $\mathcal{F}_a^b$ to be the σ-field generated by $\{S_a, \dots, S_b\}$. Then $\{S_k\}$ is φ-*mixing* if there exists a sequence $\{\varphi_1, \varphi_2, \dots\}$ for which $\lim_{n\to\infty} \varphi_n = 0$ such that

$$|P\{F_1 \cap F_2\} - P\{F_1\}P\{F_2\}| \leqslant \varphi_n P\{F_1\} \tag{12}$$

for all events $F_1 \in \mathcal{F}_0^k$ and $F_2 \in \mathcal{F}_{k+n}^\infty$.

It is not difficult to show that the arrival process to the batch server is asymptotically φ-mixing. The underlying sequence of interarrival times $\{T_k\}$ is i.i.d. and therefore stationary and φ-mixing with $\varphi_n = 0$ for all n.

To demonstrate that the sequence of distribution times is asymptotically φ-mixing we represent the distribution times at the batch server as a function of the evolution of a discrete-time Markov chain with a $2m$-dimensional state space. It is well known that a finite, homogeneous, aperiodic Markov chain is asymptotically stationary. It is also known (c.f. Billingsley 1968, p. 167–168), that a finite, homogeneous, *stationary* Markov chain is φ-mixing with

$$\varphi_n = \alpha\beta^n \tag{13}$$

for some fixed constants $\alpha > 0$ and $\beta \in (0, 1)$.

We describe the Markov chain. Each arrival epoch t_k in the original system corresponds to a transition of this embedded Markov chain. At each transition, the first m elements of the Markov chain are assigned the values realized by V_k; we will continue to refer to these first m elements as V_k. We call the second set of m elements of the state space $R_k \in \mathbf{Z}_+^m$, and at each transition, we define

$$R_k \stackrel{\text{def}}{=} \text{mod}(R_{k-1} + V_k, q\mathbf{1}),$$

where $R_0 \stackrel{\text{def}}{=} \hat{R}_0$. Thus, $\{R_k: k = 0, 1, \dots\}$ describes the evolution of the remainder term at each arrival epoch, and at each batching epoch l, $R_{Nl} = \hat{R}_l$.

For $0 \leqslant a \leqslant b$ we call $\mathcal{F}_a^b$ the σ-field generated by states a through b of the Markov chain. Then $\hat{V}_l \in \mathcal{F}_{N(l-1)}^{Nl}$, and $\hat{S}_l \in \mathcal{F}_{N(l-1)}^{Nl}$ as well, since it is a function of $\hat{V}_l$. Thus, for stationary $\{\hat{V}_l\}$ the sequence $\{\hat{S}_l\}$ is φ-mixing with φ_n defined as in (13). We note that a stationary version of the batch service system initializes the starting backlog R_0 according to the Markov chain's stationary distribution for R_k.

The asymptotically φ-mixing property of the arrival process suggests that we should be able to use such a stationary version of the batch server to determine the upper bound for $E[\hat{D}]$ and that the result should hold, no matter what the actual initial backlog, R_0. In particular, Szczotka (1990, p. 233) notes that the limiting waiting time distribution of a $G/G/1$ queueing system is the same as that of the stationary version of the system whenever: (a) the interarrival and service times of the original system can be defined as functions of a homogeneous, positive recurrent Markov chain; and (b) the stationary version has $E[\hat{S}_l] < E[\hat{T}_l]$.

We have developed just such a Markov chain as required by (a), and the following proposition demonstrates how we can maintain (b) by construction.

PROPOSITION 3. *If y^* is the unique solution to* (3), *then for the stationary sequence of distribution times $\{\bar{S}_l\}$,*

$$E[\bar{S}_l] = Ny^{*\top}\gamma + O(Ne^{-\theta N}).$$

In turn, for any $\rho = \lambda y^{*\top}\gamma < 1$ there exists an integer, $N_2^* < \infty$ such that $E[\bar{S}_l] < E[\hat{T}_l]$ for all batch sizes $N \geqslant N_2^*$.

Thus, by choosing a large enough batch size, we can ensure that the stationary version of the (upper bound $\bar{S}$ of the) batch server maintains condition (b). We can similarly demonstrate the finiteness of $E[\hat{D}^2]$ whenever $\rho < 1$.

PROPOSITION 4. *For any fixed $\rho < 1$, there exists an integer $N_3^* < \infty$ such that $E[\hat{D}^2] < \infty$ for all batch sizes $N \geqslant N_3^*$.*

Given a stationary sequence of batch distribution times, $\{\bar{S}_l\}$, we can then use (12) to develop the following bound on $\text{var}(\bar{S}_l)$.

PROPOSITION 5. *If y^* is the unique solution to* (3), *then there exists an integer $N_4^* < \infty$ and a real $0 < C_2 < \infty$, such that for any batch size $N \geqslant N_4^*$,*

$$|\text{var}(\bar{S}_1) - Ny^{*\top}\Gamma y^*| \leqslant C_2.$$

With the first two moments of $\bar{S}_l$ at our disposal, we can use (10) and Proposition 2 to find the upper bound on $E_{\pi_{q_N}}[W]$ shown in Theorem 2 (ii) and, in turn, to demonstrate the asymptotic optimality of the class of policies $\{\pi_{q_N}\}$ in heavy traffic. The proofs of Propositions 3, 4, and 5, as well as that of Theorem 2, may be found in the Appendix.

6. PRACTICAL DISPATCHING HEURISTICS

As mentioned in Section 4, the batching policies $\{\pi_{q_N}\}$ are designed primarily to provide analytically tractable upper bounds, and, as simulation results in Gans and van Ryzin (1997) have shown, they appear to have limited practical potential. Nevertheless, we can use the insights from the analysis in Section 5 to develop practical policies that appear to perform well in simulation experiments.

In this section, we describe four heuristic dispatching policies. Two are "straw" policies that are intended to

mimic naive heuristics that might be used in practice. Another two are practical policies constructed using insights from our asymptotic analysis. All four policies dynamically choose the next route to be used each time a route completes execution. Hence, unlike the batching policies, they do not require a large backlog to be present in the system to begin dispatching routes.

6.1. Two Straw Policies

One simple heuristic policy is to choose the route that will process the greatest number of loads per unit of time. The NUMBER policy follows this approach. Each time it completes the execution of a column, NUMBER uses the current backlog, Q, to calculate

$$\tau_j^{-1}\mathbf{1}^\top \min\{Q, a_j\}, \tag{14}$$

for each column, j. NUMBER then selects the column, k, that maximizes (14).

NUMBER is a variant of the "shortest processing time first" (SPT) scheduling rule, because (14) measures the number of loads delivered per unit time. The SPT rule, when applied preemptively, is known to minimize the expected delay in single-server systems (see Wolff 1989, p. 445–446).

A variant of NUMBER assigns a weight to each type of load and chooses the route that processes the maximum amount of weight per unit of time. For the WEIGHT policy, the user inputs a weight parameter w_i for each type of load before the simulation begins. To decide which column to process, WEIGHT then uses $w \in \mathbf{R}^m$ to calculate

$$\tau_j^{-1}w^\top \min\{Q, a_j\}, \tag{15}$$

for each column j and selects the column k that maximizes (15). The weights may represent some measure of the shipping capacity used by the different types of loads, such as physical weight or volume. Alternatively, the weights may represent some exogenous measure of the relative desirability of delivering each of the different types of loads.

