Evaluation of Base-Stock Policies in Multiechelon Inventory Systems with State-Dependent Demands
Part I: State-Independent Policies

Jing-Sheng Song
Graduate School of Management, University of California, Irvine, California 92717

Paul H. Zipkin
Graduate School of Business, Columbia University, New York, New York 10027

This article analyzes a model of a multiechelon inventory system. The exogenous demands form Markov-modulated Poisson processes. That is, the demand rates are functions of an underlying Markov chain. Each location follows a base-stock policy which is independent of the state of the underlying Markov chain. We employ the exogenous transit mechanism introduced by Zipkin [7] and Svoronos and Zipkin [6]. The transit times between locations have phase-type distributions. An exact procedure to compute steady-state performance measures is presented. © 1992 John Wiley & Sons, Inc.

1. INTRODUCTION

This article presents a model of a multiechelon inventory system, which is similar to that of Svoronos and Zipkin [6] and Zipkin [10]. Specifically, the system consists of several facilities or locations whose supply-demand relationships form a hierarchy: Each location places orders for replenishment with one other location, its direct predecessor. There is a single location at the highest level of the hierarchy, called the central depot, whose orders go to an outside supply source.

Thus, if we form a graph describing the system, with a node for each location and arcs representing precedence, then this graph is a rooted tree. The central depot corresponds to the root node. At the lowest level of the hierarchy are the leaves of the tree, where exogenous demands occur.

What is novel in our model is that the demands at the leaves form Markov-modulated Poisson processes. (Svoronos and Zipkin [6] assume Poisson demands, and Zipkin [10] studies compound-Poisson demands.) That is, at each leaf, the demand process forms a Poisson process with a rate depending on an underlying Markov chain. We assume all the leaves share a common underlying Markov chain.

Copyright © 1992 by John Wiley & Sons, Inc. CCC 0894-069X/92/050715-14$04.00
Feeney and Sherbrooke [2] analyzed data on the demand for aircraft engines. They found that the variance of demand over an interval of time was greater than the mean. This finding casts doubt on the most common demand model, the Poisson process. This was one reason they promoted compound-Poisson processes. The Markov-modulated Poisson process offers an alternative model, whose variance also exceeds its mean. Also, the results for this model are reasonably tractable. The choice between this and the compound-Poisson model is thus a modeling issue: Which story is more compelling in a specific application?

Burman and Smith [1] describe some examples which occur in data communications: There is a fixed number of users of the system, each of them generating a Poisson arrival stream. However, the number of active users varies over time. They modeled the number of active users as a finite-state continuous-time Markov chain. Hence the overall arrival process forms a Markov-modulated Poisson process. Similar factors affect the demand for many products. Also, demands are often responsive to general economic conditions, and we can use a continuous-time Markov chain to model the general economy.

Song and Zipkin [5] study single-location discounted-cost models with Markov-modulated Poisson demands. They show that the optimal policy for a model with linear costs is a base-stock policy, where the base-stock level is a function of the underlying Markov chain.

It is likely that an optimal policy for a multiechelon system is similarly state dependent. For simplicity (both for evaluation and implementation), we assume here, in Part I, that each location follows a base-stock, or $(S - 1, S)$, or one-for-one replenishment policy, independent of the underlying Markov chain. Such a policy is described by a single integer parameter $S \geq 0$, the base-stock level, for each location. In Part II we will study a two-echelon system, where the depot follows a state-dependent policy.

Suppose the system starts fully stocked; that is, each location’s inventory equals its base-stock level. Then, every demand at a leaf causes an order to be placed immediately at the leaf’s predecessor. This order becomes a demand at the predecessor, and so an order is placed immediately with its predecessor, etc. Thus, every demand at a leaf results in an immediate demand at all predecessors of the leaf, up to the central depot, and from there to the outside source.

When a demand occurs at any location, if the location has stock available, it releases a unit; otherwise, the demand is placed on backorder. If a resupply unit arrives at a location when there are backorders outstanding, the unit is released to fill the oldest outstanding backorder; if there are no backorders, the unit is added to inventory.

The release of a unit by a leaf results in an exogenous demand being filled, either immediately or after a delay on backorder. For higher-level nodes, however, a released unit arrives at its destination (a successor node) only after a transit time due to transportation or processing activities. These transit times may be stochastic.

