Chapter 2

The Stochastic Knapsack

The classical deterministic knapsack problem involves a knapsack of capacity C resource units and K classes of objects, with class-k objects having size b_k. Objects may be placed into the knapsack as long as the sum of their sizes does not exceed the knapsack capacity. A reward r_k is accrued whenever a class-k object is placed into the knapsack. The problem is to place the objects into the knapsack so as to maximize the total reward.

We now begin our study of the stochastic knapsack: Its objects again have heterogeneous resource requirements, but they now arrive and depart at random times. This stochastic system is of fundamental importance in modeling multiservice telecommunication technology. In Chapters 2 and 3 we assume that an arriving object is always placed into the knapsack when there is sufficient room; otherwise, the arriving object is blocked and lost. In Chapter 4 we suppose that objects have class-dependent rewards and that arriving objects can be denied access — even when there is room in the knapsack — in order to maximize the long-run average reward.

2.1 The Model and Notation

The stochastic knapsack consists of C resource units to which objects from K classes arrive. Objects from class k are distinguished by their size, b_k, their arrival rate, λ_k, and their mean holding time, $1/\mu_k$.

Class-k objects arrive at the knapsack according to a Poisson process with rate λ_k, and the K arrival processes are independent. If an arriving class-k object is admitted into the knapsack, it holds b_k resource units for a holding time that is exponentially distributed with mean $1/\mu_k$; at the end of this holding time, the b_k resource units are simultaneously released. Holding times are independent of each other and of the arrival processes. Let n_k denote the number of class-k objects in the knapsack. Then the total amount of resource utilized by the objects in the knapsack is $\mathbf{b}\cdot\mathbf{n}$, where $\mathbf{b} := (b_1, \ldots, b_K)$, $\mathbf{n} := (n_1, \ldots, n_K)$, and

$$\mathbf{b}\cdot\mathbf{n} := \sum_{k=1}^{K} b_k n_k.$$

The knapsack always admits an arriving object when there is sufficient room (see Figure 2.1). More specifically, it admits an arriving class-k object if $b_k \leq C - \mathbf{b}\cdot\mathbf{n}$; otherwise, it blocks and loses the object. Without loss of generality we assume that the sizes b_k, $k = 1, \ldots, K$, and the capacity C are all positive integers.

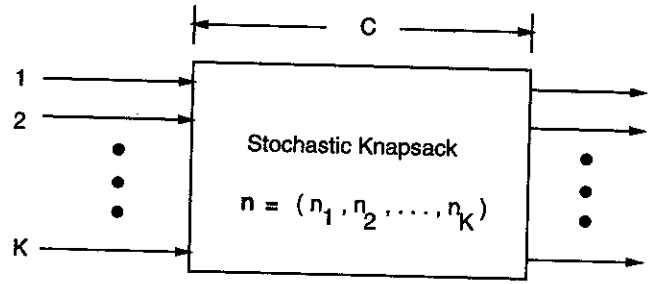

Figure 2.1: The stochastic knapsack. Objects from K different classes arrive at random and share C resource units. An object departs after its holding times.

The stochastic knapsack meets the definition of the pure loss system given in Chapter 1: An arriving object is either blocked or admitted into the knapsack and, if admitted, it remains in the knapsack for the duration of its holding time.

A Markov process captures the dynamics of the stochastic knapsack.

To define this process, let

$$\mathcal{S} := \{\mathbf{n} \in \mathcal{I}^K : \mathbf{b}\cdot\mathbf{n} \leq C\}$$

be the state space, where $\mathcal{I}$ is the set of non-negative integers. Let $X_k(t)$ be the random variable denoting the number of class-k objects in the knapsack at time t. Let $\mathbf{X}(t) := (X_1(t), \ldots, X_K(t))$ be the state of the knapsack at time t and $\{\mathbf{X}(t)\}$ be the associated stationary stochastic process. It is easily verified that this process is an aperiodic and irreducible Markov process over the finite state space $\mathcal{S}$.

We now address the equilibrium behavior of the stochastic knapsack. For each $\mathbf{n} \in \mathcal{S}$, denote $\pi(\mathbf{n})$ as the probability that the knapsack is in state $\mathbf{n}$ in equilibrium (equivalently, the long-run fraction of time that the knapsack is in state $\mathbf{n}$). Let $\rho_k := \lambda_k/\mu_k$ be the *offered load* for class-k objects. A fundamental result for the stochastic knapsack is given below.

Theorem 2.1 *The equilibrium distribution for the stochastic knapsack is*

$$\pi(\mathbf{n}) = \frac{1}{G}\prod_{k=1}^{K}\frac{\rho_k^{n_k}}{n_k!}, \qquad \mathbf{n} \in \mathcal{S}, \tag{2.1}$$

where

$$G := \sum_{\mathbf{n}\in\mathcal{S}}\prod_{k=1}^{K}\frac{\rho_k^{n_k}}{n_k!}. \tag{2.2}$$

Proof: First suppose that $C = \infty$. For each $k = 1, \ldots, K$, let $\{Y_k(t)\}$ be the stationary stochastic process denoting the number of class-k objects present in this uncapacitated system. Note that these processes form K independent birth–death processes, where the birth and death rates of the kth process are λ_k and $n_k\mu_k$, respectively. Hence the stationary distribution for this uncapacitated system is

$$\tilde{\pi}(\mathbf{n}) = \prod_{k=1}^{K}\frac{\rho_k^{n_k}}{n_k!}e^{-\rho_k}, \qquad \mathbf{n} \in \mathcal{I}^K. \tag{2.3}$$

A birth–death process is reversible and so is the (vector) joint process of independent reversible processes (see Kelly [89], Chapter 1). Hence the stationary Markov process $\{\mathbf{Y}(t)\}$ is reversible, where $\mathbf{Y}(t) :=$

$(Y_1(t),\ldots,Y_K(t))$. Now the state process of the original capacitated system, $\{\mathbf{X}(t)\}$, is a Markov process whose transition probabilities are the same as those for $\{\mathbf{Y}(t)\}$, except that they are truncated to $\mathcal{S}$ (that is, transitions from inside of $\mathcal{S}$ to outside of $\mathcal{S}$ are removed). Hence, by Corollary 1.10 of Kelly [89], the equilibrium distribution for the original capacitated system is given by (2.3) truncated to $\mathcal{S}$. But this distribution is $\pi(\mathbf{n})$, $\mathbf{n} \in \mathcal{S}$, given by (2.1). □

The expression for $\pi(\mathbf{n})$ given in Theorem 2.1 is called the product-form solution for the stochastic knapsack. We shall see in subsequent chapters that the product-form solution extends to much more general loss systems. The constant G defined in (2.2) is called the *normalization constant* for the stochastic knapsack.

Although we have assumed exponential holding time distributions, Theorem 2.1 actually holds for arbitrary distributions. This so-called insensitivity result will be formally stated in Chapter 5 where more general loss systems are considered.

Blocking Probability and Throughput

A critical performance measure for the stochastic knapsack is blocking probability. Let B_k be the probability that an arriving class-k object is blocked (equivalently, the long-run fraction of arriving class-k objects that are blocked). Since larger objects require more room than smaller objects, they have higher blocking probabilities; more precisely, $B_k > B_l$ if $b_k > b_l$. In Section 2.7 we shall see that B_k is roughly proportional to b_k for knapsacks with large C and typical traffic conditions.

Another important performance measure is throughput. Let TH_k denote the throughput of class-k objects — that is, the long-run rate at which class-k objects are admitted into the knapsack. Since class-k objects arrive at the knapsack according to a Poisson process with rate λ_k, we have $\mathrm{TH}_k = \lambda_k(1 - B_k)$. Thus, if we know B_k, we can easily determine TH_k.

Notation

For reference purposes, we now present some additional notation that will be repeatedly used throughout this book. Let X_k be the random variable denoting the number of class-k objects in the system in equilibrium. Let

$$\mathbf{X} := (X_1,\ldots,X_K)$$

be the (random) state vector, so that

$$\pi(\mathbf{n}) = P(\mathbf{X} = \mathbf{n}).$$

Define the *utilization* of the knapsack in equilibrium by

$$U := b_1X_1 + \cdots + b_KX_K.$$

Thus

$$\mathrm{UTIL} := E[U]$$

is the knapsack's average utilization. Finally, let

$$\mathcal{K} := \{1,\ldots,K\}$$

be the set of all classes.

The Erlang Loss System

The stochastic knapsack generalizes the celebrated Erlang loss system. Indeed if there is only one class and all objects have size of unity, then the stochastic knapsack reduces to the Erlang loss system. For this special case, let λ denote the arrival rate of objects, $1/\mu$ the mean holding time of an object, and $\rho := \lambda/\mu$ the offered load. Then the probability that there are c objects in the system in equilibrium, $\pi(c)$, is

$$\pi(c) = \frac{\rho^c/c!}{\sum_{c=0}^{C}\rho^c/c!}, \quad c = 0,\ldots,C. \tag{2.4}$$

It is important to note that Theorem 2.1 is a multidimensional generalization of (2.4).

The blocking probability for the Erlang loss system, denoted by $ER[\rho, C]$, is given by the Erlang loss formula (also called the Erlang-B formula):

$$ER[\rho, C] = \frac{\rho^C/C!}{\sum_{c=0}^{C}\rho^c/c!}.$$

Erlang published this result in 1917 [21].

2.2 Performance Evaluation

In order to derive an expression for the blocking probability in terms of the basic model parameters, let $\mathcal{S}_k$ be the subset of states in which the knapsack admits an arriving class-k object, that is,

$$\mathcal{S}_k := \{\mathbf{n} \in \mathcal{S} : \mathbf{b} \cdot \mathbf{n} \leq C - b_k\}.$$

As an example, Figure 2.2 illustrates the sets $\mathcal{S}$ and $\mathcal{S}_2$ for a system with capacity $C = 8$, two classes of objects, and object sizes of $b_1 = 1$ and $b_2 = 2$. The set $\mathcal{S}$ is the collection of all the black points, whereas $\mathcal{S}_2$ is the collection of black points below the broken line.

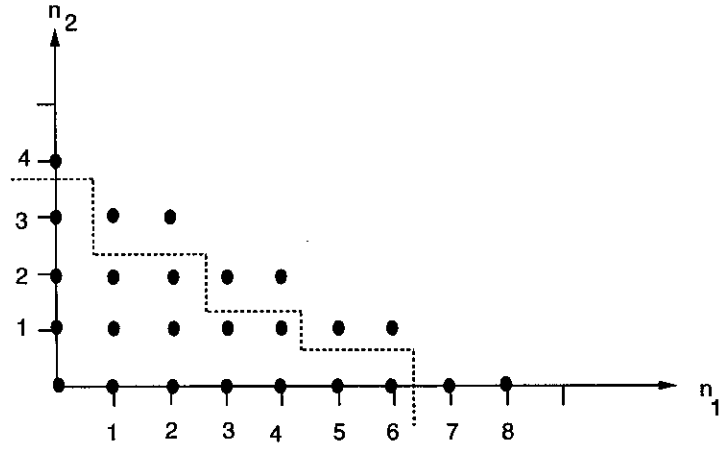

Figure 2.2: State diagram with $C = 8$, $b_1 = 1$, $b_2 = 2$. The knapsack admits arriving class-2 objects when its state is below the broken line.

Returning to the general model, because arrivals are Poisson the probability of blocking a class-k object is

$$B_k = 1 - \sum_{\mathbf{n} \in \mathcal{S}_k} \pi(\mathbf{n}).$$

This equality and Theorem 2.1 give an explicit expression for blocking

probability:

$$B_k = 1 - \frac{\sum_{\mathbf{n} \in \mathcal{S}_k} \prod_{j=1}^{K} \rho_j^{n_j}/n_j!}{\sum_{\mathbf{n} \in \mathcal{S}} \prod_{j=1}^{K} \rho_j^{n_j}/n_j!} \tag{2.5}$$

Although this result is important, it is typically impractical to brute-force sum the terms in the numerator and denominator, because the discrete state spaces $\mathcal{S}$ and $\mathcal{S}_k$ are prohibitively large for moderate values of C and K.

A Recursive Algorithm

We now present an efficient scheme for calculating the blocking probabilities which does not involve brute-force summation. Let

$$\begin{aligned} \mathcal{S}(c) &:= \{\mathbf{n} \in \mathcal{S} : \mathbf{b} \cdot \mathbf{n} = c\} \\ q(c) &:= \sum_{\mathbf{n} \in \mathcal{S}(c)} \pi(\mathbf{n}) \\ R_k(c) &:= \sum_{\mathbf{n} \in \mathcal{S}(c)} n_k \pi(\mathbf{n}). \end{aligned}$$

Note that $\mathcal{S}(c)$ is the set states for which exactly c resource units are occupied and that $q(c)$ is the probability of this event occurring in equilibrium. Let $q(c) := 0$ and $R_k(c) := 0$ for $c < 0$.