For both policies it is possible that more than one column maximizes (14) or (15). In these cases both policies use the same set of three tiebreaking rules: (a), if column k dominates column j,

$$\min\{Q, a_j\} < \min\{Q, a_k\},$$

then choose column k; otherwise (b), if test (a) results in a tie, then if the maximum residual backlog left by column k is less than that left by column j,

$$\max_{1\leq i\leq m}\{\max\{0, Q_i - a_{ik}\}\} < \max_{1\leq i\leq m}\{\max\{0, Q_i - a_{ij}\}\},$$

then choose column k; or (c), if tests (a) and (b) both result in ties, then flip a coin. Note that test (b) acts as a check that the backlogs of the different types of loads remain roughly in balance.

6.2. Two Policies Motivated by the Analysis

Our definition of work, the minimum time required by the distribution facility to clear the system backlog, suggests a greedy heuristic as a natural choice. Following this objective, the GREEDY policy always seeks to deliver the current system backlog as quickly as possible, without regard for future arrivals.

Each time it completes the execution of a column, GREEDY substitutes the current backlog, Q, in the right-hand side of (2) and solves the LP. The result of (2) determines a direct path back to the origin that requires a minimum of delivery time. GREEDY then uses the dual prices from the optimal solution of the LP (2) to select the next column it will use to process the backlog. Specifically, GREEDY chooses a column with minimum *effective* reduced cost. That is, for each column, j, GREEDY calculates the effective reduced cost as

$$\tau_j - y_t{}^\top \min\{Q_t, a_j\}, \tag{16}$$

where $y_t \in \mathbf{R}_+^m$ is the vector of dual prices associated with the optimal solution to (2). GREEDY then selects the column, k, that minimizes (16) as the next route it will use to process the backlog. The column runs for τ_k units of time, and when it completes execution, the selection process begins anew. If more than one column minimizes (16), GREEDY uses the same tiebreaking rules described for NUMBER and WEIGHT to select the next route. We note that GREEDY does not make use of the optimal basis, B, nor of any information concerning the distributions of V and T.

The last policy, CENTER, uses additional insights from our asymptotic analysis. In particular, the analysis of the batching policies suggests two objectives that a dispatching rule must accomplish if it is to be effective for high-utilization systems: (a) the policy should use the columns of B (or, more generally, zero reduced cost columns) as much as possible; (b) the policy should try to maintain the backlog "centered" within the cone of B, so it can continue to use the columns of B in the future.

Like GREEDY, at each epoch that a column completes execution, CENTER "prices out" columns to decide which route will be used next. Rather than using the "greedy" dual prices y_t as in (16), however, CENTER uses the dual prices generated by (3), y^*. That is, for each column j, CENTER calculates the effective reduced cost as

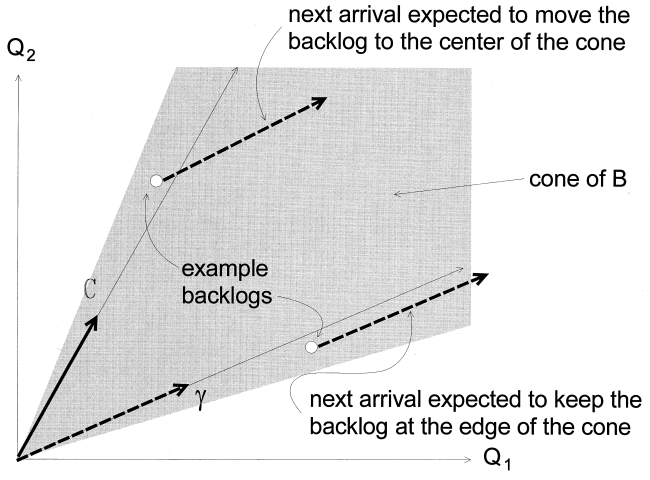
$$\tau_j - y^{*\top} \min\{Q_t, a_j\} \tag{17}$$

and selects the column k that minimizes (17). By using the dual prices from (3), y^*, CENTER increases the likelihood that the optimal basis B will be used. In particular, if the solution to (5) has a unique optimal basis then whenever at least one column, $j \in B$, has $\min\{Q_t, a_j\} = Q_t$, CENTER will choose a column of B.

To ensure that the backlog remains "centered" within the cone of B, CENTER uses a tiebreaking rule that differs from the rules used by the other three heuristics. The rule selects the column which minimizes the L_2 norm to a "centering" ray, $C \in \mathbf{R}_+^m$, that represents a favorable ratio of load types.

Before the simulation begins, the policy uses γ and B to construct C as follows: (a) let $d \in R^m$ equal $B^{-1}\gamma$; (b) let

Figure 1. Examples of the expected effect of an arrival on the backlog.



$e \in R^m$ be the "inverse" of d, where for each element, i, $e_i = 1/d_i$; then (c) let $C = Be$. Thus, if the ray of γ is near a boundary of the cone of B, then C is placed near the *opposite* boundary.

Now suppose that the policy maintains the backlog so that Q_t remains near C. By placing C near the boundary of the cone of B opposite to γ, the probability that an arrival drives the backlog out of the cone of B is reduced (for an example in two dimensions, see Figure 1).

The revised tiebreaking rules for CENTER are as follows: (a), if column k dominates column j then choose column k; (b), if test (a) results in a tie, then if the L_2 norm from the backlog to C after dispatching route k is smaller than that after dispatching route j,

$$\min_{\alpha} \sqrt{\sum_{i=1}^{m} (\alpha C_i - \max\{0, Q_t^i - a_{ik}\})^2} < \min_{\alpha} \sqrt{\sum_{i=1}^{m} (\alpha C_i - \max\{0, Q_t^i - a_{ij}\})^2},$$

then choose column k; otherwise (c), if tests (a) and (b) both result in ties, then if the maximum residual backlog left by column k is less than that left by column j then choose column k; or (d), if tests (a), (b), and (c) all result in ties, then flip a coin. Rules (a), (c), and (d) for CENTER are defined as (a), (b), and (c) were defined for the other heuristics.

7. NUMERICAL ANALYSIS

This section reports the results of three sets of simulation experiments which test the performance of the four heuristics describes in Section 6. The simulations sample the LP lower bound on system work found upon arrival under the heuristics, $\{\underline{W}_{t_k}^{\pi}: k = 1, 2, \ldots\}$, as well as the analogous quantities for the lower bound process defined by the *GI*/*GI*/1 queue described in Section 4.2 (which in this section we will call LOWER).

In each simulation run, we use the method of "batch means" (see Law and Kelton 1982, p. 295–297) with batches of size

$$M \approx 10\, \frac{\lambda^2 \sigma_T^2 + y^{*\top} \Gamma y^* / (y^{*\top} \gamma)^2}{(1-\rho)^2} \tag{18}$$

(see Whitt 1989, p. 1355–1357). For the lower bound and each of the three policies, the sample points of the work process that fall within each batch are averaged. The simulation run terminates when the 95% confidence intervals for the estimates of the population means (the average of the averages) are less than or equal to ±10% of the estimate of the population means themselves. Policies which appear to be unstable—because their population means increase with every new batch—are excluded from the ±10% stopping requirement.