We view the outside source as having ample stock, so it always releases a unit immediately in response to a demand from the central depot. However, there are transit times between the source and the depot.

The transit times along each arc are assumed to be generated by an exogenous transit mechanism. Such mechanisms are described in detail by Svoronos [4],
Svoronos and Zipkin [6], and Zipkin [7]. This is a model of production or transportation activities, where orders are processed sequentially and so do not cross in time, and where our orders comprise a negligible fraction of the total traffic or workload. This assumption implies that, in a steady-state analysis, we need to consider only the marginal, equilibrium transit-time distribution. Also, the transit times are independent of our demands and orders.

The sequential-processing assumption differentiates this model from the one used in most prior work, where the transit times on each arc are assumed to be constant or, more generally, i.i.d. random variables. See Svoronos and Zipkin [6] and Zipkin [9] for a detailed discussion.

We define notation in Section 2. Section 3 examines a model with a single location. We present an exact analysis, assuming the transit times have a (continuous) phase-type distribution. Section 4 describes a procedure for multiechelon systems, using the single-location method as a building block.

2. NOTATION

Here is the notation we use for the data of the model:

\[ A = \{A(t) : t \geq 0\} \]

= underlying Markov chain, assumed to be ergodic,

\[ Q = [q_{ab}] = \text{infinitesimal generator of } A, \]

\[ \pi = \text{stationary probability density of } A, \]

\( i, j = \) node indices,

\[ \text{Pre}(j) = \text{immediate predecessor of node } j, \]

\[ \text{Suc}(i) = \text{set of immediate successors of node } i, \]

\[ \bar{\lambda}_i(a) = \text{exogenous demand rate at node } i \text{ in state } a, \]

\[ \lambda_i(a) = \text{total demand rate at node } i \text{ in state } a \]

\[ = \begin{cases} \bar{\lambda}_i(a), & \text{Suc}(i) = \emptyset; \\ \sum_{j \in \text{Suc}(i)} \lambda_j(a), & \text{Suc}(i) \neq \emptyset, \end{cases} \]

\[ \Lambda_i = \text{diag}(\lambda_i(a))_a, \]

\[ T_i = \text{transit time before node } i, \text{ the time from the release of a unit by } \text{Pre}(i) \text{ until the receipt of the unit at } i. \]

We assume that each \( T_i \) has a continuous phase-type distribution. Specifically, we model \( T_i \) with reference to a continuous-time, finite-state, absorbing Markov chain \( H_i = \{H_i(t), t \geq 0\} \): \( T_i \) is the time until \( H_i \) is absorbed. When we talk
about the "states of $H_i$" without qualification, we shall mean the transient states only. The row vector $v_i$ specifies the initial probabilities of the (transient) states, and the matrix $U_i$ is the generator of $H_i$ (restricted to the transient states). We write $T_i \sim \text{CPH}(v_i, U_i)$ to denote a continuous phase-type distribution with these parameters.

The decision variables are

$$S_i = \text{base-stock level at node } i.$$ We assume $S_i \geq 0$.

The following denote random variables describing the equilibrium behavior of the inventory system, which depend on the data of the model and the choice of the $S_i$:

$$D_{ij} = \text{delay after node } i, \text{ the time from the order of a unit by some } j \in \text{Suc}(i) \text{ until the release of that unit by } i,$$

$$L_j = \text{total lead time at node } j = D_{ij} + T_j.$$ Denote

$$IN_i = \text{inventory at node } i$$

$$B_i = \text{backorders at node } i$$

$$K_i = \text{kena (outstanding orders) at node } i$$

$$= S_i - IN_i + B_i.$$ For any number $x$, $[x]^+ = \max\{x, 0\}$. We use $\mathbf{e}$ to denote a column vector of ones; its dimension will be clear from the context. Also, $I$ is used to denote the identity matrix, again of appropriate dimension.

3. SINGLE LOCATION MODEL

In this section we analyze a single-location system. Accordingly, we drop the node index $i$.