Corollary 2.1 *The occupancy probabilities $q(c)$, $c = 1, \ldots, C$, satisfy the following recursive equations:*

$$cq(c) = \sum_{k=1}^{K} b_k \rho_k q(c - b_k), \quad c = 0, \ldots, C.$$

Proof: We first observe that

$$cq(c) = \sum_{k=1}^{K} b_k R_k(c). \tag{2.6}$$

The derivation of the (2.6) does not rely on the product-form result of Theorem 2.1:

$$cq(c) = \sum_{\mathbf{n} \in \mathcal{S}(c)} c\pi(\mathbf{n})$$

$$= \sum_{\mathbf{n}\in\mathcal{S}(c)} \left(\sum_{k=1}^{K} b_k n_k\right) \pi(\mathbf{n})$$
$$= \sum_{k=1}^{K} b_k \sum_{\mathbf{n}\in\mathcal{S}(c)} n_k \pi(\mathbf{n}) = \sum_{k=1}^{K} b_k R_k(c).$$

It remains to derive an expression for $R_k(c)$ in terms of $q(c-b_k)$. From the product-form solution given in Theorem 2.1 we have

$$\begin{aligned} n_k\pi(\mathbf{n}) &= \frac{n_k}{G}\prod_{j=1}^{K}\frac{\rho_j^{n_j}}{n_j!} \\ &= \frac{n_k}{G}\frac{\rho_k^{n_k}}{n_k!}\prod_{j\neq k}\frac{\rho_j^{n_j}}{n_j!} \\ &= \frac{\rho_k}{G}\frac{\rho_k^{n_k-1}}{(n_k-1)!}\prod_{j\neq k}\frac{\rho_j^{n_j}}{n_j!} \\ &= \rho_k\pi(\mathbf{n}-\mathbf{e}_k), \end{aligned}$$

where $\mathbf{e}_k$ is the K-dimensional vector consisting of only zeros except for a one in the kth component. Thus

$$\begin{aligned} R_k(c) &= \sum_{\mathbf{n}\in\mathcal{S}(c)} n_k\pi(\mathbf{n}) \\ &= \rho_k \sum_{\mathbf{n}\in\mathcal{S}(c)} \pi(\mathbf{n}-\mathbf{e}_k) \\ &= \rho_k \sum_{\mathbf{n}\in\mathcal{S}(c-b_k)} \pi(\mathbf{n}) \\ &= \rho_k q(c-b_k), \end{aligned}$$

where the third equality follows from the change of variable $\mathbf{n}-\mathbf{e}_k \leftarrow \mathbf{n}$. Combining this with (2.6) gives the desired result. □

The following recursive algorithm determines the normalization constant, the occupancy probabilities, and the blocking probabilities. Its correctness is guaranteed by the above corollary.

Algorithm 2.1 *Recursive algorithm to calculate occupancy distribution and blocking probabilities for the stochastic knapsack*

1. Set $g(0) \leftarrow 1$ and $g(c) \leftarrow 0$ for $c < 0$.
2. For $c = 1, \ldots, C$, set
$$g(c) \leftarrow \tfrac{1}{c}\textstyle\sum_{k=1}^{K} b_k\rho_k g(c-b_k)\ .$$
3. Set
$$G = \textstyle\sum_{c=0}^{C} g(c).$$
4. For $c = 0, \ldots, C$, set
$$q(c) \leftarrow g(c)/G.$$
5. For $k = 1, \ldots, K$, set
$$B_k \leftarrow \textstyle\sum_{c=C-b_k+1}^{C} q(c)\ .$$

What is the computational complexity of the algorithm? Note that the bottleneck occurs in Step 2, where the unnormalized occupancy probability $g(c)$ is calculated. In order to calculate $g(c)$ for a fixed c, $O(K)$ arithmetic operations must be performed. Since $g(c)$ must be obtained for C values of c, the overall effort of Step 2, and of the algorithm as a whole, is $O(CK)$. The memory required by the algorithm is easily seen to be $O(K+C)$. Thus the computational and memory requirements are linear. Moreover, the algorithm rarely (if ever) encounters numerical problems such as imprecision or overflow. Hence we can use the recursive algorithm to determine the performance of the stochastic knapsack even when C and K are huge.

We can also use the recursive algorithm to calculate the average knapsack utilization, denoted by UTIL. Indeed, UTIL is a simple expression of the link occupancy probabilities:

$$\text{UTIL} = \sum_{c=0}^{C} cq(c).$$

Problem 2.1 For the stochastic knapsack, derive an expression for $\partial B_l/\partial\rho_k$, the derivative of the blocking probability for class-l objects with respect to ρ_k. Develop an efficient recursive algorithm to calculate $\partial B_l/\partial\rho_k$, $1 \le k, l \le K$.

2.3 Virtual Channel Establishment for ATM Multiplexers

The stochastic knapsack can accurately model the virtual channel dynamics of an asynchronous transfer mode (ATM) multiplexer (see Figure 2.3). In this section we show how it models ATM multiplexers with peak-rate admission, statistical multiplexing, and burst multiplexing.

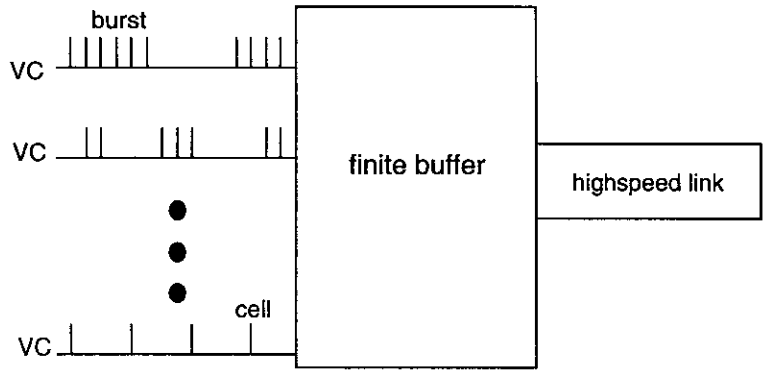

Figure 2.3: An ATM multiplexer.

We first recall and elaborate on the ATM terminology introduced in Chapter 1. A *source* is a terminal such as a telephone handset, a video player, or a multimedia computer. When a source wants to transmit information, it requests to establish a *virtual channel* (VC). Once a source has established a VC, it generates a stream of *cells*, each cell consisting of 53 bytes: 5 bytes for overhead (used for routing, error correction, priorities, etc.) and 48 bytes for the payload.[1] With a slight abuse of language, we shall write "an established VC" or sometimes more simply "a VC" for "a source with an established VC". A cell stream generated by an established VC consists of silent periods, during which no cells are generated, and activity periods, during which cells are generated either at constant or variable rate. The group of cells generated during

[1]Typically not all of the 48 bytes in the payload field carry source information; some bytes are taken by the ATM adaptation layer.

an activity period is called a *burst*. An *ATM multiplexer* is a buffer and a high-speed link; the buffer receives the cells generated by established VCs and transmits these cells, one after another, onto the high-speed link. Assume that VCs belong to a finite set of *services*. (In Section 2.8 we shall investigate a model with an infinite number — in fact, a continuum — of services.) Examples of services include voice, low- and high-quality facsimile, video conference, video on demand, file transfer, image retrieval, and LAN interconnect. The service type specifies the VC's cell generation properties and quality of service requirements, as discussed below.

Perhaps the most basic service is voice. The human mouth generates an analog voice signal which the source terminal samples, quantizes and digitizes, producing a binary stream at a rate of (say) 8,000 bytes per second. The ATM telephone then encapsulates the binary stream into ATM cells. Suppose that 40 bytes of the ATM cell carry bytes from the binary voice stream and the remaining 13 bytes carry overhead. Then a voice VC generates a cell every 5 msecs (=40/8,000 seconds) during an activity period and generates no cells during a silent period; hence its *peak rate* is 200 cells per second. For a video VC using variable bit rate (VBR) coding, the rate at which cells are generated may continuously vary but never exceeds the source's peak rate, which is again known at the outset.

We formally define the peak rate of a VC as follows. Suppose the VC generates cells with a minimum spacing of T seconds between the beginnings of successive cells. Then the VC has peak rate $1/T$ cells per second, or $1/T \times 53 \times 8$ bits per second. The units for C, the capacity of the high-speed link, can be in bits per second, cells per second, or some other block of data per second. Any convention is fine, as long as the same units are used for peak rates.

A buffer is required at the interface between the incoming cell streams and the high-speed link in order to limit the effect of *cell scale congestion* and *burst scale congestion* [126]. Cell scale congestion, due to the simultaneous arrival of cells from multiple VCs, requires a small buffer to prevent loss. Burst scale congestion occurs when the aggregate cell arrival rate, across all VCs in progress, is momentarily greater than the link capacity — as long as an arrival rate excess exists, buffer content will grow until saturation. Burst scale congestion therefore

requires a larger buffer to limit cell loss.[2]

We shall discuss three modes of operation for an ATM multiplexer: peak-rate admission, statistical multiplexing, and burst multiplexing. Peak-rate admission constrains the aggregate peak rate to be less than the transmission capacity of the high-speed link. It enforces this constraint by rejecting a VC establishment request when the VC's peak rate added to the sum of the peak rates for the established VCs exceeds the transmission capacity. Because the aggregate peak rate never exceeds the transmission capacity, peak-rate admission precludes burst scale congestion and, consequently, requires only a small buffer.

ATM with peak-rate admission resembles multirate circuit switching. Both schemes dedicate bandwidth to each connection in progress, under utilizing the transmission capacity when one or more established connections is in a silent period. ATM with peak-rate admission is, however, more flexible and easier to implement than multirate circuit switching. Indeed, ATM accommodates arbitrary and changing combinations of peak rates, whereas multirate circuit switching, once in place, supports a fixed set of peak rates. Moreover, as discussed in greater detail in the next section, multirate circuit-switching engenders difficult synchronization, signaling, and network management problems because the signaling network must track the positions of the slots in a frame for each multirate connection.

The statistical multiplexing mode permits the aggregate peak rate to exceed the transmission capacity. It can utilize the link more efficiently, allowing the link to transmit at its maximum rate, even when some of the established VCs are silent. It may, however, cause unacceptable cell loss or delay for one or more of the established VCs. A VC's allowable cell loss and delay are specified by its quality of service (QoS) requirements; for example, the QoS requirement for a service might be that the fraction of cells lost be less than 10^{-6}. To guarantee that all the VCs in progress meet their QoS requirements it is necessary, as with peak-rate admission, to reject certain VC establishment requests.

To implement statistical multiplexing, we must determine whether a

[2]Although a very large buffer may render cell loss negligible, it does not prevent unacceptable cell delays.

given collection of established VCs meet the QoS requirements. Peak-rate admission obviates this problem at the expense of reduced efficiency. In Section 4.7 we shall discuss a specific statistical multiplexing scheme, *service separation*, which greatly alleviates this problem while preserving statistical multiplexing for VCs belonging to the same service.

The burst multiplexing mode also permits the sum of the peak rates for the established VCs to exceed the transmission capacity; but it does not permit the established VCs to transmit bursts at will. Specifically, an established VC can only transmit a burst if the peak rate of the burst plus the sum of the peak rates of the bursts in progress is less than the link capacity. If this condition is violated, then the burst is either lost or stored in the terminal buffer for transmission at a later time. A service's allowable burst loss and burst delay are specified by its QoS requirements; for example, the QoS requirement for a service might be that the fraction of bursts blocked be less than 10^{-4}. To guarantee that all the VCs in progress meet their QoS requirements it is necessary, as with peak-rate admission and statistical multiplexing, to reject certain VC establishment requests. As does peak-rate admission, burst multiplexing precludes burst scale congestion and, consequently, requires only a small buffer. The burst multiplexing mode can be implemented with Boyer's fast reservation protocol [19].

In order to model the three modes of operation in the context of the stochastic knapsack, we introduce some notation and terminology. Let C denote the transmission capacity of the high-speed link, K denote the number of services, and $b_1, \ldots, b_K$ denote the peak rates for the K services. The *VC profile* is $(n_1, \ldots, n_K)$, where n_k is the current number of established service-k VCs. Since VCs arrive and depart, the VC profile changes with time. We assume that service-k VCs make establishment requests according to a Poisson process with rate λ_k. But we permit the holding time of a service-k VC to have an arbitrary distribution with mean $1/\mu_k$.