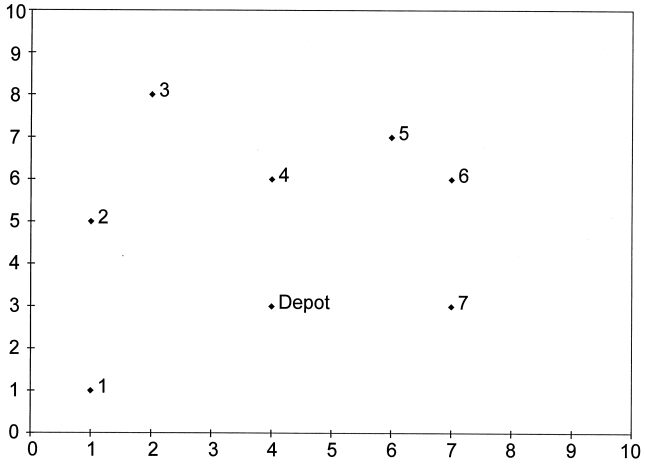
In all simulation runs, we use *i.i.d.* exponential interarrival times. Then by PASTA (see Wolff 1989, p. 293–297) we may interpret the arrival averages calculated by the simulation to be unbiased estimates for the analogous time averages, $E_{\pi}[W]$. Within each of the three sets of simulation experiments, we vary only the mean of the interarrival time distribution T to achieve ρ's of 0.8, 0.9, 0.95, and 0.99.

7.1. Simulation Examples

An Example Load Consolidation Problem. The first two sets of experiments simulate a load consolidation problem. In the problem, four types of loads—"sizes" 51, 26, 12, and 3—arrive at random intervals to a source location and wait for delivery to a destination. A truck of capacity 100 delivers the backlog of loads waiting to be shipped by traveling from the source to the destination and back. The time required to complete the round trip is one. The set of feasible routes is defined by the set of all packings of the truck for which the aggregate size of the loads being packed does not exceed the truck's capacity. While there are 367 feasible packings of the truck, only 30 are not dominated (if $a_j > a_k$ then j dominates k).

An Example One-Warehouse, Multiple-Retailer Problem. The third set of experiments simulates a one-warehouse, multiple-retailer problem. In the problem, there are seven retailers, whose locations are shown in Figure 2. Each of the seven retailers experiences random demand for two types of products, one of size 2 and the other of size 3. Therefore, there are 14 distinct load types enumerated in the set covering model, one for each product-location pair. The warehouse has one truck of capacity 5 with which it replenishes all seven retailers. Again, feasible delivery routes are ones which ship products whose aggregate size does not exceed the truck's capacity. In total there are 91 feasible routes that visit one or two locations, 77 of which are not dominated. In this problem the various routes required different amounts of time to execute. The time required to complete a route equals the length of the TSP tour that visits the route's locations.

Figure 2. One-warehouse, multiple-retailer problem used in simulations.



In all three sets of simulations the WEIGHT policy uses the loads' sizes as their weights when it decides which loads in the backlog to deliver. All other aspects of WEIGHT, as well as the other policies, are implemented as previously described.

Arrival Statistics. For all three examples, exactly one load arrives into the system at an arrival epoch. Therefore, each γ_i equals simply the conditional probability that a type-i load arrives, given there has been an arrival (see Table 1). In addition, the coefficient of variation of each type of load's arriving quantity, calculated as $\sqrt{(1-\gamma_i)/\gamma_i}$, is greater than or equal to 1.5 for all loads except type 4 in Example 1. The fact that only one type arrives at a time also implies that for all examples, the arriving quantities are negatively correlated across types, with correlation for types i and j equal to $-\sqrt{\gamma_i\gamma_j/(1-\gamma_i)(1-\gamma_j)}$.

In Example 1, the arrival probabilities obtain dual prices that are directly proportional to the loads' sizes. In this case, we expect the performance of WEIGHT to be similar to that of CENTER, since both policies effectively use the same primary decision-making rule. In turn, we can attribute differences between the two policies' performances to the effect of their different tiebreaking rules.

The probabilities for Example 2 lead to zero dual prices for loads types 2, 3, and 4. Here, the smaller loads are not "in heavy traffic." That is, as it attempts to minimize the backlogs of the largest loads, the vehicle has ample capacity to deliver the smaller loads. Finally, note that Example 3's arrival probabilities are all of the same order of magnitude. At the same time they are constructed *not* to be perfectly symmetric across types. (See Table 1.)

Table 1. γs used in simulations and their associated y^*s.

Type i	Example 1		Example 2		Example 3	
	γ_i	y_i^*	γ_i	y_i^*	γ_i	y_i^*
1	0.0413	0.5152	0.3	1.0	0.066	3.61
2	0.0810	0.2576	0.2	0.0	0.073	3.61
3	0.1755	0.1212	0.3	0.0	0.067	2.56
4	0.7021	0.0303	0.2	0.0	0.081	3.00
5	—	—	—	—	0.063	2.24
6	—	—	—	—	0.074	3.42
7	—	—	—	—	0.083	3.00
8	—	—	—	—	0.064	3.61
9	—	—	—	—	0.072	3.61
10	—	—	—	—	0.083	8.21
11	—	—	—	—	0.051	3.00
12	—	—	—	—	0.085	6.71
13	—	—	—	—	0.056	5.06
14	—	—	—	—	0.082	3.00

7.2. Simulation Results

The simulation results are summarized in Tables 2 and 3. Note that the CENTER policy consistently outperforms both the two straw policies and the GREEDY policy. Furthermore, as system utilization increases, the LP lower bound on the policy's expected backlog appears to approach the performance of the lower bound process, though the optimality gap in Example 3 is quite wide, even at $\rho = 0.99$.

Table 2 presents the raw simulation results: confidence intervals for LOWER and for the LP lower bounds on expected work under the four policies. Note that for each example, crude *upper bounds* on system performance under all of the policies can be calculated from Table 2 by adding the sum of the m largest τ_j's to the lower bounds. For Examples 1 and 2, this sum equals 4 ($m \times 1$), and for Example 3 it equals 219.2.

With exponential interarrival times, the lower bound process behaves as an $M/G/1$ queue. Therefore, as a check on the validity of the simulation results for the lower bound process, we compare LOWER to its corresponding analytical expectation. Table 2 shows that in all cases the $M/G/1$ expectation falls within the 95% confidence interval of the simulation mean for LOWER.

For each simulation run, Table 3 shows the relative frequency with which each policy uses the columns of the optimal basis B, as well as the relative frequency with which it uses any column whose reduced cost in (5) equals zero ($\tau_j - y^{*\top}a_j = 0$). While the columns of B are always included in this second set, there will be zero-reduced-cost columns which are not a part of B whenever multiple optimal solutions to (5) exist. Indeed, in Examples 1 and 2 the solution of (5) has multiple optimal solutions; for these examples we use this larger set to capture the use of *all* of the columns which efficiently deliver the backlog of system work.

In Example 1, in which the loads' dual prices are proportional to their sizes, the WEIGHT, GREEDY, and CENTER policies all perform well. As utilization increases, all three policies' LP lower bounds continue to converge to LOWER. As noted earlier, CENTER's advantage over WEIGHT can be explained by its "centering" tie-breaking rule, which actively seeks to maintain an advantageous mix of load types in the backlog. GREEDY's ability to dynamically redefine its "preferred" routes, based on the mix of load types in the backlog, evidently allows it to manage the mix of loads in the backlog better than WEIGHT as well.

Table 2. Simulation results for LOWER and the four policies.