To summarize the model, there is an underlying continuous-time Markov chain $A$, which is ergodic with stationary density $\pi$. Given $A(t) = a$, the demand process forms a Poisson process with rate $\lambda(a)$. We follow a base-stock policy with base-stock level $S \geq 0$. Orders are filled after a random transit time $T$; here, this is the full lead time, i.e., $L = T$. We assume $T \sim \text{CPH}(v, U)$, and $v \mathbf{e} = 1$, so $T > 0$ with probability 1. Also, we write $L \sim \text{CPH}(\gamma, \Gamma)$, and let $Z = \{Z(t), t \geq 0\}$ be the Markov chain whose time to absorption has the distribution of $L$. Since $L = T$, we have $\gamma = v, \Gamma = U$ and $Z = H$. We introduce this extra notation to simplify our discussion below of multiechelon models.

The steady-state behavior of the system is characterized by the random variables $IN, B$ and $K = S - IN + B$. Also, the delay $D$ means the time a demand must wait on backorder.
The variable $K$, it turns out, has the same distribution as the lead time demand. This result can be demonstrated using standard inventory-theoretic arguments, as in Zipkin [7]. Using a similar approach as in Part 5 of Zipkin [8], we can show that $K$ has a discrete phase-type distribution, which we write

$$K \sim \text{DPH}(\omega, \Omega).$$

Here,

$$\Omega = [I \otimes \Lambda - I \otimes Q - \Gamma \otimes I]^{-1}(I \otimes \Lambda), \quad \omega = (\gamma \otimes \pi)\Omega.$$

Since $B = [K - S]^+$, it follows (see Neuts [3, p. 47]) that $B \sim \text{DPH}(\omega \Omega^S, \Omega)$. Thus,

$$E[K] = \omega(I - \Omega)^{-1}e,$$
$$E[B] = \omega(I - \Omega)^{-1}\Omega^S e,$$

Before proceeding further, we establish two lemmas which we will use frequently in the rest of the article.

**Lemma 3.1:** Let $F(t), G(t), M_1, M_2$ be matrices with appropriate dimensions satisfying the following differential equations and boundary conditions:

$$\frac{d}{dt} F(t) = F(t)M_1, \quad t \geq 0,$$
$$\frac{d}{dt} G(t) = F(t)M_2, \quad t \geq 0,$$
$$F(0) = I, \quad G(0) = 0.$$

Suppose $M_1$ is nonsingular and $\lim_{t \to \infty} F(t) = 0$. Then

$$G = \lim_{t \to \infty} G(t) = -M_1^{-1}M_2.$$

**Proof:** Observe $-I = \int_0^\infty \frac{d}{dt} F(t) \ dt = \int_0^\infty F(t) \ dt \ M_1$. So, $\int_0^\infty F(t) \ dt = -M_1^{-1}$. Thus, $G = \int_0^\infty \frac{d}{dt} G(t) \ dt = \int_0^\infty F(t) \ dt \ M_2 = -M_1^{-1}M_2$. \qed

Let $t_1$ be the first demand epoch, and $\{N(t): t \geq 0\}$ be the demand process. Let the discrete-time Markov chain $W = \{W[n]: n \geq 0\}$ be the process $A$ embedded at demand epochs.
LEMMA 3.2. The chain \( W \) has one-step transition matrix
\[
\Phi = (\Lambda - Q)^{-1}\Lambda
\]
and equilibrium probability vector
\[
\alpha = (\pi \Lambda e)^{-1}\pi \Lambda = \left( \frac{\pi(a)\Lambda(a)}{\sum_b \pi(b)\Lambda(b)} \right)_a.
\]

PROOF: Define
\[
\Psi_{ab}(t) = \Pr\{N(t) = 0, A(t) = b| N(0) = 0, A(0) = a\},
\]
\[
\Phi_{ab}(t) = \Pr\{t_1 \leq t, A(t_1) = b| N(0) = 0, A(0) = a\},
\]
\[
\Psi(t) = [\Psi_{ab}(t)], \quad \Phi(t) = [\Phi_{ab}(t)].
\]
Then, for \( t \geq 0 \), we have
\[
\Psi'(t) = \Psi(t)[Q - \Lambda],
\]
\[
\Phi'(t) = \Psi(t)\Lambda,
\]
and initial conditions \( \Psi(0) = I \otimes I, \Phi(0) = 0 \otimes 0 \). Applying Lemma 3.1, the transition matrix of \( W \) is
\[
\Phi = \lim_{t \to \infty} \Phi(t) = (\Lambda - Q)^{-1}\Lambda.
\]
Now, we verify that \( \alpha \) is the equilibrium distribution corresponding to \( \Phi \). First, it is obvious that \( \alpha \) is a probability vector. Now, recall that \( \pi Q = 0 \). So,
\[
\alpha \Phi = (\pi \Lambda e)^{-1}\pi \Lambda[(\Lambda - Q)^{-1}\Lambda]
= (\pi \Lambda e)^{-1}[(\pi(\Lambda - Q))[\Lambda - Q]^{-1}\Lambda]
= (\pi \Lambda e)^{-1}\pi \Lambda = \alpha.
\]