A Knapsack Model for Peak-Rate Admission

Peak-rate admission admits a new service-k VC if and only if

$$b_k + \sum_{l=1}^{K} b_l n_l \leq C,$$

where $(n_1, \ldots, n_K)$ is the current VC profile. Consequently, at all times the VC profile $(n_1, \ldots, n_K)$ satisfies

$$\sum_{k=1}^{K} b_k n_k \leq C.$$

The ATM multiplexer with peak-rate admission perfectly matches the stochastic knapsack. The Recursive Algorithm 2.1 efficiently calculates the blocking probability, B_k, for service-k VC requests. It also can determine throughputs, average utilization, and derivatives of these performance measures.

A Knapsack Model for Statistical Multiplexing

Now suppose that the ATM system operates in the statistical multiplexing mode. This mode permits VC profiles $(n_1, \ldots, n_K)$ satisfying

$$\sum_{k=1}^{K} b_k n_k > C,$$

where the b_k's are again the peak rates and C is the capacity of the high-speed link. The performance of ATM with statistical multiplexing is now characterized by two types of measures. The first, referred to as *cell performance*, is the cell loss and delay due to cell accumulation and overflow in the buffer. The second, referred to as *connection performance*, is the rejection probability of VC establishment requests. There is a clear tradeoff between cell and connection performance: If we admit (respectively reject) more VC connection requests, the buffer will become more (respectively less) congested.

The performance of an ATM multiplexer operating under the statistical multiplexing mode also depends on the scheduling policy that is

employed to transmit cells. In order to fix ideas, and to focus on admission and not on scheduling, throughout this discussion we assume that cells are transmitted in order of arrival, independently of their service type. We consider more elaborate scheduling policies in Section 4.7.

Cells and VC establishment requests arrive at two entirely different time scales; the former occurs on the order of milliseconds whereas the latter occurs on the order of seconds or even minutes. Therefore the VC profile is quasi-static with respect to the cell arrival processes. This observation motivates the following definition: A VC profile $\mathbf{n} = (n_1, \ldots, n_K)$ is said to be *allowable* if the QoS requirements are met for all K services when $\mathbf{n}$ is permanent (that is, no new VC establishments or VC departures). Let Λ denote the set of allowable VC profiles.

The *admission policy* determines whether or not an arriving VC is accepted. We require the admission policy to be a function only of the current VC profile and the service type of the arriving VC; in particular, we do not permit the policy to take into account the buffer content at the VC arrival instant. The restriction to this class of policies is again motivated by the great difference in time scales between cell arrivals and VC request arrivals. Under this restriction, an admission policy can be defined by a mapping $\mathbf{f} = (f_1, \ldots, f_K)$, where $f_k : \mathcal{I}^K \to \{0, 1\}$ and $f_k(\mathbf{n})$ takes the value 0 (respectively 1) if an arriving service-k VC is rejected (respectively admitted) when the current profile is $\mathbf{n}$. For policy $\mathbf{f}$, let $\mathcal{S}(\mathbf{f})$ denote the set of profiles in $\mathcal{I}^K$ that are visited infinitely often. The set $\mathcal{S}(\mathbf{f})$ is called the *admission region* of the policy $\mathbf{f}$. We say that $\mathbf{f}$ is an *allowable policy* if $\mathcal{S}(\mathbf{f}) \subseteq \Lambda$. Thus, when the system operates under an allowable policy, every possible VC profile meets the QoS requirements.

In Chapter 4 we shall consider optimizing, over the class of all admission policies, the performance of the stochastic knapsack. For this discussion, however, we limit our attention to a subclass of policies, namely, linear policies. A policy $\mathbf{f}$ is said to be a *linear policy* if there exists positive numbers $b_1^e, \ldots, b_K^e$ such that

$$f_k(\mathbf{n}) = 1 \quad \Leftrightarrow \quad \mathbf{b}^e \cdot \mathbf{n} + b_k^e \leq C,$$

where $\mathbf{b}^e = (b_1^e, \ldots, b_K^e)$. Under a linear policy, the admission region is

$$\mathcal{S}(\mathbf{f}) = \{\mathbf{n} : \mathbf{b}^e \cdot \mathbf{n} \leq C\}.$$

Thus, if the ATM system is operated in the statistical multiplexing mode under a linear policy, the stochastic knapsack can again be applied to determine the VC blocking probabilities and other performance measures of interest. Observe that peak-rate admission is the linear policy with $\mathbf{b}^e = \mathbf{b}$.

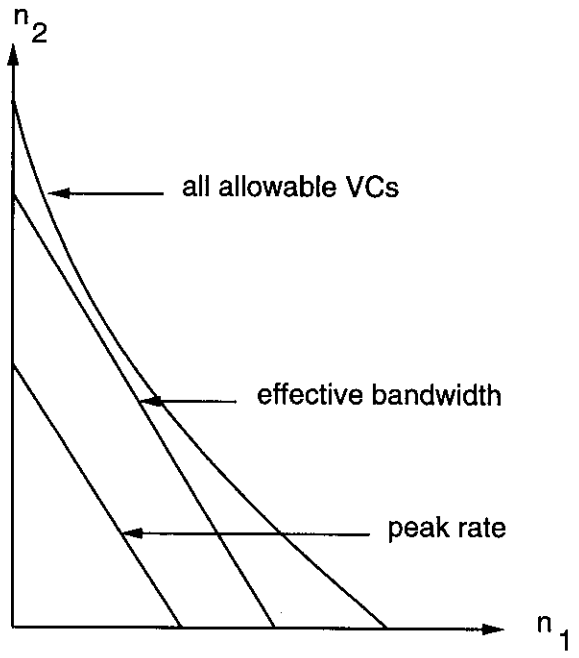

Figure 2.4: The boundaries of the admission regions for three allowable policies.

Suppose that $\mathbf{f}$ is an allowable linear policy with associated vector $\mathbf{b}^e = (b_1^e, \ldots, b_K^e)$. We say that b_k^e is the *effective bandwidth* of service k. Figure 2.4 shows the shapes and relationships of the boundaries of the admission regions for three admission policies: the policy based on peak rates; the policy based on effective bandwidths; and the policy based on admitting all VCs as long as the profile remains allowable. The boundaries for the peak rate admission and effective bandwidth admission are

linear, whereas the boundary for all allowable VCs is typically nonlinear and convex. The peak rate boundary is below the effective bandwidth boundary, which is below the boundary for all allowable VCs.

A Knapsack Model for Burst Multiplexing

Again let b_k denote the peak rate of the service-k VCs and C the capacity of the link. The burst multiplexing mode permits VC profiles $\mathbf{n} = (n_1, \ldots, n_K)$ satisfying

$$\sum_{k=1}^{K} b_k n_k > C,$$

but an established VC can no longer transmit a burst at will. To be more specific, let m_k denote the number of class-k VCs that are currently transmitting bursts; clearly, $0 \leq m_k \leq n_k$. If an established service-k VC wants to transmit a new burst, it may do so if and only if

$$b_k + \sum_{l=1}^{K} b_l m_l \leq C.$$

If the above condition is violated, then the burst is blocked — that is, it is either lost or stored in the terminal buffer for transmission at a later time.

As with the statistical multiplexing mode, if the VC admission policy is linear, we can define effective bandwidth and use the stochastic knapsack theory to determine VC blocking probability. In Chapter 3 we shall see how burst blocking probabilities can be assessed with a generalized version of the stochastic knapsack.

2.4 Contiguous Slot Assignment

Multirate circuit switching is another transport scheme that enables services with different bandwidth requirements to be integrated over the same transmission links. Since the late 1980s multirate circuit switching has been offered by many telephone companies, being a popular technology for providing video conference services on a dial-up basis.

We discuss multirate circuit switching in the context of time division multiplexing (TDM). A TDM frame consists of C slots, all of which contain the same number of bits. A call in progress occupies one or more slots in a TDM frame, and as the frame cycles around, the call stays in the same slot positions for its entire duration.[3] If an arriving call requires b slots per frame and less than b slots are available, then the arriving call is rejected. The stochastic knapsack is an appropriate model for this system if *flexible slot assignment* is used — that is, if the slots of a wideband call ($b \geq 2$) can be arbitrarily scattered across the TDM frame.

But most multirate cirucuit switching systems in operation or planned for deployment require *contiguous slot assignment*: each wideband call is required to occupy contiguous slots in the TDM frame. This restriction greatly simplifies the slot assignment and tracking that are needed at both ends of the transmission system.

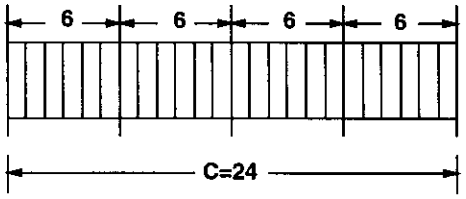

Figure 2.5: A contiguous allocation scheme. The TDM frame consists of four groups of six slots. A wideband call occupies six contiguous slots in a group.

As an example, consider a TDM frame consisting of $C = 24$ slots. Suppose the TDM system is to integrate two classes of calls — the narrowband calls, requiring $b_n = 1$ slot per frame, and the wideband calls, requiring $b_w = 6$ slots per frame. One possible contiguous allocation scheme is illustrated in Figure 2.5. In this scheme, when a wideband call arrives, it must be inserted into one of the four contiguous groups of slots shown in Figure 2.5; if all four of the groups have at least one occupied slot, then an arriving wideband call is blocked. In order to

[3]For the stochastic knapsack we use the natural terminology object; for ATM we use VC; for multirate circuit switching and telephone networks we use call.

reduce the blocking of wideband calls, it is desirable to *pack* the narrowband calls — that is, an arriving narrowband call is placed into the unfilled group with the largest number of occupied slots.

From the perspective of connection blocking, multirate circuit switching with flexible slot assignment closely resembles ATM with peak-rate admission. But the two schemes have important differences. Multirate circuit switching has potentially higher information transport because it lacks ATM's overhead in the cell header. On the other hand, ATM efficiently adapts to new services and bandwidth requirements and, because it does not track slots, simplifies the network signaling.

Performance Analysis of Contiguous Slot Assignment

With flexible slot assignment, we can easily assess performance with the Recursive Algorithm 2.1. However, contiguous slot assignment, which cannot be modeled as a stochastic knapsack, makes the analysis much more difficult, although not intractable. Focusing on contiguous slot assignment, we now discuss a Markov process analysis which was originally developed by Ramaswami and Rao [117]. We present this analysis in the context of Figure 2.5, that is, we assume $C = 24$, two classes with $b_n = 1$ and $b_w = 6$, and each wideband call occupies one of four contiguous groups. We also assume that calls arrive according to Poisson processes and have exponential holding times. Finally we assume that packing is employed for the narrowband calls.

The state of the system is described by the vector

$$\mathbf{n} = (n_w, n_0, \ldots, n_6),$$

where n_w is the number of wideband calls in the frame, n_0 is the number of groups with all slots idle, and for $i = 1, \ldots, 6$, n_i is the number of groups with exactly i slots occupied by narrowband calls. The stochastic process corresponding to this state vector is clearly a finite-state Markov process; its state space is

$$\mathcal{S} := \{\mathbf{n} \in \mathcal{I}^7 : n_w + n_0 + n_1 + \cdots + n_6 = 4\}.$$

The number of states in $\mathcal{S}$ is the same as the number of ways of throwing four indistinguishable balls into eight urns, which is equal to 330. Denote $\mathcal{S}_l$ for the set of states that have l wideband calls.

The infinitesimal generator for this Markov process takes the form

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{00} & \mathbf{Q}_{01} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{Q}_{10} & \mathbf{Q}_{11} & \mathbf{Q}_{12} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{21} & \mathbf{Q}_{22} & \mathbf{Q}_{23} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Q}_{32} & \mathbf{Q}_{33} & \mathbf{Q}_{34} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Q}_{43} & \mathbf{Q}_{44} \end{bmatrix},$$

where $\mathbf{Q}_{lm}$ is a matrix containing the rates from states in $\mathcal{S}_l$ to the states in $\mathcal{S}_m$.[4] It is a straightforward exercise to specify the terms in the $\mathbf{Q}_{lm}$ matrix (see [117] and Section 4.6). The equilibrium probabilities $\boldsymbol{\pi} = (\pi(\mathbf{n}), \mathbf{n} \in \mathcal{S})$ are the solutions to

$$\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}$$

along with

$$\sum_{\mathbf{n}\in\mathcal{S}} \pi(\mathbf{n}) = 1.$$

Computing $\boldsymbol{\pi}$ is fairly efficient because the matrix $\mathbf{Q}$ is block diagonal and the $\mathbf{Q}_{lm}$'s are sparse; efficient computational procedures that exploit this special structure are discussed in [117]. The equilbrium probabilities determine the equilibrium performance measures of greatest interest [117].