Ex	ρ	$M/G/1$[1]	LOWER[2]	NUMBER[3]	WEIGHT[3]	GREEDY[3]	CENTER[3]
1	0.80	0.462	0.458 (±0.036)	2.62 (±0.26)	1.37 (±0.06)	1.33 (±0.05)	1.30 (±0.04)
	0.90	1.04	1.05 (±0.06)	550 (±54.8)	2.32 (±0.12)	2.03 (±0.07)	1.95 (±0.07)
	0.95	2.19	2.32 (±0.23)	—[4]	4.47 (±0.41)	3.38 (±0.25)	3.26 (±0.24)
	0.99	11.4	12.2 (±1.22)	—[5]	20.6 (±1.87)	13.5 (±1.23)	13.1 (±1.22)
2	0.80	2.00	2.16 (±0.19)	3.82 (±0.38)	3.15 (±0.27)	2.80 (±0.20)	2.70 (±0.19)
	0.90	4.50	4.63 (±0.25)	27.5 (±2.74)	9.10 (±0.66)	5.44 (±0.25)	5.18 (±0.25)
	0.95	9.49	9.31 (±0.65)	12,200 (±1,220)	334 (±33.3)	10.41 (±0.67)	9.85 (±0.65)
	0.99	49.5	51.5 (±5.08)	—[4]	—[4]	53.0 (±5.09)	52.0 (±5.08)
3	0.80	9.40	9.07 (±0.68)	79.0 (±7.87)	48.1 (±4.60)	55.1 (±4.43)	46.3 (±3.36)
	0.90	21.2	20.6 (±1.43)	9,790 (±959)	5,570 (±557)	185 (±10.8)	108 (±7.70)
	0.95	44.7	44.1 (±4.39)	—[4]	—[4]	522 (±38.3)	202 (±13.7)
	0.99[6]	232	221 (±23.4)	—[5]	—[5]	—[4]	765 (±68.3)

[1]Solution to the Pollaczek-Khintchine formula for an M/G/1 queue with the same statistics as LOWERs.
[2]95% confidence interval for the steady state expectation of system work.
[3]95% confidence interval for the steady state expectation of the LP lower bound on system work.
[4]System appears to be unstable.
[5]Policy not included in the simulation.
[6]Simulation run terminated after running for 168 hours on an HP 9000/735 work station.

Observe that the fraction of instances in which WEIGHT, GREEDY, and CENTER use zero-reduced-cost columns continues to grow as ρ increases in Example 1, exceeding 94% for all three policies at $\rho = 0.99$. In contrast, the use of these columns stalls at about 65% for NUMBER, and this policy quickly becomes unstable. Also note that fraction of instances in which any of the policies uses the columns of B is quite low, typically under 5% of the total. Indeed, because the dual prices are proportional to the loads' weights, every column in the A-matrix has a reduced cost equal to or close to zero. In total there are 19 columns with reduced costs of zero, only four of which are included in B.

In Example 2, GREEDY and CENTER appear to converge quickly to the lower bound, falling within 1% to 3% of LOWER at $\rho = 0.99$. Furthermore, they are the only policies that remain stable at this utilization. It is not surprising that NUMBER and WEIGHT do not perform well in this case because their weighting schemes do not correspond to the loads' dual prices.

Again, Table 3 shows that as system utilization increases, GREEDY and CENTER increasingly use zero-reduced-cost columns, exceeding 99% at $\rho = 0.99$, while NUMBER's and WEIGHT's use of these columns does not grow consistently with increases in ρ, stalling out at about 90% and 95%, respectively. In Example 2 there are two classes of routes: those that ship a large load and have a reduced cost of zero, and those that do not and have a reduced cost of one. The columns of B represent four of the seven columns of A that have a reduced cost of zero.

In Example 3 the performance of all of the policies is poorest when compared to LOWER. At $\rho = 0.99$ the average lower bound on CENTER's backlog is significantly above LOWER, and the other three policies are unstable. Nevertheless, CENTER's performance, though poor with respect to the lower bound, is strongest relative to that of the other three heuristics. From Table 3 we see that CENTER's use of the optimal basis is high, surpassing 99% at $\rho = 0.99$, while that of the second-best, GREEDY, does not exceed 96%.

The results of Table 3 confirm the insights provided by our heavy traffic analysis: the greater the load on the distribution system, the more critical it becomes to use zero-reduced-cost delivery routes (such as those in the optimal basis B) as much as possible. Indeed, the results of Example 3 show that at high utilization, even a 3% discrepancy in the use of these

Table 3. Relative frequency with which efficient columns are used.

		NUMBER		WEIGHT		GREEDY		CENTER	
Ex	ρ	B	all	B	all	B	all	B	all
1	0.80	3.6%	63.9%	2.7%	60.9%	2.6%	60.9%	3.1%	61.4%
	0.90	3.7%	66.8%	2.9%	72.2%	2.7%	71.9%	3.3%	73.7%
	0.95	3.8%	65.3%	5.2%	81.9%	3.2%	81.4%	3.8%	84.9%
	0.99	—	—	16.4%	95.0%	3.3%	94.3%	4.1%	96.7%
2	0.80	53.7%	83.0%	55.7%	83.6%	57.0%	84.1%	57.7%	83.9%
	0.90	59.8%	90.1%	63.0%	91.0%	65.9%	91.2%	66.7%	91.4%
	0.95	60.5%	90.5%	66.6%	95.0%	71.1%	95.4%	72.1%	95.4%
	0.99	60.2%	89.6%	66.7%	94.4%	76.5%	99.3%	77.0%	99.2%
3	0.80	78.2%	78.2%	81.0%	81.0%	79.6%	79.6%	90.8%	90.8%
	0.90	86.9%	86.9%	87.8%	87.8%	86.2%	86.2%	94.2%	94.2%
	0.95	92.9%	92.9%	91.4%	91.4%	91.6%	91.6%	96.5%	96.5%
	0.99	—	—	—	—	95.7%	95.7%	99.1%	99.1%

"efficient" columns may make the difference between a stable and an unstable system (e.g., Example 3 at $\rho = 0.99$). By this measure, the CENTER policy performs quite well.

8. CONCLUSIONS AND PRACTICAL IMPLICATIONS

We believe that our approach to stochastic distribution problems is promising. It offers the broad modeling flexibility inherent in set-covering formulations and is able to accommodate essentially arbitrarily complex routing constraints. Moreover, set-covering formulations are used frequently in many practical implementations of routing algorithms (see for example Balinsky and Quandt 1964, Ceria et al. 1995, Cullen et al. 1981, Desrochers et al. 1992), so our results fit naturally within the current state of practice. Similarly, we are able to capture a wide array of distribution policies. The only significant restriction is that dynamically altering routes during the course of execution (as is done in Dror et al. 1989 and Jaillet and Odoni 1988, for example) is not allowed.

Our analysis also provides significant insights. The asymptotic analysis succinctly reveals the factors that drive congestion at high system utilization. The vector of dual prices y^* defines a simple stability condition, and they play a central role in characterizing the optimal system work in heavy traffic. Indeed, our analysis demonstrates that as the offered load grows to match the distribution network's capacity (as measured by ρ), the incremental burden placed on the system by the arrival of each load is described exactly by the dual prices y^*.

In addition, the analysis has implications for the design of effective dispatching heuristics. The asymptotic analysis shows that to stabilize and minimize the backlog in high-utilization systems, an effective dispatching policy must (1) consistently use routes whose reduced cost, $\tau_j - y^{*\top} a_j$, is zero (*efficient routes*); and (2) maintain a mix of loads which is roughly centered in the cone of the optimal basis B, so that efficient routes can be used in the future.