Next we turn to the analysis of the customer delay \( D \). Define \( \tau_S(n) = \) time between demands \( n \) and \( n + S \), \( L(n) = \) lead time initiated by the \( n \)th demand, \( D(n + S) = \) delay experienced by demand \( n + S \).
Suppose we start with \( S \) as the initial inventory. Then the order initiated by the \( n \)th demand will be used to fill the \( (n + S) \)th demand. So
\[
D(n + S) = [L(n) - \tau_S(n)]^+.
\]
Letting \( \tau_S \) denote the steady-state time between \( S \) demands, we have
\[
D = [L - \tau_S]^+.
\]
Recall $Z = \{Z(t), t \geq 0\}$ is the Markov-chain representation of the lead time $L$. Evidently, $D$ is the remaining time until $Z$ is absorbed after the random time $\tau_3$. Let $\rho(t)$ denote the density of $Z(t)$, and suppose we can determine $\rho(\tau_3)$. Then, $D \sim \text{CPH}(\rho(\tau_3), \Gamma)$. That is, $D$ results from restarting the process $Z$ with an initial density equal to the density $\rho(\tau_3)$. Also, $\Pr\{D = 0\}$ is precisely the defect of $\rho(\tau_3)$, i.e., $1 - \rho(\tau_3) e$. Our problem, then, is to determine $\rho(\tau_3)$.

Consider the discrete-time Markov chain $(X, W) = \{(X[n], W[n]), n \geq 0\}$, where $X[n]$ and $W[n]$ are just $Z(t)$ and $A(t)$ observed at the $n$th demand epoch, respectively, and the 0th demand is chosen as a typical demand. From the above analysis, $\rho(\tau_3)$ is precisely the marginal density of $X[S]$, assuming that $X[0] \sim Z(0) \sim \gamma$ and $W[0] \sim A(0) \sim \alpha$. So the key step is to find the one-step transition matrix of $(X, W)$, say $V$.

Recall $t_1$ is the first demand epoch, and write

$$V_{uavb} = \Pr\{Z(t_1) = v, A(t_1) = b | Z(0) = u, A(0) = a\}.$$ 

Then, $V = [V_{uavb}]_{uavb}$. Observe

$$V = \lim_{t \to \infty} V(t),$$

where

$$V_{uavb}(t) = \Pr\{t_1 \leq t, Z(t_1) = v, A(t_1) = b | Z(0) = u, A(0) = a\}$$

and

$$V(t) = [V_{uavb}(t)]_{uavb}.$$ 

Our task is then to analyze $V_{uavb}(t)$ and its limit. For this purpose, define for $t \geq 0$,

$$R_{uavb}(t) = \Pr\{Z(t) = v, A(t) = b, N(t) = 0 | Z(0) = u, A(0) = a, N(0) = 0\},$$

$$R(t) = [R_{uavb}(t)]_{uavb}.$$ 

Then using the generator of the joint process, we can write differential equations in these quantities:

$$R(0) = I \otimes I, \quad V(0) = 0 \otimes 0,$$

$$R'(t) = R(t)[\Gamma \oplus Q - I \otimes \Lambda],$$

$$V'(t) = R(t)(I \otimes \Lambda).$$
Applying Lemma 3.1, we obtain

\[ V = -[\Gamma \oplus Q - I \otimes \Lambda]^{-1}(I \otimes \Lambda) \]

\[ = [I \otimes \Lambda - \Gamma \otimes I - I \otimes Q]^{-1}(I \otimes \Lambda) \]

\[ = \Omega. \] (4)

Thus, we have actually shown that the one-step transition matrix of the chain \((X, W)\) is exactly the matrix \(\Omega\) which we use to compute \(E[K]\) and \(E[B]\) in (1). The following result is immediate.