Ramaswami and Rao observed that the blocking probability for wideband calls is significantly greater with contiguous slot assignment than with flexible slot assignment. This increase in blocking results, of course, from the need for a contiguous group of six free slots, rather than six free slots anywhere in the frame, upon arrival of a wideband call. Furthermore, because contiguous slot assignment blocks a greater fraction of wideband calls, arriving narrowband calls see more free slots and, hence, have less blocking. This reduction in blocking, however, is usually insignificant because the probability of blocking for narrowband calls is minute even for flexible slot assignment in most practical circumstances.

[4] Of course the diagonal terms in $\mathbf{Q}_{lm}$ are defined so that the sum across each row is zero.

Problem 2.2 Determine the cardinalities of $\mathcal{S}_l$, $l = 0, \ldots, 4$. Write out explicitly the matrices $\mathbf{Q}_{lm}$, $0 \le l, m \le 4$.

Approximate Performance Analysis of Contiguous Slot Assignment

In many applications wideband calls have both significantly longer holding times and significantly smaller arrival rates than narrowband calls. For example, in a system integrating voice and video conference calls, we expect the video calls to arrive relatively infrequently but to have relatively long durations. In these applications, there is typically many narrowband arrivals and departures before the next wideband arrival or departure.

Observing this disparity in time scales, Reiman and Schmitt [121] propose a natural approximation in which the narrowband occupancy process is assumed to reach equilibrium between wideband events (arrivals and departures). The approximation is an application of the more general theory of nearly completely decomposable (NCD) Markov chains [38]. We now summarize the technique as applied to the multiplexing system of Figure 2.5.

Let λ_n and μ_n denote the arrival and departure rates for the narrowband calls. Fix $\bar{\lambda}_w$, $\bar{\mu}_w$, and $\epsilon > 0$. Let $\lambda_w := \epsilon\bar{\lambda}_w$ and $\mu_w := \epsilon\bar{\mu}_w$ be the arrival and departure rates for the wideband calls. Let $B_n(\epsilon)$ and $B_w(\epsilon)$ denote the blocking probability of the narrowband and wideband calls. The NCD regime corresponds to $\epsilon \to 0$. The approximation procedure is to approximate $B_n(\epsilon)$ and $B_w(\epsilon)$ by $\tilde{B}_n$ and $\tilde{B}_w$, where

$$\tilde{B}_n := \lim_{\epsilon\to 0} B_n(\epsilon) \qquad \tilde{B}_w := \lim_{\epsilon\to 0} B_w(\epsilon).$$

The appealing feature of this approximation is that $\tilde{B}_n$ and $\tilde{B}_w$ are substantially easier to calculate than $B_n(\epsilon)$ and $B_w(\epsilon)$.

Following Reiman and Schmitt, we now show how to calculate $\tilde{B}_n$ and $\tilde{B}_w$. First consider the case when $\epsilon = 0$. Then there are no wideband arrivals or departures; thus, the off-diagonal matrices, $\mathbf{Q}_{lm}$, $l \neq m$, are now all zero matrices. Let $\boldsymbol{\pi}_l = (\pi_l(\mathbf{n}),\ \mathbf{n} \in \mathcal{S}_l)$ solve

$$\boldsymbol{\pi}_l\mathbf{Q}_{ll} = \mathbf{0}$$

along with

$$\sum_{\mathbf{n}\in\mathcal{S}_l} \pi_l(\mathbf{n}) = 1.$$

Then $\boldsymbol{\pi}_l$ is the equilibrium distribution of the Markov process that starts with l wideband calls, but has no wideband arrivals or departures. If a wideband call were permitted to arrive to this system (with $\epsilon = 0$), it would be blocked with probability

$$\tilde{B}_l^w = \sum_{\mathbf{n}\in\mathcal{S}_l:n_0=0} \pi_l(\mathbf{n}).$$

Now that we have the blocking probability for wideband calls conditioned on the number of wideband calls in progress, we determine the same quantity for the narrowband calls. Given l wideband calls in progress, the narrowband calls see an Erlang loss system with capacity $C = 24 - 6l$ and offered load $\rho_n = \lambda_n/\mu_n$; thus, the probability that a narrowband call is blocked is $ER[\rho_n, 24 - 6l]$.

Next, for $\epsilon > 0$ but small, the number of wideband calls in the system is approximately a birth–death process with birth rates $\bar{\lambda}_w\epsilon(1 - B_w^l)$, $l = 0,\ldots,3$, and death rates $\bar{\mu}_w\epsilon l$, $l = 1,\ldots,4$. Let $(\eta_0,\ldots,\eta_4)$ denote the stationary distribution of this birth–death process. (It does not depend on ϵ.) Then it can be shown from the NCD theory (see Simon and Ando [147]) that the limiting probabilities are computed by unconditioning the conditional probabilities, that is,

$$\hat{B}_n = \sum_{l=0}^{4} \eta_l ER[\rho_n, 24 - 6l]$$

and

$$\hat{B}_w = \sum_{l=0}^{4} \eta_l \tilde{B}_w^l.$$

The most difficult part of this approximation procedure is computing the equilibrium probabilities $(\boldsymbol{\pi}_l(\mathbf{n}),\ \mathbf{n} \in \mathcal{S}_l)$ for each $l = 0, 1,\ldots,4$. This calls for solving a linear system with $|\mathcal{S}_l|$ unknowns for $l = 0,\ldots,4$. Nevertheless, solving these five systems of linear equations is substantially easier than solving the original system of linear equations with $|\mathcal{S}|$ unknowns.

For given λ_n, μ_n, λ_w, μ_w with $\lambda_n >> \lambda_w$ and $\mu_n >> \mu_w$, how does the engineer employ this procedure to approximate blocking probabilities? First the engineer determines the blocking probabilities for $\epsilon > 0$ conditioned on the presence of l wideband calls, as just described. These conditional probabilities depend on λ_n and μ_n but not on λ_w and μ_w. The engineer then solves the birth–death equations with birth rate $\lambda_w(1 - B_w^l)$ and death rate $\mu_w l$. The engineer then unconditions the conditional probabilities, as just described. Reiman and Schmitt give several numerical examples which illustrate the accuracy of the approximation.

2.5 Stochastic Comparisons

Up to this point our focus has been on computing performance measures of the stochastic knapsack. We have seen that a simple recursive algorithm does this quite efficiently.

The qualitative behavior of the stochastic knapsack is also of great interest. In particular, we would like to understand the behavior of the various performance measures as the arrival rate, service rate, and knapsack capacity increase. In order to facilitate an in-depth study of the knapsack's qualitative behavior, in this section we collect several important results from stochastic comparisons. We shall see that a particular form of stochastic comparison — namely, the likelihood ratio ordering — gives valuable insight into the qualitative behavior.

Throughout this section we shall write "increasing" for "non-decreasing"; similarly we shall write "decreasing" for "non-increasing". Let X and Y be two discrete random variables with support $\mathcal{I}$ (the non-negative integers).[5] The random variable X is said to be *stochastically larger* than the random variable Y, written $X \geq_{st} Y$, if

$$P(X \geq n) \geq P(Y \geq n) \text{ for all } n \in \mathcal{I}.$$

We will repeatedly make use of the fact that $X \geq_{st} Y$ if and only if $E[f(X)] \geq E[f(Y)]$ for all increasing functions $f(\cdot)$ defined over $\mathcal{I}$ (for

[5] A discrete random variable X is said to have support $\mathcal{T}$ if $P(X = n) > 0$ for all $n \in \mathcal{T}$ and $\sum_{n\in\mathcal{T}} P(X = n) = 1$.

example, see [143]). In particular, if $X \geq_{st} Y$ then $E[X^t] \geq E[Y^t]$ for $t \geq 0$.

Although stochastically larger is an important concept, it is often cumbersome to work with when dealing with product-form stochastic networks whose equilibrium probabilities have normalization constants. This has motivated researchers to consider another form of stochastic comparison, the likelihood ratio ordering. The random variable X is said to be larger than the random variable Y in the sense of the *likelihood ratio ordering*, written $X \geq_{lr} Y$, if

$$\frac{P(X=n+1)}{P(X=n)} \geq \frac{P(Y=n+1)}{P(Y=n)} \quad \text{for all} \quad n \in \mathcal{I}.$$

One of the appealing features of the likelihood ratio ordering is that if there are complex normalization constants in the distributions of X and Y, they are canceled out when taking the above ratio, which often leads to a trivial verification of $X \geq_{lr} Y$. Another feature is that larger in the likelihood ratio sense implies larger in the stochastic sense:

Lemma 2.1 *Suppose $X \geq_{lr} Y$. Then $X \geq_{st} Y$.*

Proof: Let $m := \min\{n : P(X=n) \geq P(Y=n)\}$. If $m = 0$, then X and Y have the same distribution and the result holds trivially. Henceforth suppose that $m \geq 1$. It follows from the definition of m that $P(X=n) < P(Y=n)$ for all $n = 0, 1, \ldots, m-1$. Thus

$$P(X \leq n) < P(Y \leq n) \quad \text{for all} \quad n = 0, 1, \ldots, m-1$$

and hence

$$P(X > n) \geq P(Y > n) \quad \text{for all} \quad n = 0, 1, \ldots, m-1. \tag{2.7}$$

From the definition of m we have

$$P(X=m) \geq P(Y=m). \tag{2.8}$$

Because $X \geq_{lr} Y$ we also have $P(X=n+1) \geq P(Y=n+1) \cdot P(X=n)/P(Y=n)$ for all $n \geq m$, which, when combined with (2.8) and a simple inductive argument, gives

$$P(X \geq n) \geq P(Y \geq n) \quad \text{for all} \quad n \geq m. \tag{2.9}$$

Combining (2.7) and (2.9) gives the desired result. □

We now introduce some notation and definitions that will simplify the subsequent derivations. For a random variable X defined on the discrete state space $\mathcal{I}$ let

$$r_X(n) := \begin{cases} 0 & n = 0 \\ \frac{P(X=n-1)}{P(X=n)} & n = 1, 2, \ldots \end{cases}$$

We refer to $r_X(\cdot)$ as the *ratio function* of the random variable X. Note that

$$X \geq_{lr} Y \quad \Leftrightarrow \quad r_X(n) \leq r_Y(n), \qquad n \in \mathcal{I}.$$

We say that a random variable X has the *increasing ratio (IR) property* if $r_X(\cdot)$ is an increasing function over $\mathcal{I}$.[6] Because the stochastic knapsack has objects with heterogeneous sizes, we shall also need the following generalization of the IR property: For a positive integer b a random variable X is said to have the *IR(b) property* if $r_X(n+b) \geq r_X(n)$ for all $n \in \mathcal{I}$ (see Figure 2.6).

Up to this point we have supposed that random variables X and Y have support on the non-negative integers $\mathcal{I}$. Now suppose that $\{i_0, i_1, \ldots, i_M\}$ is the support of X, where $i_0 < i_1 < \cdots < i_M$ are integers and $M \leq \infty$. In this case we define the ratio function for X as

$$r_X(n) = \begin{cases} 0 & n = 0 \\ \frac{P(X=i_{n-1})}{P(X=i_n)} & n = 1, \ldots, M \\ \infty & n > M, \end{cases}$$

and the IR property is defined accordingly. If the random variable Y also has this support, then $X \geq_{lr} Y$ is still meaningfully defined as $r_X(n) \leq r_Y(n)$ for all $n = 1, \ldots, M$. Moreover, it is easily seen that $X \geq_{lr} Y$ continues to imply $X \geq_{st} Y$ with this extended definition of likelihood ratio ordering.

We shall also need to make use of the following result. Its proof is straightforward, but tedious; it can be found in Ross and Yao [140].

[6] In the literature, when X has the increasing ratio property, it is sometimes said to have the Polya frequency of order 2 property.

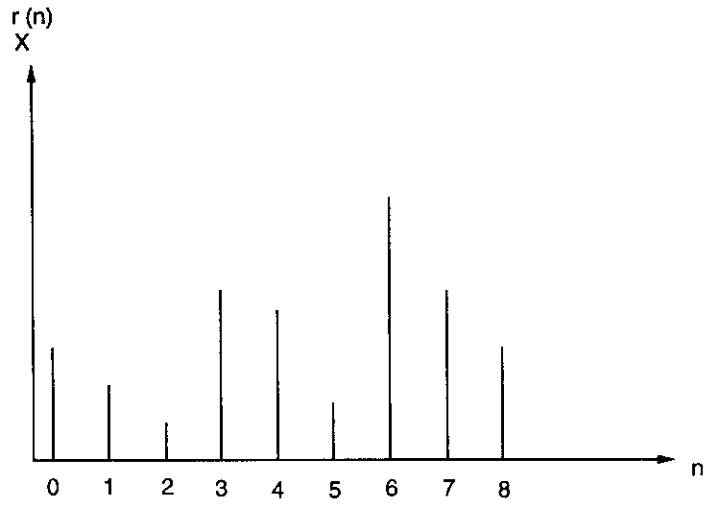

Figure 2.6: The ratio function for a random variable with the IR(3) property. By definition, $r_X(n+3) \geq r_X(n)$ for all n.