These rules are quite practical in a number of ways. First, while the number of feasible routes for a distribution facility n may be large (in fact, it may grow exponentially quickly with the number of load types), the number of *efficient* routes will typically be much smaller, on the order of the number of load types. Moreover, these routes can be generated and evaluated on-line, using the dual prices y^* and column-generation-like techniques. Thus, it is easy to maintain a list of efficient routes. Then, simple dispatching rules, such as those used by CENTER, can be used to select those routes in the list that best direct the backlog toward the center of the cone. Second, computing the input data for these rules is straightforward. To compute the dual prices y^* and find the optimal basis B, one solves the $m \times n$ LP, (3), once. This preprocessing step requires only an estimate of γ, as well as τ and the columns of A. Again, not all n routes need be explicitly enumerated to solve the LP; rather, column generation techniques may be used.

We believe these simple rules can be used to enhance the performance of traditional, static routing algorithms without the need for drastic changes in solution methodology. Often static problems are solved repeatedly, in a rolling horizon fashion, and only a subset of the routes generated by the algorithm are actually dispatched. By estimating the long-run statistics of arriving loads and solving for y^* using (3), one can easily estimate the "time efficiency" of each route using the reduced cost $\tau_j - y^{*\top} a_j$. This estimate of time efficiency can then be combined with traditional cost and service-level criteria to evaluate which routes are most attractive to dispatch. Without such congestion information, however, one runs the risk of poor long-run service performance. Indeed, our simulation results show that simply adding a sensible throughput-oriented criterion (such as NUMBER and WEIGHT) does not even guarantee system stability. The dual information y^* and the concept of centering appear to be critical for controlling congestion.

Finally, the assumption of a single vehicle is not restrictive. Indeed, with m identical vehicles we can consider policies which only dispatch all m vehicles simultaneously on the same route. This effectively increases each element a_{ij} by a factor of m, and hence reduces the resulting y^* by a factor $1/m$. The lower bound can similarly be modified to prove asymptotic optimality in the m-vehicle case. Again, these dual prices can then be used in more practical policies like the CENTER policy to dynamically price out routes to dispatch as each vehicle becomes available. While we have not tested this policy in simulations, we believe its behavior would be similar to that of the single-vehicle case.

APPENDIX: PROOFS OF PROPOSITIONS

The proofs of Propositions 1 and 4 use Chernoff's bounds on the tail probabilities of sums of i.i.d. random variables.

LEMMA 3 (from Chernoff, found in Gans and van Ryzin 1997). *Let X be a random variable with $0 \leq X \leq C$*, a.s., *for some $C < \infty$. Suppose S_n is the sum of n* i.i.d. *samples of X. Then (i) for each $a > E[X]$ there exists a $\theta > 0$ such that $P\{S_n \geq na\} = O(e^{-\theta n})$; (ii) for each $a < E[X]$ there exists a $\theta > 0$ such that $P\{S_n \leq na\} = O(e^{-\theta n})$.*

Proof of Proposition 1

$$
\begin{aligned}
P\{E_l\} &= P\{\hat{V}_l \notin \mathcal{N}(N\gamma, N\epsilon)\} \\
&\leq P\left\{ \exists\, i : \hat{R}^i_{l-1} + \sum_{k=1}^{N} V^i_{N(l-1)+k} - \hat{R}^i_l < N(\gamma_i - \epsilon) \right\} \\
&\quad + P\left\{ \exists\, i : \hat{R}^i_{l-1} + \sum_{k=1}^{N} V^i_{N(l-1)+k} - \hat{R}^i_l > N(\gamma_i + \epsilon) \right\} \\
&\leq \sum_{i=1}^{m} P\left\{ \hat{R}^i_{l-1} + \sum_{k=1}^{N} V^i_{N(l-1)+k} - \hat{R}^i_l < N(\gamma_i - \epsilon) \right\} \\
&\quad + \sum_{i=1}^{m} P\left\{ \hat{R}^i_{l-1} + \sum_{k=1}^{N} V^i_{N(l-1)+k} - \hat{R}^i_l > N(\gamma_i + \epsilon) \right\} \\
&\leq \sum_{i=1}^{m} P\left\{ \sum_{k=1}^{N} V^i_{N(l-1)+k} - q < N(\gamma_i - \epsilon) \right\} \\
&\quad + \sum_{i=1}^{m} P\left\{ \sum_{k=1}^{N} V^i_{N(l-1)+k} + q, > N(\gamma_i + \epsilon) \right\}. \qquad (19)
\end{aligned}
$$

Now for any i, $P\{\Sigma_{k=1}^{N} V^i_{N(l-1)+k} - q < N(\gamma_i - \epsilon)\} = P\{\Sigma_{k=1}^{N} V^i_{N(l-1)+k} < N(\gamma_i - (\epsilon - q/N))\}$, and for $N \geq \lceil 2q/\epsilon \rceil$, $\epsilon - q/N > \epsilon/2$. Then for all $N \geq \lceil 2q/\epsilon \rceil$ we have

$$P\left\{ \sum_{k=1}^{N} V^i_{N(l-1)+k} - q < N(\gamma_i - \epsilon) \right\} \leq P\left\{ \sum_{k=1}^{N} V^i_{N(l-1)+k} < N(\gamma_i - \epsilon/2) \right\}, \quad (20)$$

and the same argument shows that whenever $N \geq \lceil 2q/\epsilon \rceil$

$$P\left\{ \sum_{k=1}^{N} V^i_{N(l-1)+k} + q > N(\gamma_i + \epsilon) \right\} \leq P\left\{ \sum_{k=1}^{N} V^i_{N(l-1)+k} > N(\gamma_i + \epsilon/2) \right\}. \quad (21)$$

Together, Equations (19), (20), and (21) give us

$$P\{E_l\} \leq \sum_{i=1}^{m} P\left\{ \sum_{k=1}^{N} V^i_{N(l-1)+k} < N(\gamma_i - \epsilon/2) \right\} + \sum_{i=1}^{m} P\left\{ \sum_{k=1}^{N} V^i_{N(l-1)+k} > N(\gamma_i + \epsilon/2) \right\} \quad (22)$$

for all $N \geq \lceil 2q/\epsilon \rceil$. Because $E[V^i_k] = \gamma_i$, we may apply the Chernoff bounds of Lemma 3 to complete the proof. □

The proofs of Propositions 2 and 4 also use the following corollary to Proposition 1.

COROLLARY 1. *Given the conditions of Proposition* 1, *there exists a sequence of independent events*, $\{F_l: l = 1, 2, \dots\}$, *such that for any batch*, l, $\{E_l\} \subseteq \{F_l\}$ *and* $P\{F_l\} = O(e^{-\theta N})$.

PROOF. Let $\{F_l\} \stackrel{\text{def}}{=} \{\exists i: \Sigma_{k=1}^{N} V^i_{N(l-1)+k} - q < N(\gamma_i - \epsilon) \cup \Sigma_{k=1}^{N} V^i_{N(l-1)+k} + q > N(\gamma_i + \epsilon)\}$. Then $\{E_l\} \subseteq \{F_l\}$ and the right-hand side of (19) is also an upper bound on $\{F_l\}$. □

Proof of Proposition 2

Recall that for policy π_{q_N}, $\bar{S}_l < NC_3$ *a.s.* and that NC_3 provides an upper bound on the time required to clear the accumulator at any arbitrary epoch. In addition, Lemma 2 can be used to provide a bound on the average time required to clear the backlog at the batch server. Using these bounds together, we find

$$E_{\pi_{q_N}}[W] \leq \frac{E[\hat{S}_l \hat{D}_l]}{E[\hat{T}_l]} + \frac{E[\hat{S}^2]}{2E[\hat{T}_l]} + NC_3 \leq \frac{\lambda E[\bar{S}_l \hat{D}_l]}{N} + \frac{\lambda N^2 C_3^2}{N} + NC_3 = \frac{\lambda E[\bar{S}_l \hat{D}_l]}{N} + O(N). \quad (23)$$