**PROPOSITION 3.3:** Let \(a(n)\) be the density of \((X[n], W[n])\) restricted to transient states of \(Z\). Then

\[ a(0) = \gamma \otimes \alpha, \]

\[ a(n) = a(0)\Omega^n. \]

Observe

\[ \rho(\tau) = \sigma(S)(I \otimes e) = (\gamma \otimes \alpha)\Omega^S(I \otimes e) \]

is the marginal density of \(X\) of the joint chain \((X, W)\) after \(S\) steps, and hence the density of \(Z(\tau)\). We have shown the following proposition.

**PROPOSITION 3.4:** \(D \sim CPH[(\gamma \otimes \alpha)\Omega^S(I \otimes e), \Gamma]\). Thus,

\[ E[D] = -(\gamma \otimes \alpha)\Omega^S(I \otimes e)\Gamma^{-1}e. \]

Let \(m\) be the number of states of the process \(A\) and \(l\) the number of transient states of the process \(Z\). Then, to obtain \(\Omega\) we need to invert an \(ml \times ml\) matrix, which requires \(O(ml^3)\) arithmetic operations. Now, \(\Omega\) is itself an \(ml \times ml\) matrix. As we know, the multiplication of two matrices of this size needs \(O(m^3l^3)\) operations.

It is evident that the complexity of computing \(E[D]\) for a single, large \(S\) is dominated by the matrix multiplications \(\Omega^S\), that is \(O(\log(S)m^3l^3)\). The term \(\log(S)\) is the number of matrix multiplications needed to compute \(\Omega^S\). (We can compute \(\Omega^2 = \Omega \Omega, \Omega^4 = \Omega^2 \Omega^2, \) etc.)

For smaller \(S\) we can compute the vector \((\gamma \otimes \alpha)\Omega^S\) recursively; this requires \(O(Sm^2l^2)\) steps. Together with the computation of \(\Omega\), the complexity becomes \(O(Sm^2l^2) + O(m^3l^3)\).

Generally speaking, there is no useful special structure of \(\Omega\) to simplify the computation. However, \(\Omega\) inherits any special structure of the matrix \(\Gamma\). For instance, if \(\Gamma\) is triangular, then \(\Omega\) is block triangular. This block structure can be exploited in computing \(\Omega\) and powers of \(\Omega\); Neuts [3] discusses such computational issues.
4. MULTIECHELON MODEL

Now we consider the multiechelon model. Recall, we assume the same state-process \( A \) for each of the nodes at the lowest echelon. Also, each node follows a state-independent base-stock policy. Then, demand at each node in the upper echelons is also a Markov-modulated Poisson process with this same state process \( A \).

Our basic approach is to use the single-location model recursively to analyze each node, starting with the root node and working down toward the leaves, as in Svoronos and Zipkin [6] and Zipkin [10]. The link between a node \( i \) and its successor \( j \in \text{Suc}(i) \) is provided by the identity

\[
L_j = D_{ij} + T_f.
\]

Starting with \( L_i \), we analyze node \( i \) to characterize \( D_{ij} \); this gives us \( L_j \), so we can continue the analysis at node \( j \). (Since \( E[B_i] \) and \( E[I_N] \) can be derived from \( E[D_f] \), we omit a direct analysis of backorders.)

Note, Lemma 3.2 applies to each node \( i \) in the network, with \( \Lambda_i \) replacing \( \Lambda \). Correspondingly, we shall use \( \Phi_i \) and \( \alpha_i \) to denote the matrix \( \Phi \) and the vector \( \alpha \), respectively.

4.1. Proportional-Rate Case

In this subsection, we assume there exist \( \mu_i, \kappa_a \) such that

\[
\bar{\lambda}(a) = \mu_i \kappa_a.
\]

For example, \( \mu_i \) may represent the individual demand rate at location \( i \), while \( \kappa_a \) may represent the number of the active users; or, \( \mu_i \) may be the initial demand rate and \( \kappa_a \) represents a fraction or multiple of the initial demand rate. As we shall see in subsection 4.3, in this case \( D_{ij} \) is independent of \( j \). (So, we shall drop the second subscript \( j \) from the notation \( D_{ij} \) in this subsection.) Consequently, the analysis becomes much simpler.