Lemma 2.2 *Let $\{Y_k,\ k = 1,\ldots,K-1\}$ be a set of independent random variables, each random variable with the increasing ratio property. Let Y_K and Y'_K be two random variables that are independent of $\{Y_k,\ k = 1,\ldots,K-1\}$ such that $Y_K \geq_{lr} Y'_K$. Let $\{b_k,\ k = 1,\ldots,K\}$ be a set of positive integers such that b_k is a divisor of b_{k+1} for $k = 1,\ldots,K-1$. Let $V := b_1Y_1 + \ldots + b_KY_K$ and $V' := b_1Y_1 + \cdots + b_KY'_K$. Then V and V' have the IR(b_K) property and $V \geq_{lr} V'$.*

Note that the conclusion of Lemma 2.2 applies only to the last index K, not to the other indices $k \neq K$.

2.6 Monotonicity Properties for the Stochastic Knapsack

We need the following notation in order to apply the stochastic comparison concepts of the previous section to the stochastic knapsack. Recall

that X_k denotes the equilibrium number of class-k objects in the knapsack and that $\mathbf{X} := (X_1,\ldots,X_K)$. Let $\{Y_1,\ldots,Y_K\}$ be a collection of independent Poisson random variables with the kth Poisson random variable having parameter ρ_k:

$$P(Y_k = n) = e^{-\rho_k}\frac{\rho_k^n}{n!}, \qquad n \in \mathcal{I}.$$

Note that $r_{Y_k}(n) = n/\rho_k$, so that Y_k has the increasing ratio property. The following result follows directly from the product-form solution for the stochastic knapsack (given in Theorem 2.1).

Corollary 2.2 *For all $\mathbf{n} \in \mathcal{S}$ we have*

$$P(\mathbf{X} = \mathbf{n}) = \frac{\prod_{k=1}^K P(Y_k = n_k)}{P(b_1Y_1 + \cdots + b_KY_K \leq C)}.$$

The random variable Y_k can therefore be thought of as the "unconstrained cousin" of X_k.

Before determining the fundamental qualitative structure of the stochastic knapsack, it is convenient to introduce some additional notation. Recall that $\mathcal{K} = \{1,\ldots,K\}$; any subset $\mathcal{G} \subseteq \mathcal{K}$ is referred to as a *group* of classes. Let

$$U_{\mathcal{G}} := \sum_{k\in\mathcal{G}} b_kX_k, \quad U := U_{\mathcal{K}}, \quad U_{(k)} := U - b_kX_k,$$
$$V_{\mathcal{G}} := \sum_{k\in\mathcal{G}} b_kY_k, \quad V := V_{\mathcal{K}}, \quad V_{(k)} := V - b_kY_k.$$

Note that $U_{\mathcal{G}}$ is the amount of knapsack resource utilized by objects from group $\mathcal{G}$. As mentioned earlier, the random variable U is the utilization of the knapsack; the average utilization is UTIL $= E[U]$. The random variables $V_{\mathcal{G}}$, V, and $V_{(k)}$ can be thought of as the "unconstrained cousins" of $U_{\mathcal{G}}$, U, and $U_{(k)}$, respectively.

Monotonicity with Respect to Offered Load

What happens to TH$_l$ when we increase λ_k? If the object sizes b_k, $k \in \mathcal{K}$, are all equal we might expect TH$_l$ to decrease with λ_k when

$k \neq l$. This conjecture can be false, however, if the object sizes are different: by increasing the arrival rate for class-k objects the blocking of the "wide" objects may increase, allowing for more class-l objects to be admitted. We shall study this issue in some detail. But first consider the behavior of TH_k when λ_k is increased — it is more easily characterized.

Theorem 2.2 *The number of class-k objects in the knapsack, X_k, is increasing with respect to ρ_k in the likelihood ratio ordering. Furthermore, TH_k is increasing in λ_k.*

Proof: From Corollary 2.2 we have

$$P(X_k = n) = \frac{P(Y_k = n)P(V_{(k)} \leq C - b_k n)}{P(V \leq C)}$$

from which we have

$$\frac{P(X_k = n+1)}{P(X_k = n)} = \frac{\rho_k}{n_k + 1} \frac{P(V_{(k)} \leq C - b_k n + b_k)}{P(V_{(k)} \leq C - b_k n)}. \tag{2.10}$$

(Note that the normalization constant, $P(V \leq C)$, has been conveniently cancelled out in the above expression.) Since the distribution of $V_{(k)}$ does not involve ρ_k, the first statement follows directly. For the second statement, recall that likelihood ratio ordering implies stochastic ordering. This implies that $E[X_k]$ is increasing with respect to ρ_k. Hence, $\mu_k E[X_k] = \mathrm{TH}_k$ is increasing in λ_k. □

It is important to note that the above result does not require any restriction on the object sizes, $b_1, \ldots, b_K$. For the remainder of this section we suppose that the following restriction is in force.

Divisibility Condition: For $k = 1, \ldots, K-1$, b_k is a divisor of b_{k+1}.

The following result is of fundamental importance.

Theorem 2.3 *Let $\mathcal{G}$ be a nonempty group of classes and let l be its largest element. Denote $\mathcal{H} = \mathcal{K} - \mathcal{G}$. Then $U_{\mathcal{G}}$ is increasing and $U_{\mathcal{H}}$ is decreasing in ρ_l in the likelihood ratio ordering. In particular, the knapsack utilization, U, is increasing in ρ_K in the likelihood ratio ordering.*

Proof: Let $i_0, i_1, \ldots, i_M$ be the support of $U_{\mathcal{G}}$, where $i_n < i_{n+1}$, $n = 0, \ldots, M-1$. From Corollary 2.2 it follows that

$$P(U_{\mathcal{G}} = i_n) = \frac{P(V_{\mathcal{G}} = i_n)P(V_{\mathcal{H}} \leq C - i_n)}{P(V \leq C)}.$$

In order to apply the theory of likelihood ratio ordering, it is convenient to replace the inequality in the numerator of the above expression with an equality. To this end let Y_0 be a random variable independent of $\{Y_1, \ldots, Y_K\}$ such that

$$P(Y_0 = n) = \begin{cases} a^{-1} & n = 0, \ldots, L \\ a^{-1}\alpha^{n-L} & n \geq L+1, \end{cases}$$

where $a = L + 1/(1-\alpha)$, $0 < \alpha < 1$ and $L \geq C$. Note that Y_0 has the increasing ratio property. Also note that for all $0 \leq d \leq C$

$$P(V_{\mathcal{H}} + Y_0 = d) = \sum_{c=0}^{d} P(V_{\mathcal{H}} = c)P(Y_0 = d - c) = \frac{1}{a} P(V_{\mathcal{H}} \leq d).$$

Thus

$$P(U_{\mathcal{G}} = i_n) = \frac{P(V_{\mathcal{G}} = i_n)P(V_{\mathcal{H}} + Y_0 = C - i_n)}{aP(V \leq C)},$$

from which we have

$$r_{U_{\mathcal{G}}}(n) = \frac{P(U_{\mathcal{G}} = i_{n-1})}{P(U_{\mathcal{G}} = i_n)} = \frac{r_{V_{\mathcal{G}}}(n)}{\prod_{m=i_{n-1}}^{i_n - 1} r_{V_{\mathcal{H}}+Y_0}(C-m)}. \tag{2.11}$$

Now increasing ρ_l will increase Y_l in the likelihood ratio ordering. Thus, by Lemma 2.2, increasing ρ_l will increase $V_{\mathcal{G}}$ in the likelihood ratio ordering. Since $V_{\mathcal{H}} + Y_0$ does not involve ρ_l, it then follows from (2.11) that $U_{\mathcal{G}}$ is increasing with respect to ρ_l in the likelihood ratio ordering. The fact that $U_{\mathcal{H}}$ is decreasing is similarly proved by interchanging $\mathcal{G}$ and $\mathcal{H}$ in (2.11) and by defining $\{i_0, \ldots, i_M\}$ to be the support of $U_{\mathcal{H}}$. □

If the "widest" objects have their arrival rates increased, we might expect them to occupy more room in the knapsack, thereby reducing the presence of objects from the other classes. This intuition can be confirmed by letting $\mathcal{H} = \{k\}$ in Theorem 2.3.

Corollary 2.3 *For $k = 1, \ldots, K-1$, the number of class-k objects in the knapsack, X_k, is decreasing in ρ_K in the likelihood ratio ordering.*

We now need to appeal to the *elasticity property*, which states that

$$\frac{\partial B_l}{\partial \rho_k} = \frac{\partial B_k}{\partial \rho_l} \text{ for all } 1 \leq k, l \leq K. \tag{2.12}$$

The proof of this result is provided in Chapter 5 (Corollary 5.2) for a much more general loss system. Combining Corollary 2.3, the elasticity property, and the equality $\mathrm{TH}_k = \lambda_k(1 - B_k)$ gives the following result.

Corollary 2.4 *For $k = 1, \ldots, K-1$, TH_k is decreasing and B_k is increasing in λ_K. For $k = 1, \ldots, K-1$, TH_K is decreasing and B_K is increasing in λ_k.*

The Divisibility Condition is crucial for the validity of Theorem 2.3 and the subsequent corollaries. To see this, consider the stochastic knapsack with three classes of objects. Suppose that $C = 4$, $b_1 = 1$, $b_2 = 2$, and $b_3 = 3$. We claim that TH_1 can actually increase with λ_3, contradicting Corollary 2.4. The argument goes roughly as follows (it can easily be made rigorous; see the proof of Theorem 2.4). Suppose that $\rho_k := \lambda_k/\mu_k$ is very large for all three classes so that $P(U = 4) \approx 1$. Further suppose that ρ_2 is much larger than $\max(\rho_1, \rho_3)$ so that $P(X_1 = 0, X_2 = 2, X_3 = 1) \approx 1$. In this case, $\mathrm{TH}_1 = \mu_1 E[X_1] \approx 0$. Now keep ρ_1 and ρ_2 fixed, but increase ρ_3 so that it becomes much larger than ρ_2. We will then have $P(X_1 = 1, X_2 = 0, X_3 = 1) \approx 1$ and $\mathrm{TH}_1 = \mu_1 E[X_1] \approx \mu_1$, establishing the claim.

The preceding corollary fails to address the behavior of B_l when ρ_k is increased with $1 \leq k, l \leq K-1$. We would expect B_l to increase with ρ_k when the system is being operated at $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)$ for at least certain values of $\boldsymbol{\rho}$. What is perhaps surprising is that the opposite may be true for other values of $\boldsymbol{\rho}$.

Theorem 2.4 *Fix k, l such that $1 \leq k, l \leq K-1$. In addition to the Divisibility Condition, suppose that b_K is a divisor of C and that $b_K \geq 2b_{K-1}$. Then there exists $\boldsymbol{\rho}^+ = (\rho_1^+, \ldots, \rho_K^+)$ and $\boldsymbol{\rho}^- = (\rho_1^-, \ldots, \rho_K^-)$, with $\rho_k^+ > 0$, $\rho_k^- > 0$, $k = 1, \ldots, K$, such that*

$$\frac{\partial B_l}{\partial \rho_k}(\boldsymbol{\rho}^+) > 0, \qquad \frac{\partial B_l}{\partial \rho_k}(\boldsymbol{\rho}^-) < 0.$$

Proof: Recall that

$$B_l = 1 - G_l/G, \tag{2.13}$$

where G is the normalization constant defined in Section 2.1 and

$$G_l := \sum_{\mathbf{n} \in \mathcal{S}_l} \prod_{k=1}^{K} \frac{\rho_k^{n_k}}{n_k!}.$$

From (2.13) and the definition of the normalization constants it is easily seen that $0 < B_l < 1$ and, because b_k is a divisor of C (since b_k is a divisor of b_K), $B_l \to 1$ as $\rho_k \to \infty$. Hence, there exists a $\boldsymbol{\rho}^+$ at which $\partial B_l/\partial \rho_k$ is strictly positive.