Observe that in Corollary 1 the event $\{F_l\}$ is independent of $\hat{D}_l$, therefore

$$\begin{aligned} E[\bar{S}_l \hat{D}_l] &= E[(y^{*\top} \hat{V}_l + NC_3 \mathbf{1}_{\{E_l\}}) \hat{D}_l] \\ &\leq E[(y^{*\top} \hat{V}_l + NC_3 \mathbf{1}_{\{F_l\}}) \hat{D}_l] \\ &= E[(y^{*\top} \hat{V}_l) \hat{D}_l] + O(Ne^{-\theta N}) E[\hat{D}_l]. \end{aligned}$$

Expanding $\hat{V}_l$ and using $\mathbf{0} \leq \hat{R}_l \leq \mathbf{1}q$, we then find

$$\begin{aligned} E[\bar{S}_l \hat{D}_l] &\leq E\left[y^{*\top} \left(\hat{R}_{l-1} + \sum_{k=1}^{N} V_{(l-1)N+k} - \hat{R}_l \right) \hat{D}_l \right] + O(Ne^{-\theta N}) E[\hat{D}_l] \\ &= Ny^{*\top} \gamma E[\hat{D}_l] + E[y^{*\top} (\hat{R}_{l-1} - \hat{R}_l) \hat{D}_l] + O(Ne^{-\theta N}) E[\hat{D}_l] \\ &\leq Ny^{*\top} \gamma E[\hat{D}_l] + y^{*\top} \mathbf{1}q E[\hat{D}_l] + O(Ne^{-\theta N}) E[\hat{D}_l] \\ &= (Ny^{*\top} \gamma + y^{*\top} \mathbf{1}q + O(Ne^{-\theta N})) E[\hat{D}_l]. \quad (24) \end{aligned}$$

Then substituting the right-hand side of (24) for $E[\bar{S}_l \hat{D}_l]$ in (23) we find

$$\begin{aligned} E_{\pi_{q_N}}[W] &\leq \frac{\lambda (Ny^{*\top} \gamma + y^{*\top} \mathbf{1}q + O(Ne^{-\theta N})) E[\hat{D}_l]}{N} + O(N) \\ &= \left(\rho + \frac{\lambda y^{*\top} \mathbf{1}q}{N} + O(e^{-\theta N}) \right) E[\hat{D}_l] + O(N). \end{aligned}$$

Then for any fixed $\rho < 1$ we can find a finite integer N_1^* so that for all $N \geq N_1^*$, we have $\rho + \lambda y^{*\top} \mathbf{1}q/N + O(e^{-\theta N}) < 1$, and $E_{\pi_q N}[W] \leq E[\hat{D}] + O(N)$. □

Proof of Proposition 3

From Proposition 1 we know that $P\{E_l\} = O(e^{-\theta N})$. Then given stationary $\{\hat{R}_l\}$, we have

$$\begin{aligned} E[\bar{S}_l] &= E\left[y^{*\top} \hat{R}_{l-1} + \sum_{k=1}^{N} y^{*\top} V_{N(l-1)+k} - y^{*\top} \hat{R}_l + NC_3 \mathbf{1}_{\{E_l\}} \right] \\ &= y^{*\top} E[\hat{R}_{l-1}] + \sum_{k=1}^{N} y^{*\top} \gamma - y^{*\top} E[\hat{R}_l] + NC_3 P\{E_l\} \\ &= Ny^{*\top} \gamma + O(Ne^{-\theta N}). \end{aligned}$$

Then for any $\rho = \lambda y^{*\top} \gamma < 1$ there exists an integer, $N_2^* > 0$ such that $\lambda O(e^{-\theta N_2^*}) < (1 - \rho)$ and $E[\bar{S}_l] = Ny^{*\top} \gamma + O(Ne^{-\theta N}) < N/\lambda = E[\hat{T}_l]$ for all $N \geq N_2^*$. □

Proof of Proposition 4

A simple sample-path upper bound on the delay of an arbitrary batch arriving to the batch server, $\hat{D}_l$, is the length of the busy period, B, into which the batch arrived. Then $E[\hat{D}_l^2] \leq E[B^2] = E[(\Sigma_{l=1}^{\hat{n}} \hat{S}_l)^2]$, where the random variable $\hat{n}$ equals the number of batches served in the busy period into which the delayed batch arrives. For batches of size N, the crude bound $\hat{S}_l \leq NC_3$ allows us to write $E[\hat{D}_l^2] \leq N^2 C_3^2 E[\hat{n}^2]$.

Now defining the random variable n to be the number of batches served in an arbitrary busy period, we have

$$P\{\hat{n} = m\} = \frac{mP\{n = m\}}{\sum_{l=1}^{\infty} lP\{n = l\}}$$

so that

$$E[\hat{n}^2] = \sum_{m=1}^{\infty} m^2 P\{\hat{n} = m\} = \frac{E[n^3]}{E[n]}.$$

Thus, if we can show that for any fixed $\rho < 1$, $P\{n = m\} = O(e^{-\theta m})$ for some $\theta > 0$, we will have shown that $E[\hat{n}^2] < \infty$ and in turn that $E[\hat{D}^2] < \infty$ as well. Then

$$\begin{aligned}
P\{n = m\} &\leq P\{n > m - 1\} \leq P\left\{ \sum_{l=1}^{m-1} (\bar{S}_l - \hat{T}_l) > 0 \right\} \\
&= P\left\{ \sum_{l=1}^{m-1} \left(y^{*\top} \left(\hat{R}_{l-1} + \sum_{k=1}^{N} V_{(l-1)N+k} - \hat{R}_l \right) \right.\right. \\
&\qquad \left.\left. + NC_3 \mathbf{1}_{\{E_l\}} - \hat{T}_l \right) > 0 \right\} \\
&= P\left\{ \sum_{l=1}^{m-1} \left(\hat{T}_l - \sum_{k=1}^{N} y^{*\top} V_{(l-1)N+k} - NC_3 \mathbf{1}_{\{E_l\}} \right) \right. \\
&\qquad \left. < y^{*\top} (\hat{R}_0 - \hat{R}_{m-1}) \right\} \\
&\leq P\left\{ \sum_{l=1}^{m-1} \left(\hat{T}_l - \sum_{k=1}^{N} y^{*\top} V_{(l-1)N+k} - NC_3 \mathbf{1}_{\{E_l\}} \right) \right. \\
&\qquad \left. < y^{*\top} \mathbf{1} q \right\}. \qquad (25)
\end{aligned}$$

Now Corollary 1 shows that a sample-path upper bound on the right-hand side of (25) may be constructed by replacing the indicator functions of the events $\{E_1, \ldots, E_{m-1}\}$ by those of the independent events $\{F_1, \ldots, F_{m-1}\}$, which gives us $P\{n = m\} \leq P\{\sum_{l=1}^{m-1} (\hat{T}_l - \sum_{k=1}^{N} y^{*\top} V_{(l-1)N+k} - NC_3 \mathbf{1}_{\{F_l\}}) < y^{*\top} \mathbf{1} q\}$.