We use \( n_i \) to index the demands at node \( i \) and define

\[
\tau_S(n_i) = \text{time between demands } n_i \text{ and } n_i + S_i,
\]

\[
L_i(n_i) = \text{lead time initiated by demand } n_i,
\]

\[
D_i(n_i + S_i) = \text{delay of demand } n_i + S_i.
\]

As in the single-location model, we have

\[
D_i(n_i + S_i) = [L_i(n_i) - \tau_S(n_i)]^+.
\]

and in equilibrium,

\[
D_i = [L_i - \tau_S]^+.
\]

Our task, then, is to compute the distribution of \( D_i \) for each node \( i \).
Consider a pair of nodes $i$ and $j$ with $j \in \text{Suc}(i)$. If $i$ is the root node, we can analyze $D_i$ as in the single-location model. In general, we shall assume $D_i$ has been analyzed and use this information to analyze $D_j$.

View $D_j$ as $D_j(n_j + S_j)$ where $n_j + S_j$ is a demand under equilibrium conditions. So we can also treat $n_j$ as an equilibrium demand. Now $n_j$ at $j$ corresponds to some demand $n_i + S_i$ at $i$. Thus, we may write

$$L_i(n_i) = D_i(n_i + S_i) + T_j(n_j),$$

regarding $n_i + S_i$ as an equilibrium demand at $i$. In sum, the experience of demands in equilibrium at $i$ can be used, through (6) and then (7), to describe the equilibrium lead time at $j$.

Specifically, we assume $L_i \sim \text{CPH}(\gamma_i, \Gamma_i)$, where $Z_i$ is the continuous-time Markov chain whose absorption determines $L_i$. Also, $(X_i, W_i)$ is defined as the joint process $(Z_i, A)$ embedded at demand epochs during the lead time, analogous to $(X, W)$ given previously. Then, by an analysis similar to that of Section 3, $(X_i, W_i)$ has transition matrix

$$\Omega_i = (I \otimes \Lambda_i - \Gamma_i \otimes I - I \otimes Q)^{-1}(I \otimes \Lambda_i).$$

Let $\sigma_i(0)$ be the initial density which reflects the equilibrium condition when $L_i$ and $\tau_S$ begin. If $i$ is the root node, we set $\sigma_i(0) = \gamma_i \otimes \alpha_i$, as in the single-location model, where $\gamma_i \sim u_i$. Otherwise, assume $\sigma_i(0)$ has been determined by analysis of Pre$(i)$.

Now

$$\sigma_i(S_i) = \sigma_i(0)\Omega_i^S$$

gives the joint density of $(X_i, W_i)$ when $D_i$ begins, and $D_i$ is the time to absorption of $Z_i$ with an initial density derived from $\sigma_i(S_i)$. Also, $L_j = D_i + T_j$ can be represented as the time until absorption of a chain $Z_j$ with generator

$$\Gamma_j = \begin{bmatrix} \Gamma_j & -\Gamma_j u_i w_j \\ 0 & U_j \end{bmatrix}.$$

Thus, we can define processes $W_j = \{W_j[n_j]\}$ and $X_j = \{X_j[n_j]\}$, where $W_j$ and $X_j$ are $A$ and $Z_j$ observed by demand $n_j$, respectively. As before, the joint chain $(X_j, W_j)$ has transition matrix

$$\Omega_j = (I \otimes \Lambda_j - \Gamma_j \otimes I - I \otimes Q)^{-1}(I \otimes \Lambda_j).$$

In order to analyze $D_j$, we must determine the initial density $\sigma_j(0)$ for $(X_j, W_j)$. Since $n_j$ corresponds to some $n_i + S_i$, $W_j$ begins where $W_i$ terminates at time $\tau_S$, so we identify $W_j[0]$ with $W_i(S_i)$. As for $X_j$, there are two groups of states, reflected in the partition of $\Gamma_j$, one group inherited from $X_i$, and a new group reflecting $T_j$. If $X_j[0]$ is in the first group, then $Z_i$ was not absorbed at $\tau_S$ and $D_i > 0$. The vector $\sigma_i(S_i)$ thus describes $(X_i[0], W_i[0])$ for this group of
states. If \(X_j[0]\) is in the second group, then \(D_i = 0\), and \(L_i\) starts with \(T_i\), according to the probability vector \(v_j\). For these states \((X_j[0], W_j[0])\) has density
\[
v_j \otimes (\alpha_j - \sigma_i(S_i)(e \otimes I)).
\]
Here, \(\alpha_j\) is the equilibrium density of \(W_i\), from which we deduce, state by state, the probability that \(D_i > 0\). Notice, \(\alpha_j = \alpha_0\) in this case, because of (3) and (5). In sum,
\[
\sigma(0) = \{\sigma_i(S_i), v_j \otimes (\alpha_j - \sigma_i(S_i)(e \otimes I))\}.
\]
Now we can compute
\[
\sigma_i(S_i) = \sigma_i(0)\Omega_i^S,
\]
and proceed to analyze Suc\((j)\). Applying this approach recursively, we can analyze the entire network from top to bottom.