From (2.13) it also follows that $\partial B_l/\partial \rho_k$ is strictly negative if and only if

$$G\frac{\partial G_l}{\partial \rho_k} - G_l\frac{\partial G}{\partial \rho_k} > 0. \tag{2.14}$$

Let $\tilde{\boldsymbol{\rho}} := (0, \ldots, 0, \rho_K)$, where $\rho_K > 0$. It is easily seen that G and $\partial C/\partial \rho_k$ evaluated at $\tilde{\boldsymbol{\rho}}$ are given by

$$G|_{\tilde{\boldsymbol{\rho}}} = \phi\left(\lfloor \frac{C}{b_K} \rfloor\right)$$

and

$$\frac{\partial G}{\partial \rho_k}|_{\tilde{\boldsymbol{\rho}}} = \phi\left(\lfloor \frac{C - b_k}{b_K} \rfloor\right),$$

where $\lfloor x \rfloor$ is the largest integer less than or equal to x and

$$\phi(n) := \sum_{m=0}^{n} \frac{\rho_K^m}{m!}.$$

The same results hold for G_l and $\partial G_l/\partial \rho_k$ with C replaced by $C - b_l$. Thus (2.14) holds true at $\tilde{\boldsymbol{\rho}}$ if and only if

$$\phi\left(\frac{C}{b_K}\right)\phi\left(\lfloor \frac{C - b_l - b_k}{b_K} \rfloor\right) > \phi\left(\lfloor \frac{C - b_l}{b_K} \rfloor\right)\phi\left(\lfloor \frac{C - b_k}{b_K} \rfloor\right). \tag{2.15}$$

Since $B_K \geq 2b_{K-1}$ and $b_k, b_l \leq b_{K-1}$, we have

$$\lfloor \frac{C - b_l - b_k}{b_K} \rfloor = \lfloor \frac{C - b_l}{b_K} \rfloor = \lfloor \frac{C - b_k}{b_K} \rfloor = \frac{C}{b_K} - 1.$$

Therefore (2.15) is equivalent to $\phi(C/b_K) > \phi(C/b_K - 1)$, which clearly holds true. Hence (2.14) holds true at $\tilde{\boldsymbol{\rho}}$.

To complete the proof we note that G, $\partial G/\partial \rho_k$, G_l, and $\partial G_l/\partial \rho_k$ are all continuous functions over of $\boldsymbol{\rho}$ over $[0,\infty)^K$. (This follows immediately from the definition of G and G_l.) Hence (2.14) holds true for some $\boldsymbol{\rho}^-$ such that $\rho_j^- > 0$, $j = 1,\ldots,K$. □

It follows from Corollary 2.4 and Theorem 2.4 that the Jacobian matrix for blocking probabilities takes the form

$$\left[\frac{\partial B_l}{\partial \lambda_k}\right]_{1\le l,k\le K} = \begin{bmatrix} * & * & * & * & + \\ * & * & * & * & + \\ * & * & * & * & + \\ * & * & * & * & + \\ + & + & + & + & + \end{bmatrix}$$

where the $+$, $-$, and $*$ signify that the corresponding term is positive, negative, and positive or negative, respectively. From the above matrix, the identity $\mathrm{TH}_k = \lambda_k(1 - B_k)$, and Theorem 2.2, we also have

$$\left[\frac{\partial TH_l}{\partial \lambda_k}\right]_{1\le l,k\le K} = \begin{bmatrix} + & * & * & * & - \\ * & + & * & * & - \\ * & * & + & * & - \\ * & * & * & + & - \\ - & - & - & - & + \end{bmatrix}.$$

Monotonicity with Respect to Knapsack Capacity

We now consider the behavior of the performance measures when the knapsack capacity is increased. The following result does not require the Divisibility Condition.

Theorem 2.5 *The knapsack utilization, U, increases in the likelihood ratio ordering as the knapsack capacity, C, is increased. Consequently, the average utilization, UTIL, increases with C.*

Proof: For a system with capacity C, let $\{i_0,\ldots,i_M\}$ be the support of U. Let U' denote the utilization of the system with capacity $C+1$,

and let $\{i_0,\ldots,i_{M'}\}$ be the corresponding support, where $M' \ge M$. It is straightforward to show that

$$r_U(n) = r_V(n), \qquad n = 1,\ldots,M$$

and

$$r_{U'}(n) = r_V(n), \qquad n = 1,\ldots,M'.$$

Thus, $r_{U'}(n) \le r_U(n)$, $n = 1,\ldots,M'$, which completes the proof. □

Thus, with respect to utilization, we always gain something when increasing the knapsack capacity. With the monotonicity and divisibility conditions in force, we also have the following complementary result.

Theorem 2.6 *Let $\mathcal{G}$ be a group of classes such that $\phi \subset \mathcal{G} \subset \mathcal{K}$. Denote $\mathcal{H} = \mathcal{K} - \mathcal{G}$ and l the largest element in $\mathcal{H}$. Then increasing C by b_l increases $U_{\mathcal{G}}$ in the likelihood ratio ordering.*

Proof: For the system with capacity C, let $\{i_0,\ldots,i_M\}$ be the support of $U_{\mathcal{G}}$. For the system with capacity $C + b_l$, designate all parameters with a prime (that is, M', $U'_{\mathcal{G}}$, etc.). Also, let Y_0 be defined as in the proof of Theorem 2.3 (with $L \ge C + b_K$). From (2.11) we have

$$r_{U_{\mathcal{G}}}(n) = \frac{r_{V_{\mathcal{G}}}(n)}{\prod_{m=i_{n-1}}^{i_n-1} r_{V_{\mathcal{H}}+Y_0}(C-m)}, \qquad n = 1,\ldots,M,$$

and

$$r_{U'_{\mathcal{G}}}(n) == \frac{r_{V_{\mathcal{G}}}(n)}{\prod_{m=i_{n-1}}^{i_n-1} r_{V_{\mathcal{H}}+Y_0}(C+b_l-m)}, \qquad n = 1,\ldots,M'.$$

Therefore $U_{\mathcal{G}} \le_{lr} U'_{\mathcal{G}}$ if

$$r_{V_{\mathcal{H}}+Y_0}(C+b_l-m) \ge r_{V_{\mathcal{H}}+Y_0}(C-m), \qquad m = 0,\ldots,C.$$

But by Lemma 2.2, $V_{\mathcal{H}} + Y_0$ has the IR(b_l) property, so that the above relation holds true. □

Corollary 2.5 *Suppose the knapsack capacity C is increased by b_K. Then TH_k increases and B_k decreases for all $k \in \mathcal{K}$.*

Proof: With $\mathcal{G} = \{k\}$, it follows from the previous theorem that X_k, $1 \leq k \leq K-1$, increases in likelihood ratio ordering when C is increased by b_K and that X_K increases in the likelihood ratio ordering when C is increased by b_{K-1} and hence by b_K. The proof is then completed by invoking Lemma 2.1 and the identities $\mathrm{TH}_k = \mu_k E[X_k]$ and $\mathrm{TH}_k = \lambda_k(1 - B_k)$. □

Problem 2.3 With the Divisibility Condition in force, give an example illustrating that the throughput of a class can decrease when the knapsack capacity increases by one unit. With the Divisibility Condition *not* in force, give an example illustrating that the throughput of a class can decrease when the knapsack capacity increases by the size of the widest customer.

Problem 2.4 Suppose we operate an ATM system in the statistical multiplexing mode with linear policy $\mathbf{f}$ (see Section 2.3). Let b_k^e, $k \in \mathcal{K}$, denote the associated effective bandwidths. Suppose we then replace $\mathbf{f}$ with a new linear policy $\tilde{\mathbf{f}}$ with effective bandwidths $\tilde{b}_k = ab_k^e$, $k \in \mathcal{K}$, where $0 < a < 1$. Show that the average link utilization increases.

2.7 Asymptotic Analysis of the Stochastic Knapsack

Since the stochastic knapsack typically has a large capacity for applications of practical interest, we are motivated to study its asymptotic behavior as its capacity is increased to infinity. With any luck, the algorithms to calculate blocking probabilities and the qualitative theory for monotonicity will simplify.

For a natural asymptotic regime, we shall see that the blocking probability of an object is proportional to its size. This appealing result can be used as a "rule of thumb" for dimensioning the loss system. We shall also see, in contrast with an earlier monotonicity result, that blocking always increases in an asymptotic sense when the offered loads increase.

The Erlang Loss System

What is an interesting and meaningful asymptotic regime for the Erlang loss system? Because transmission capacity has been growing by leaps and bounds, a regime with $C \to \infty$ is compelling. But while increasing C the regime should also increase ρ, not only because call volumes have been rapidly growing, but also because blocking would approach zero very quickly if ρ were held fixed.

Consider a sequence of Erlang loss systems indexed by $C = 1, 2, \ldots$, where the Cth system has capacity C and offered load $\rho^{(C)}$. We focus our attention on asymptotic regimes for which both C and $\rho^{(C)}$ go to infinity; furthermore, we assume that the following limit exists:

$$\rho^* := \lim_{C \to \infty} \frac{\rho^{(C)}}{C}.$$

Let $B(C)$ denote the blocking probability for the Cth system. It is well known (for example, see [151]) that

$$\lim_{C \to \infty} B(C) = \begin{cases} 0 & \rho^* \leq 1 \\ 1 - 1/\rho^* & \rho^* > 1. \end{cases}$$

This result has an appealing fluid interpretation. Consider a pipe of capacity 1 to which a flow ρ^* is offered. If the flow is no greater than the pipe capacity (that is, $\rho^* \leq 1$), then the entire flow passes and there is no blocking. On the other hand, if the flow is greater than the pipe capacity (that is, $\rho^* > 1$), the excess flow $\rho^* - 1$ fails to pass through the pipe, so that the fraction $(\rho^* - 1)/\rho^* = 1 - 1/\rho^*$ is blocked.

It is also well known that $B(C)$ converges to zero exponentially fast (for example, see [81]) when $\rho^* < 1$. Although blocking also goes to zero for the case $\rho^* = 1$, we shall see that it does so relatively slowly. It is also known that when $\rho^* > 1$ the number of free resource units approaches a geometric distribution with parameter $1/\rho^*$ — that is, the probability that there are c free resource units approaches $(1 - 1/\rho^*)(1/\rho^*)^c$.

These results suggest the following guideline: we should dimension the capacity so that $C \approx \rho$ when ρ is large. Indeed, if $C >> \rho$ then the system is overdimensioned because C can be reduced without significantly increasing the blocking probability. If $C << \rho$, the system is underdimensioned because blocking probability is large.

The asymptotic regime is said to be *under loaded, critically loaded,* or *over loaded* depending on whether $\rho^* < 1$, $\rho^* = 1$, or $\rho^* > 1$. Assuming that the Erlang loss system is to be well-dimensioned, the case of critical loading is of greatest practical interest; thus we hereafter suppose that $\rho^{(C)}/C$ converges to 1. In particular, we suppose that

$$\frac{\rho^{(C)}}{C} = 1 - \frac{\alpha}{\sqrt{C}}, \tag{2.16}$$

where α is an arbitrary, but fixed real number.

Let $X(C)$ be the equilibrium number of objects in the Cth system, and consider the asymptotic behavior of $X(C)$ as $C \to \infty$. Because $\rho^{(C)} \to \infty$ we expect $X(C) \to \infty$, which is not a terribly interesting result. Therefore, to gain some insight into the system's asymptotic behavior we need to normalize $X(C)$ before taking the limit. To this end, consider the normalized random variable

$$\hat{X}(C) := \frac{X(C) - \rho^{(C)}}{\sqrt{C}}.$$

Owing to (2.16), we can write the normalized random variable as

$$\hat{X}(C) = \frac{X(C)}{\sqrt{C}} + \alpha - \sqrt{C}.$$

Since $0 \le X(C) \le C$, it follows that

$$\alpha - \sqrt{C} \le \hat{X}(C) \le \alpha.$$

Therefore, if the distribution of $\hat{X}(C)$ converges to the distribution of a random variable $\hat{X}$, then we would certainly expect $P(\hat{X} \le \alpha) = 1$.

This specific normalization is chosen because the mean and the variance of $\hat{X}(C)$ remain finite in the limit, a result we will state more precisely in the subsequent theorem. But first we need some new notation. For any random variable Y, denote $F_Y(\cdot)$ for its distribution function and $f_Y(\cdot)$ for its density function. Let Z be the standard normal random variable. Let $\hat{X}$ be the random variable whose density is that of Z conditioned on the event $\{Z \le \alpha\}$, that is,

$$f_{\hat{X}}(x) = \begin{cases} f_Z(x)/P(Z \le \alpha) & x \le \alpha \\ 0 & x > \alpha. \end{cases}$$

The following result is well known (for example, see [159], [87], or [151]) and can be proved by taking the limit of the known distribution of $\hat{X}(C)$ and applying Stirling's formula.