Observe that the left-hand side of the inequality inside the bounding probability is the sum of $(m - 1)$ *i.i.d.* random variables $X_l \stackrel{\text{def}}{=} \hat{T}_l - \sum_{k=1}^{N} y^{*\top} V_{(l-1)N+k} - NC_3 \mathbf{1}_{\{F_l\}}$ with expectation $E[X] = N(\lambda - y^{*\top} \gamma - C_3 O(e^{-\theta N}))$. Then we have $P\{n = m\} \leq P\{\sum_{l=1}^{m-1} X_l < y^{*\top} \mathbf{1} q\} = P\{\sum_{l=1}^{m-1} X_l < (m - 1)\, y^{*\top} \mathbf{1} q/(m - 1)\}$.

Finally, we note that there exists an $N_3^* < \infty$ such that $E[X] > 0$ for all $N \geq N_3^*$. For any arbitrary $N \geq N_3^*$, in turn, there exists an $m^* < \infty$ such that for all $m \geq m^*$ we have $0 < y^{*\top} \mathbf{1} q/(m - 1) < a$ for some $a \in (0, E[X])$. Then we write $P\{n = m\} \leq P\{\sum_{l=1}^{m-1} X_l < (m - 1)a\}$. Applying the Chernoff bounds of Lemma 3 we complete the proof. □

Proof of Proposition 5

$$\begin{aligned}
&\operatorname{var}(\bar{S}_1) \\
&= \operatorname{var}\left(\sum_{k=1}^{N} y^{*\top} V_k + y^{*\top} R_0 - y^{*\top} R_N + NC_3 \mathbf{1}_{\{E_1\}} \right) \\
&= \operatorname{var}\left(\sum_{k=1}^{N} y^{*\top} V_k \right) + \operatorname{var}(y^{*\top} R_0) + \operatorname{var}(y^{*\top} R_N) \\
&\quad + \operatorname{var}(NC_3 \mathbf{1}_{\{E_1\}}) + 2 \operatorname{cov}\left(\sum_{k=1}^{N} y^{*\top} V_k, y^{*\top} R_0 \right) \\
&\quad - 2 \operatorname{cov}\left(\sum_{k=1}^{N} y^{*\top} V_k, y^{*\top} R_N \right) \\
&\quad + 2 \operatorname{cov}\left(\sum_{k=1}^{N} y^{*\top} V_k, NC_3 \mathbf{1}_{\{E_1\}} \right) - 2 \operatorname{cov}(y^{*\top} R_0, y^{*\top} R_N) \\
&\quad + 2 \operatorname{cov}(y^{*\top} R_0, NC_3 \mathbf{1}_{\{E_1\}}) - 2 \operatorname{cov}(y^{*\top} R_N, NC_3 \mathbf{1}_{\{E_1\}}).
\end{aligned}$$

We proceed to bound each term of the right-hand side. For the first variance we have $\operatorname{var}(\Sigma_{k=1}^{N} y^{*\top} V_k) = Ny^{*\top} \Gamma y^*$, because the V_ks are independent of each other. For the next two variances we recall that $\{R_k\}$ is stationary so that $\operatorname{var}(y^{*\top} R_0) = \operatorname{var}(y^{*\top} R_N)$. Furthermore, because each R_k is bounded by $q - 1$, and y^* is nonnegative, $\operatorname{var}(y^{*\top} R_0) \leq E[(y^{*\top} R_0)^2] \leq (y^{*\top} \mathbf{1} q)^2$. Similarly, for the fourth covariance term, we have $|\operatorname{cov}(y^{*\top} R_0, y^{*\top} R_N)| \leq (y^{*\top} \mathbf{1} q)^2$, because again, both R_0 and R_N are bounded by $q - 1$. For the last variance we see that $\operatorname{var}(NC_3 \mathbf{1}_{\{E_1\}}) = N^2 C_3^2 \operatorname{var}(\mathbf{1}_{\{E_1\}}) \leq N^2 C_3^2 P\{E_1\} = O(N^2 e^{-\theta N})$. For the first covariance term, we have $\operatorname{cov}(\Sigma_{k=1}^{N} y^{*\top} V_k, y^{*\top} R_0) = 0$, because arriving quantities are independent of previous remainder terms. The third covariance term yields

$$\begin{aligned}
&\left| \operatorname{cov}\left(\sum_{k=1}^{N} y^{*\top} V_k, NC_3 \mathbf{1}_{\{E_1\}} \right) \right| \\
&\leq NC_3 \sum_{k=1}^{N} \sum_{i=1}^{m} y_i^* |\operatorname{cov}(V_k^i, \mathbf{1}_{\{E_1\}})| \\
&\leq NC_3 \sum_{k=1}^{N} \sum_{i=1}^{m} y_i^* E[V_k^i \mathbf{1}_{\{E_1\}}] \leq N^2 C_1 C_3 y^{*\top} \mathbf{1} P\{E_1\} \\
&= O(N^2 e^{-\theta N}),
\end{aligned}$$

because each V_k^i is bounded above by C_1. The same analysis for the last two covariance terms yields $|\operatorname{cov}(y^{*\top} R_0, NC_3 \mathbf{1}_{\{E_1\}})| = O(Ne^{-\theta N})$, and $|\operatorname{cov}(y^{*\top} R_N, NC_3 \mathbf{1}_{\{E_1\}})| = O(Ne^{-\theta N})$. Finally, we bound the second covariance term. We have

$$\begin{aligned}
&\left| \operatorname{cov}\left(\sum_{k=1}^{N} y^{*\top} V_k, y^{*\top} R_N \right) \right| \\
&\leq \sum_{k=1}^{N} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i^* y_j^* |E[V_k^i R_N^j] - E[V_k^i] E[R_N^j]| \\
&\leq \sum_{k=1}^{N} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i^* y_j^* \sum_{s=0}^{C_1} \sum_{t=0}^{q-1} s \cdot t |P\{V_k^i = s \cap R_N^j = t\} \\
&\qquad - P\{V_k^i = s\} P\{R_N^j = t\}|
\end{aligned}$$

$$\leqslant \sum_{k=1}^{N} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i^* y_j^* \sum_{s=0}^{C_1} \sum_{t=0}^{q-1} s \cdot t \alpha \beta^{N-k} P\{V_k^i = s\},$$

because $V_k \in \mathcal{F}_0^k$, $R_N \in \mathcal{F}_N^\infty$, and $\{(V_k, R_k)\}$ is φ-mixing with $\varphi_k = \alpha\beta^k$. Then using the trivial bound $P\{V_k^* = s\} \leq 1$ we have

$$\left|\operatorname{cov}\left(\sum_{k=1}^{N} y^{*\top} V_k, y^{*\top} R_N\right)\right| \leq \alpha(y^{*\top}\mathbf{1})^2 C_1 q \sum_{k=1}^{N} \beta^{N-k} \leq \frac{\alpha(y^{*\top}\mathbf{1})^2 C_1 q}{1-\beta}.$$

Collecting the bounds on the absolute values of all terms except for the first variance, we note that we can find a $C_2 > 0$ and an integer $N_4^* < \infty$ such that for all $N \geq N_4^*$ the sums of the absolute values of the terms are less than C_2. Then for all $N \geq N_4^*$ we have $|\operatorname{var}(\bar{S}_1) - Ny^{*\top}\Gamma y^*| \leq C_2$. This completes the proof. □