At each level we have
\[
D_i \sim \text{CPH}[\sigma_i(S_i)(I \otimes e), \Gamma_i].
\]
and
\[
E[D_i] = -\sigma_i(S_i)(I \otimes e)\Gamma_i^{-1}e.
\]
Observe that the matrix \(\Gamma_i\) here is independent of the choice of \(S_i\). This is used only to determine the initial vectors.

Suppose we wish to evaluate a single policy, which means evaluating (14) for all \(i\). We can get a rough idea of the complexity of the entire computation. First, suppose we have completed the computation of \(E[D_j]\) from the top node 0 down to the node \(i\), and proceed to compute \(E[D_j]\) for \(j \in \text{Suc}(i)\). From (11)–(14), it is easy to see that the complexity of computing \(E[D_j]\) is dominated by that of \(\sigma_i(S_i)\), which in turn is dominated by the computation of \(\Omega_i^S\). By the analysis following Proposition 3.4, this is \(O(\log(S_j)m^3I_j^e)\), where \(I_j\) is the number of transient states of the process \(Z_j\). Suppose we have total of \(J + 1\) nodes in the entire network, with root \(j = 0\). Then the total number of operations needed in the whole procedure is bounded by
\[
\sum_{j=0}^{J} O(\log(S_j)m^3I_j^e).
\]

### 4.2. Nonproportional-Rate Case

Now suppose we do not have (5). Since \(D_{ij}\) depends on \(j\) in general, the analysis becomes more complex. To keep the notation simple, we start with the two-echelon model, whose root is \(i = 0\).
Suppose the set of leaves is $\text{Suc}(0) = \{1, 2, \ldots, J\}$. Note, $D_{0j} = [L_0 - \tau_{S_0}]^+$, where $\tau_{S_0}$ is the time between demand units $n_0$ and $n_0 + S_0$ at node 0, and $n_0$ is an equilibrium demand, conditional on demand $n_0 + S_0$ coming from node $j$. Let $C_0[n]$ be the index of the leaf from which demand $n$ comes. Then, we need to compute the density of $(X_0[S_0], W_0[S_0]|C_0[S_0] = j)$, which we can write as

$$\sigma_0[S_0] = \frac{1}{\Pr[C_0[S_0] = j]} (\gamma_0 \otimes \alpha_0) \Omega_0^{S_0-1} \Omega_{0j}. \tag{15}$$

Here

$$\Omega_0 = (I \otimes \Lambda_0 - \Gamma_0 \oplus Q)^{-1} (I \otimes \Lambda_0) \tag{16}$$

is the one-step transition matrix of the chain $(X_0, W_0)$, and $\Omega_{0j}$ is the one-step transition matrix of the same chain assuming the transition is triggered by a demand from node $j$. (Hence, $\Omega_0 = \sum_{j=1}^J \Omega_{0j}$.)

Now, we analyze the matrix $\Omega_{0j}$. For this purpose, define

$$N_j(t) = \text{number of demands in } [0, t] \text{ at node } j,$$

$$N_0(t) = \sum_{j=1}^J N_j(t),$$

$$t_1 = \text{first arrival epoch at node } 0.$$

For $t \geq 0$, let

$$R(u, a, v, b, t) = \Pr\{Z_0(t) = v, A(t) = b, N_0(t) = 0$$

$$|Z_0(0) = u, A(0) = a, N_0(0) = 0\},$$

$$V_{0j}(u, a, v, b, t) = \Pr\{t \leq t_1, Z_0(t_1) = v, A(t_1) = b, N_j(t_1) = 1$$

$$|Z_0(0) = u, A(0) = a, N_0(0) = 0\},$$

$$R(t) = [R(u, a, v, b, t)]_{u=v},$$

$$V_{0j}(t) = [V_{0j}(u, a, v, b, t)]_{u=v}.$$

Clearly, $\Omega_{0j} = \lim_{t \to \infty} V_{0j}(t)$. Observe,

$$\frac{d}{dt} R(t) = R(t)[\Gamma_0 \oplus Q - I \otimes \Lambda_0],$$

$$\frac{d}{dt} V_{0j}(t) = R(t)(I \otimes \Lambda_j),$$
and

\[ R(0) = I \otimes I \quad V_0(0) = 0 \otimes 0. \]