Theorem 2.7 *Suppose the asymptotic regime is critically loaded. Then distribution of $\hat{X}(C)$ converges to the distribution of $\hat{X}$. Furthermore, the moments of $\hat{X}(C)$ converge to the corresponding moments of $\hat{X}$.*

The above result enables us to determine the rate at which $B(C)$ converges to zero. [7]

Corollary 2.6 *If the asymptotic regime is critically loaded, then*

$$B(C) = \frac{\delta}{\sqrt{C}} + o(1/\sqrt{C}),$$

where

$$\delta := \frac{f_Z(\alpha)}{F_Z(\alpha)}.$$

Proof: From Little's formula and the definition of $\hat{X}(C)$ we have

$$B(C) = 1 - \frac{E[X(C)]}{\rho^{(C)}} = -\frac{\sqrt{C}E[\hat{X}(C)]}{\rho^{(C)}}.$$

Thus, by critical loading and Theorem 2.7,

$$\lim_{C\to\infty} \sqrt{C}B(C) = -E[\hat{X}],$$

where

$$E[\hat{X}] = \frac{\int_{-\infty}^{\alpha} x f_Z(x)\mathrm{d}x}{\int_{-\infty}^{\alpha} f_Z(x)\mathrm{d}x} = \frac{-f_Z(\alpha)}{F_Z(\alpha)} = -\delta. \quad \square$$

Thus, neglecting the term $o(1/\sqrt{C})$, we can say that $B(C)$ converges to zero at rate $\delta/\sqrt{C}$ when the offered load satisfies (2.16). Note that this rate sharply contrasts with the exponential convergence rate for the case $\rho^* < 1$.

[7] A function $f(n)$ is said to be $o(g(n))$ if $\lim_{n\to\infty} f(n)/g(n) = 0$.

The Stochastic Knapsack

Now consider a sequence of stochastic knapsacks with the Cth knapsack having capacity C and offered loads $\rho_k^{(C)}$, $k \in \mathcal{K}$. We assume that the object sizes are held fixed at b_k, $k \in \mathcal{K}$, for each of these knapsacks. Let $B_k(C)$ be the probability of blocking a class-k object for the Cth knapsack. The following discussion of the asymptotic behavior of the stochastic knapsack will closely parallel the discussion for the Erlang loss system.

We consider asymptotic regimes for which the following limits exist:

$$\rho_k^* = \lim_{C\to\infty} \frac{\rho_k^{(C)}}{C}, \quad k \in \mathcal{K}.$$

Let

$$\rho^* := \sum_{k=1}^{K} b_k \rho_k^*.$$

For the stochastic knapsack, the under loaded, critically loaded, and over loaded regimes are defined with respect to ρ^* exactly as they are for the Erlang loss system.

Kelly [87] studied this asymptotic regime for a class of loss systems for which the stochastic knapsack is a special case. It follows from his results that

$$\lim_{C\to\infty} B_k(C) = \begin{cases} 0 & \rho^* \le 1 \\ 1-(1-a^*)^{b_k} & \rho^* > 1, \end{cases}$$

where a^* is the unique solution to

$$\sum_{k=1}^{K} b_k \rho_k^* (1-a^*)^{b_k} = 1. \tag{2.17}$$

As with the Erlang loss system, there is an interesting fluid interpretation for this result. In this fluid interpretation, the offered flow of class-k objects is ρ_k^* per unit time and each object consists of b_k atoms; hence the total offered flow of atoms is ρ^* per unit time. The pipe again has capacity 1, so that it can admit atoms at a rate of 1 per unit time. If $\rho^* \le 1$, all atoms pass, and hence there is no blocking. Now suppose

$\rho^* > 1$, so that the pipe overflows and blocks atoms. Let a^* denote the probability that an atom is blocked. Assuming that a class-k object is admitted if and only if all its atoms are admitted, and assuming that atoms are blocked independently of each other, then the probability that a class-k object is admitted is $(1-a^*)^{b_k}$. Furthermore, since the offered flow of class-k atoms is $b_k\rho_k^*$ per unit time, the admitted flow for class-k atoms is $b_k\rho_k^*(1-a^*)^{b_k}$, the sum of which should be equal to the capacity of the pipe; hence, with this fluid interpretation, a^* is the solution to (2.17).

As for the Erlang loss model, we can argue that the critical loaded case, $\rho^* = 1$, is of greatest practical interest. Focusing on critical loading, we further assume that

$$\frac{\sum_{k=1}^{K} b_k \rho_k^{(C)}}{C} = 1 - \frac{\alpha}{\sqrt{C}}, \tag{2.18}$$

where α is a fixed, but arbitrary real number. Let $X_k(C)$ be the number of class-k objects in the knapsack in equilibrium for the Cth knapsack. Let

$$\hat{X}_k(C) := \frac{X_k(C) - \rho_k^{(C)}}{\sqrt{C}}$$

be the associated normalized random variable. Combining (2.18) with the fact $0 \le \sum_{k=1}^{K} b_k X_k(C) \le C$ gives

$$\alpha - \sqrt{C} \le \sum_{k=1}^{K} b_k \hat{X}_k(C) \le \alpha.$$

Thus if the distribution of $\hat{\mathbf{X}}(C) := (\hat{X}_1(C), \ldots, \hat{X}_K(C))$ converges to the distribution of a random vector $\hat{\mathbf{X}} = (\hat{X}_1, \ldots, \hat{X}_K)$, we would certainly expect $P(\mathbf{b}\cdot\hat{\mathbf{X}} \le \alpha) = 1$. Let $\mathbf{Y} := (Y_1, \ldots, Y_K)$ be a vector of independent random variables, where Y_k has the normal distribution with mean 0 and variance ρ_k^*. Let $\hat{\mathbf{X}} := (\hat{X}_1, \ldots, \hat{X}_K)$ be a vector of random variables with joint density function given by

$$f_{\hat{\mathbf{X}}}(\mathbf{x}) = \begin{cases} f_{\mathbf{Y}}(\mathbf{x})/P(\mathbf{b}\cdot\mathbf{Y} \le \alpha) & \text{if } \mathbf{b}\cdot\mathbf{x} \le \alpha \\ 0 & \text{otherwise.} \end{cases}$$

Note that the distribution of $\hat{\mathbf{X}}$ is that of $\mathbf{Y}$ conditioned on $\{\mathbf{b}\cdot\mathbf{Y} \le \alpha\}$. The proof of the following result, which again involves taking the limit

of the known distribution for $\hat{\mathbf{X}}(C)$ and applying Stirling's formula, can be found in Kelly [87].

Theorem 2.8 *Suppose the asymptotic regime is critically loaded. The distribution of $\hat{\mathbf{X}}(C)$ converges to the distribution of $\hat{\mathbf{X}}$. Furthermore, the moments of $\hat{\mathbf{X}}(C)$ converge to the respective moments of $\hat{\mathbf{X}}$.*

The above result enables us to determine the asymptotic behavior of the blocking probabilities. Again let Z denote the standard normal random variable.

Corollary 2.7 *For all $k \in \mathcal{K}$,*

$$B_k(C) = \frac{b_k \delta}{\sqrt{C}} + o(1/\sqrt{C}),$$

where

$$\delta := \frac{f_Z(\alpha/\sigma)}{\sigma F_Z(\alpha/\sigma)}$$

and

$$\sigma^2 := \sum_{k=1}^{K} b_k^2 \rho_k^*.$$

Proof: Mimicking the proof of Corollary 2.6, we obtain

$$B_k(C) = -\frac{\sqrt{C} E[\hat{X}_k(C)]}{\rho_k^{(C)}}.$$

Thus, from Theorem 2.8,

$$\lim_{C \to \infty} \sqrt{C} B_k(C) = -\frac{E[\hat{X}_k]}{\rho_k^*}, \tag{2.19}$$

where

$$\begin{aligned} E[\hat{X}_k] &= \frac{E[Y_k 1(\mathbf{b} \cdot \mathbf{Y} \le \alpha)]}{P(\mathbf{b} \cdot \mathbf{Y} \le \alpha)} \\ &= \frac{\int_{-\infty}^{\alpha} E[Y_k | \mathbf{b} \cdot \mathbf{Y} = x] f_{\mathbf{b} \cdot \mathbf{Y}}(x) \mathrm{d}x}{P(\mathbf{b} \cdot \mathbf{Y} \le \alpha)}. \end{aligned}$$

Since $\mathbf{b} \cdot \mathbf{Y}$ has a normal distribution with mean 0 and variance σ^2, it follows that

$$P(\mathbf{b} \cdot \mathbf{Y} \le \alpha) = F_Z(\alpha/\sigma)$$

and that (for example, see Feller [47], page 72)

$$E[Y_k | \mathbf{b} \cdot \mathbf{Y} = x] = \frac{\mathrm{cov}(Y_k, \mathbf{b} \cdot \mathbf{Y})}{\mathrm{var}(\mathbf{b} \cdot \mathbf{Y})} x = \frac{b_k \rho_k^*}{\sigma^2} x.$$

Thus,

$$\begin{aligned} E[\hat{X}_k] &= \frac{b_k \rho_k^*}{\sigma^2 F_Z(\alpha/\sigma)} \int_{-\infty}^{\alpha} x f_{\mathbf{b} \cdot \mathbf{Y}}(x) \mathrm{d}x \\ &= -\frac{b_k \rho_k^* f_Z(\alpha/\sigma)}{\sigma F_Z(\alpha/\sigma)} = -b_k \rho_k^* \delta. \end{aligned} \tag{2.20}$$

Combining (2.19) and (2.20) gives the desired result. □

Neglecting the term $o(1/\sqrt{C})$, Corollary 2.7 implies that the *blocking probability of an object is asymptotically proportional to its size.* Moreover, as for the Erlang loss system, the blocking probabilities converge to zero with order $1/\sqrt{C}$.

An Approximation Procedure

Corollary 2.7 leads to a natural procedure to approximate blocking probabilities for a knapsack with fixed C and ρ_k's. First we solve for α in (2.18):

$$\alpha \leftarrow \sqrt{C} - \frac{\sum_{k=1}^{K} b_k \rho_k}{\sqrt{C}}.$$

We also let

$$\sigma^2 \leftarrow \frac{\sum_{k=1}^{K} b_k^2 \rho_k}{C}.$$

Next, we obtain δ as specified in Corollary 2.7. We then approximate blocking probabilities with an explicit and simple formula:

$$B_k \approx b_k \frac{\delta}{\sqrt{C}}, \qquad k \in \mathcal{K}.$$

As an example, consider three stochastic knapsacks each with $K = 3$, $b_1 = 1$, $b_2 = 6$, and $b_3 = 24$. The capacities and offered loads for the three knapsacks are specified in Table 2.1. A simple calculation gives $\alpha = 0$ and $\sigma^2 = 31/3$ for all three knapsacks. Table 2.2 compares the exact blocking probabilities, obtained from the recursive algorithm, with the approximate blocking probabilities. Note that the approximation is quite accurate, particularly for large C.

	C	ρ_k
Knapsack I	10	$10/3b_k$
Knapsack II	100	$100/3b_k$
Knapsack III	1,000	$1,000/3b_k$

Table 2.1: Parameters for three knapsacks.

		B_1	B_2	B_3
Knapsack I	Exact	.073	.467	1.00
	Approx	.078	.471	1.88
Knapsack II	Exact	.022	.135	.544
	Approx	.025	.149	.596
Knapsack III	Exact	.008	.046	.183
	Approx	.008	.047	.188

Table 2.2: Exact versus approximate blocking probabilities.

Reiman [120] gives additional computational results. His results show that the approximation remains accurate if $\alpha \neq 0$ as long as $|\alpha|$ is small (roughly less than 1).

Asymptotic Monotonicity

Recall from Section 2.6 that the stochastic knapsack exhibits bizarre behavior as the offered loads are varied: the partial derivative $\partial B_l/\partial B_k$ can be either positive or negative for a fixed k and l. Following the arguments of Reiman [120] we now show that this bizarre behavior vanishes in a certain sense as the knapsack capacity becomes very large.

We first introduce the following more explicit assumption concerning the growth rates of the offered loads:

$$\frac{\rho_k^{(C)}}{C} = \rho_k^* + \frac{\beta_k}{\sqrt{C}}, \qquad k \in \mathcal{K},$$

where β_k, $k \in \mathcal{K}$, are fixed but arbitrary real numbers. Compatibility with (2.18) requires

$$\alpha = -\sum_{k=1}^{K} b_k \beta_k.$$

What is the behavior of the blocking probabilities as the offered loads are increased? Since we must keep $\sum_{k=1}^{K} b_k \rho_k^*$ equal to 1 to remain in the critical loaded regime, we shall increase the offered load for class k by increasing the parameter β_k. Of course, increasing β_k will decrease α.