Proof of Theorem 2

Part (i). Suppose there exists a unique optimal solution, y^*, to (3). Then from Proposition 3, Proposition 5, and (9) we know that for any fixed $\rho = \lambda y^{*\top}\gamma < 1$ we can find an $N^* \stackrel{\text{def}}{=} \max\{N_1^*, N_2^*, N_3^*, N_4^*\}$ such that for all batch sizes $N \geq N^*$ (10) is well-defined and the system is stable. In turn, from (10) and Proposition 2 we have

$$\begin{aligned} E_{\pi_{q_N}}[W] &\leq \frac{\operatorname{var}(\bar{S}_l) + \operatorname{var}(\hat{T}_l)}{2(E[\hat{T}_l] - E[\bar{S}_l])} + \frac{E[\hat{T}_l] - E[\bar{S}_l]}{2} + O(N) \\ &\leq \frac{Ny^{*\top}\Gamma y^* + C_2 + N\sigma_T^2}{2\left(\frac{N}{\lambda} - Ny^{*\top}\gamma + O(Ne^{-\theta N})\right)} + \frac{N(1-\rho)}{2\lambda} + O(N) \\ &= \frac{\lambda(y^{*\top}\Gamma y^* + \sigma_T^2)}{2(1-\rho-\lambda O(e^{-\theta N}))} + \frac{C_2}{N(1-\rho-\lambda O(e^{-\theta N}))} + O(N). \end{aligned}$$

Part (ii): From part (i) we have

$$\begin{aligned} (1-\rho)E_{\pi_{q_N}}[W] &\leq \frac{\lambda(y^{*\top}\Gamma y^* + \sigma_T^2)}{2} \times \left[\frac{1-\rho}{1-\rho-\lambda O(e^{-\theta N})}\right] + \frac{C_2}{N} \\ &\quad \times \left[\frac{1-\rho}{1-\rho-\lambda O(e^{-\theta N})}\right] + (1-\rho)O(N). \end{aligned}$$

Then setting $N \stackrel{\text{def}}{=} \lceil (1-\rho)^{-b} \rceil$ for some fixed $b \in (0, 1)$ and letting $\rho \to 1$ by scaling T so that $\lambda \to 1/y^{*\top}\gamma$, we have

$$\lim_{\rho\to 1}\left[\frac{1-\rho}{1-\rho-\lambda O(e^{-\theta N})}\right] = 1,$$

$$\lim_{\rho\to 1}\frac{C_2}{N} \times \left[\frac{1-\rho}{1-\rho-\lambda O(e^{-\theta N})}\right] = 0,$$

and $\lim_{\rho\to 1}(1-\rho)O(N) = 0$. This completes the proof. □

ACKNOWLEDGMENT

The research of both authors was supported by internal grants from the Columbia University Graduate School of Business and the Wharton School of the University of Pennsylvania.

REFERENCES

Balinsky, M. L., R. E. Quandt. 1964. On a large integer program for a delivery problem. *Oper. Res.* **12** 300–304.

Ballou, R. H. 1985. *Business Logistics Management*. Prentice Hall, Englewood Cliffs, NJ.

Bambos, N., J. Walrand. 1993. Scheduling and stability aspects of a general class of parallel processing systems. *Advances in Appl. Probab.* **25** 176–202.

Bazaraa, M. S., J. Jarvis, H. D. Sherali. 1990. *Linear Programming and Network Flows*. John Wiley & Sons, New York.

Bertsimas, D. J., D. Simchi-Levi. 1996. A new generation of vehicle routing research: robust algorithms, addressing uncertainty. *Oper. Res.* **44** 286–304.

——, G. J. van Ryzin. 1991. A stochastic and dynamic vehicle routing problem in the Euclidean plane. *Oper. Res.* **39** 601–615.

——, ——. 1993a. Stochastic and dynamic vehicle routing in the Euclidean plane: the multiple-server, capacitated vehicle case. *Oper. Res.* **41** 60–76.

——, ——. 1993b. Stochastic and dynamic vehicle routing with general interarrival and service time distributions. *Adv. in Appl. Probab.* **25** 947–978.

Billingsley, P. 1968. *Convergence of Probability Measures*. John Wiley & Sons, New York.

Ceria, S., P. N. A. Sassano. 1996. A Lagrangian-based heuristic for large-scale set covering problems. *Math. Programming* (Forthcoming).

Coffman, E. G. Jr., A. A. Puhalskii, M. I. Reiman. 1995a. Polling systems with zero switchover times: a heavy-traffic averaging principle. *Ann. Appl. Probab.* **5** 681–719.

——, ——, ——. 1998. Polling systems in heavy traffic: a Bessel process limit. *Math. Oper. Res.* **23** 257–304.

Cullen, F., J. Jarvis, D. Ratliff. 1981. Set partitioning based heuristics for interactive routing. *Networks* **11** 125–143.

Daley, D. J., A. Ya. Kreinin, C. D. Trengrove. 1992. Inequalities concerning the waiting time in single server queues: a survey. U. N. Bhat, I. V. Basawa, eds. *Queueing and Related Models*. Clarendon Press, Oxford, UK.

Desrochers, M., J. Desrosiers, M. Solomon. 1992. A new optimization algorithm for the vehicle routing problem with time windows. *Oper. Res.* **40** 342–354.

Dror, M., G. Laporte, P. Trudeau. 1989. Vehicle routing with stochastic demands: properties and solution frameworks. *Transp. Sci.* **23** 166–176.

Gans, N. F., G. J. van Ryzin. 1997. Optimal control of a multiclass, flexible queueing system. *Oper. Res.* **45** 677–693.

Giles, F. R., W. R. Pulleyblank. 1979. Total dual integrality and integer polyhedra. *Linear Algebra and its Appl.* **25** 191–196.

Higginson, J. K., J. H. Bookbinder. 1994a. Markovian decision processes in shipment consolidation. *Trans. Sci.* **29** 242–255.

——, ——. 1994b. Policy recommendations for a shipment-consolidation problem. *J. Bus. Logist.* **15** 87–111.

Jaillet, P., A. Odoni 1988. The probabilistic vehicle routing problem. B. L. Golden, A. A. Assad, eds. *Vehicle Routing: Methods and Studies*. Elsevier North-Holland, Amsterdam.

Kleinrock, L. 1976. *Queueing Systems*. Vol. 2, John Wiley & Sons, New York.

Law, A. M., W. D. Kelton. 1982. *Simulation Modeling and Analysis*. McGraw-Hill, New York.

Minkoff, A. S. 1993. A Markov decision model and decomposition heuristic for dynamic vehicle dispatching. *Oper. Res.* **41** 77–90.

Powell, W. B. 1995. A stochastic formulation of the dynamic assignment problem, with an application to truckload motor carriers. *Trans. Sci.* **30** 195–219.

——, P. Humblet. 1986. The bulk service queue with a general control strategy: theoretical analysis and a new computational procedure. *Oper. Res.* **34** 267–275.

——, P. Jaillet, A. Odoni. 1995. Stochastic and dynamic networks and routing. Ball et al., eds. *Network Routing*, Chapter 3. Handbooks in OR & MS, vol. 8. Elsevier North-Holland, Amsterdam.

Psaraftis, H. N. 1988. Dynamic vehicle routing problems. B. L. Golden, A. A. Assad, eds. *Vehicle Routing: Methods and Studies*. Elsevier North-Holland, Amsterdam.

—— 1995. Dynamic vehicle routing: status and prospects. *Ann. Oper. Res.* **61** 143–164.

Reiman, M. I., R. Rubio, L. M. Wein. 1996. Heavy traffic analysis of the dynamic stochastic inventory-routing problem. *Trans. Sci.* (Forthcoming).

Szczotka, W. 1990. Exponential approximation of waiting time and queue size for queues in heavy traffic. *Adv. in Appl. Probab.* **22** 230–240.

Whitt, W. 1989. Planning queueing simulations. *Management Sci.* **35** 1341–1366.

Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs, NJ.