Applying Lemma 3.1 yields

\[ \Omega_{0j} = (I \otimes \Lambda_0 - \Gamma_0 \oplus Q)^{-1}(I \otimes \Lambda_i). \]  \hspace{1cm} (17)

Next, we compute \( Pr\{C_0[S_0] = j\} \). Recall, \( W_0[0] \sim \alpha_0 \), where \( \alpha_0 \) is the equilibrium distribution of \( W_0 \) (Lemma 3.2). So, \( W_0[S_0] \sim \alpha_0 \). Thus

\[
Pr\{C_0[S_0] = j\} = \sum_a Pr\{W_0[S_0] = a\}Pr\{C_0[S_0] = j|W_0[S_0] = a\}
\]

\[
= \sum_a \frac{\lambda_i(a)}{\lambda_o(a)} \alpha_o(a)
\]

\[
= \alpha_0 \Lambda_0^{-1} \Lambda_j e,
\]

which is independent of \( S_0 \).

Combining (15)–(18) yields

\[
\sigma_{0j}[S_0] = [\alpha_0 \Lambda_0^{-1} \Lambda_j e]^{-1}(\gamma_0 \otimes \alpha_0)\Omega_{0j}^{-1} \Omega_{0j}
\]

\[
= [\alpha_0 \Lambda_0^{-1} \Lambda_j e]^{-1}(\gamma_0 \otimes \alpha_0)[(I \otimes \Lambda_0 - \Gamma_0 \oplus Q)^{-1}(I \otimes \Lambda_i)]^{-1}
\]

\[
\times (I \otimes \Lambda_0 - \Gamma_0 \oplus Q)^{-1}(I \otimes \Lambda_i). \]  \hspace{1cm} (19)

The vector \( \sigma_{0j}[S_i] \) plays the same role as \( \sigma_j[S_i] \) does in the case of proportional demand rates. Specifically, we may use the former in place of the latter to define \( \sigma_j[0] \). Then we can analyze \( j \) as before.

This same approach can be extended to general multiechelon systems. At each node \( i \) the state space of \( C_i \) must include all the individual leaves supplied by node \( i \), not just Suc(\( i \)).

### 4.3. Reduction in the Proportional-Rate Case

Next we want to show that in the proportional-rate case, we do not have to distinguish which leaf a typical demand at node \( i \) comes from; i.e., \( D_{ij} \) is independent of \( j \). To keep the notation simple, we again focus on the two-echelon model. We need only show that \( \sigma_{0j}[S_i] \) is independent of \( j \).

Suppose now (5) holds. Let \( \hat{K} = \text{diag}(\kappa_a) \). Then

\[
\Lambda_j = \mu_j \hat{K}, \quad j = 0, 1, \ldots, J. \]  \hspace{1cm} (20)
From (19) it suffices to show that \([\alpha_0 \Lambda_0^{-1} \Lambda_j]^{-1} \Lambda_j\) is independent of \(j\). Notice, \(\Lambda_0^{-1} \Lambda_j = (\mu_j / \mu_0) I\), and \(\alpha_0 e = 1\). So

\[
[\alpha_0 \Lambda_0^{-1} \Lambda_j]^{-1} \Lambda_j = (\mu_0 / \mu_j)[\alpha_0 e]^{-1} \mu_j \bar{K} = \mu_0 \bar{K} = \Lambda_0,
\]

which is independent of \(j\). Note, (21) also verifies that \(\alpha_0 [S_0]\) is identical to \(\alpha_0 [S_0]\) derived in Section 4.1, provided (5) holds.

REFERENCES


Manuscript received December 26, 1990
Revised manuscript received December 5, 1991
Accepted February 24, 1992