From Corollary 2.7 we know that the asymptotic normalized blocking probability for class l is given by

$$\hat{B}_l := \lim_{C\to\infty} \sqrt{C} B_l(C) = b_l \delta.$$

From the definition of δ it follows that

$$\hat{B}_l = \frac{b_l}{\sigma} h\left(\frac{1}{\sigma}\sum_{i=1}^{K} b_i \beta_i\right),$$

where $h(\cdot)$ is the "hazard rate" of the standard normal distribution, that is,

$$h(x) := \frac{f_Z(x)}{1 - F_Z(x)}.$$

It is known that $h(\cdot)$ is strictly increasing, so that

$$\frac{\partial \hat{B}_l}{\partial \beta_k} = \frac{b_k b_l}{\sigma^2} h'\left(\frac{1}{\sigma}\sum_{i=1}^{K} b_i \beta_i\right) > 0, \quad 1 \leq k, l \leq K.$$

Thus, the Jacobian matrix for the normalized asymptotic blocking probabilities has an appealing form:

$$\left[\frac{\partial \hat{B}_l}{\partial \beta_k}\right]_{1\leq l,k\leq K} = \begin{bmatrix} + & + & + & + \\ + & + & + & + \\ + & + & + & + \\ + & + & + & + \end{bmatrix}.$$

2.8 The Stochastic Knapsack with Continuous Sizes*

The stochastic knapsack model requires the object sizes to take on values in a finite set. In this section we allow the object sizes to take on any value in the continuum $(0, C]$.[8]

We suppose that the knapsack has size C and that objects arrive according to a Poisson process. Denote λ for the arrival rate of objects. The sizes of arriving objects are independent and identically distributed random variables. Denote

$$F(b), \qquad 0 < b \leq C,$$

for the distribution function of the size of an arriving object. An arriving object of size b is admitted into the knapsack if and only if b or more resource units are available. Object sojourn times are independent and exponentially distributed with mean $1/\mu$.

The above paragraph formally defines the stochastic knapsack with continuous sizes. Note that if $F(\cdot)$ is concentrated on a finite set, then the model becomes an ordinary stochastic knapsack.

Let $L(t)$ denote the number of objects in the knapsack at time t. Let $b_1, b_2, \ldots, b_{L(t)}$ denote the sizes of the objects in the knapsack at time t. The state of the knapsack at time t is the unordered set

$$\xi(t) = \{b_1, \ldots, b_{L(t)}\}.$$

The state space is

$$\mathcal{S} = \cup_{l \in \mathcal{I}} S_l,$$

where

$$S_0 = \phi$$

and for $l \geq 0$

$$S_l := \{\{b_1, \ldots, b_l\}, \ 0 \leq b_i \leq C, \ i = 1, \ldots, l\}.$$

This knapsack model with parameters λ, μ, and $F(\cdot)$ (and an initial state) define a Markov process $\{\xi(t)\}$ taking values in $\mathcal{S}$.

[8]Sections with an asterisk (*) can be skipped on first reading.

Let $\rho := \lambda/\mu$. Let $1(A)$ denote the indicator function for the event A, that is, $1(A) = 1$ if A is true and $1(A) = 0$ if A is false. Let h be a random variable for the sample space $\mathcal{S}$, that is, h is a (measurable) function from $\mathcal{S}$ to the reals. The following result is due to Robert [124].

Theorem 2.9 *In equilibrium and for any random variable h, the expected value of h for the stochastic knapsack with continuous sizes is*

$$E[h] = \frac{1}{G}[h(\phi) + \sum_{l=1}^{\infty} \frac{\rho^l}{l!} \int_{[0,\infty]^l} h(\{b_1, \ldots, b_l\}) 1(b_1 + \cdots + b_l \leq C) \prod_{i=1}^{l} \mathrm{d}F(b_i) \}, \tag{2.21}$$

where

$$G := \sum_{l=0}^{\infty} \frac{\rho^l}{l!} \sigma_l(C)$$

and

$$\sigma_l(c) = \begin{cases} \int_{[0,\infty]^l} 1(b_1 + \cdots + b_l \leq c) \prod_{i=1}^{l} \mathrm{d}F(b_i) & l \geq 1 \\ 1 & l = 0. \end{cases}$$

Before proving Theorem 2.9 we illustrate the result with some examples. If we set $h(\phi) = 0$ and

$$h(\{b_1, \ldots, b_l\}) = 1(l = m),$$

it follows from Theorem 2.9 that

$$\begin{aligned} P(\text{“}m \text{ objects in knapsack”}) &= \frac{1}{G} \frac{\rho^m}{m!} \sigma_m(C) \\ &= \frac{\frac{\rho^m}{m!} \sigma_m(C)}{\sum_{l=0}^{\infty} \frac{\rho^l}{l!} \sigma_l(C)}. \end{aligned}$$

When the knapsack is in state $\{b_1, \ldots, b_l\}$, the objects utilize $b_1 + \cdots + b_l$ resource units. Let U denote the random variable for knapsack utilization. If we set $h(\phi) = 1$ and

$$h(\{b_1, \ldots, b_l\}) = 1(b_1 + \cdots + b_l \leq c),$$

then it follows from Theorem 2.9 that

$$P(U \le c) = \frac{\sum_{l=0}^{\infty} \frac{\rho^l}{l!}\sigma_l(c)}{\sum_{l=0}^{\infty} \frac{\rho^l}{l!}\sigma_l(C)}. \tag{2.22}$$

It is important to note that $P(U = 0) = 1/G > 0$. Also note that we can obtain a good approximation for $P(U \le c)$ by truncating the sums in (2.22) to a finite limit; indeed the terms satisfy the bound

$$\frac{\rho^l}{l!}\sigma_l(c) \le \frac{[\rho F(c)]^l}{l!},$$

which decays rapidly.

There is a useful probabilistic interpretation of (2.22). Let U_l, $l \ge 0$, be independent random variables each with distribution function $F(\cdot)$. Let L be a Poisson random variable with parameter ρ, independent of U_l, $l \ge 0$. Then

$$\begin{aligned} P\left(\sum_{l=1}^{L} U_l \le c\right) &= e^{-\rho}\left[1 + \sum_{l=1}^{\infty} \frac{\rho^l}{l!} \int_{[0,\infty]^l} 1(b_1 + \cdots + b_l \le c) \prod_{i=1}^{l} \mathrm{d}F(b_i)\right] \\ &= e^{-\rho} \sum_{l=0}^{\infty} \frac{\rho^l}{l!}\sigma_l(c). \end{aligned}$$

Thus from (2.22), the knapsack utilization is a normalized random sum of independent random variables:

$$P(U \le c) = \frac{P(U_1 + \cdots + U_L \le c)}{P(U_1 + \cdots + U_L \le C)}. \tag{2.23}$$

We also note that with $\mu_k = \mu$ for all $k \in \mathcal{K}$, Theorem 2.9 is a generalization of the product-form result of Theorem 2.1 for the stochastic knapsack. Specifically, if we concentrate $F(\cdot)$ on $\{b_1, \ldots, b_K\}$, then Theorem 2.9 specializes to Theorem 2.1.

Proof of Theorem 2.9 Let $\mathcal{H}$ be the set of all (measurable) real-valued functions defined on $\mathcal{S}$. Let $\mathbf{Q} : \mathcal{H} \to \mathcal{H}$ be the infinitesimal generator [122] for the Markov process $\{\xi_t,\ t \ge 0\}$, that is, for $\xi = \{b_1, \ldots, b_l\}$,

$$\begin{aligned} \mathbf{Q}(h)(\xi) &= \mu \sum_{i=1}^{l} [h(\xi - \{b_i\}) - h(\xi)] \\ &+ \lambda \int_0^{\infty} [h(\xi \cup \{b\}) - h(\xi)] 1(b_1 + \cdots + b_l + b \le C) \mathrm{d}F(b). \end{aligned}$$

Let h^1 be the function on $\mathcal{S}$ defined by $h^1(\xi) = 1$ for all $\xi \in \mathcal{S}$. It suffices to show [122] that (2.21) satisfies

$$E[h^1] = 1 \tag{2.24}$$

and that for all $h \in \mathcal{H}$

$$E[\mathbf{Q}(h)] = 0. \tag{2.25}$$

It is convenient to introduce the following notation:

$$\mathbf{b}_l = \{b_1, \ldots, b_l\},$$

$$1(\mathbf{b}_l) = 1(b_1 + \cdots + b_l \le C),$$

and

$$\mathrm{d}F(\mathbf{b}_l) = \prod_{i=1}^{l} \mathrm{d}F(b_i).$$

We first establish (2.24). Replacing h with h^1 in (2.21) gives

$$\begin{aligned} E[h^1] &= \frac{1}{G}\left[1 + \sum_{l=1}^{\infty} \frac{\rho^l}{l!} \int_{[0,\infty]^l} 1(\mathbf{b}_l) \mathrm{d}F(\mathbf{b}_l)\right] \\ &= \frac{G}{G} = 1. \end{aligned}$$

Turning now to (2.25), first note

$$\mathbf{Q}(h)(\phi) = \lambda \int_0^C h(\{b\}) \mathrm{d}F(b) - \lambda h(\phi).$$

Replacing h with $\mathbf{Q}(h)$ in (2.21) and invoking the above equation gives

$$\begin{aligned} G \cdot E[\mathbf{Q}(h)] &= \lambda \int_0^C h(\{b\}) \mathrm{d}F(b) - \lambda h(\phi) \\ &+ \sum_{l=1}^{\infty} \frac{\rho^l}{l!} \int_{[0,\infty]^l} \mathbf{Q}(h)(\mathbf{b}_l)\ 1(\mathbf{b}_l) \mathrm{d}F(\mathbf{b}_l) \\ &= \lambda \int_0^C h(\{b\}) \mathrm{d}F(b) - \lambda h(\phi) \\ &+ \sum_{l=1}^{\infty} \mu \frac{\rho^l}{l!} \sum_{i=1}^{l} \int_{[0,\infty]^l} [h(\mathbf{b}_l - \{b_i\}) - h(\mathbf{b}_l)] 1(\mathbf{b}_l) \mathrm{d}F(\mathbf{b}_l) \\ &+ \sum_{l=1}^{\infty} \lambda \frac{\rho^l}{l!} \int_{[0,\infty]^{l+1}} [h(\mathbf{b}_{l+1}) - h(\mathbf{b}_l)] 1(\mathbf{b}_{l+1}) \mathrm{d}F(\mathbf{b}_{l+1}). \end{aligned}$$

Considering the symmetry along coordinates, we obtain

$$G \cdot E[\mathbf{Q}(h)] = \lambda \int_0^C h(\{b\})\mathrm{d}F(b) \tag{2.26}$$
$$- \lambda h(\phi) \tag{2.27}$$
$$+ \sum_{l=1}^{\infty} \mu l \frac{\rho^l}{l!} \int_{[0,\infty]^l} h(\mathbf{b}_{l-1})1(\mathbf{b}_l)\mathrm{d}F(\mathbf{b}_l) \tag{2.28}$$
$$- \sum_{l=1}^{\infty} \mu l \frac{\rho^l}{l!} \int_{[0,\infty]^l} h(\mathbf{b}_l)1(\mathbf{b}_l)\mathrm{d}F(\mathbf{b}_l) \tag{2.29}$$
$$+ \sum_{l=1}^{\infty} \lambda \frac{\rho^l}{l!} \int_{[0,\infty]^{l+1}} h(\mathbf{b}_{l+1})1(\mathbf{b}_{l+1})\mathrm{d}F(\mathbf{b}_{l+1}) \tag{2.30}$$
$$- \sum_{l=1}^{\infty} \lambda \frac{\rho^l}{l!} \int_{[0,\infty]^{l+1}} h(\mathbf{b}_l)1(\mathbf{b}_{l+1})\mathrm{d}F(\mathbf{b}_{l+1}). \tag{2.31}$$

Making the change of variables $l \leftarrow l+1$ in (2.28) and adding it to (2.27) plus (2.31) gives

$$\sum_{l=0}^{\infty} \mu\rho \frac{\rho^l}{l!} \int_{[0,\infty]^{l+1}} h(\mathbf{b}_l)1(\mathbf{b}_{l+1})\mathrm{d}F(\mathbf{b}_{l+1})$$
$$- \sum_{l=0}^{\infty} \lambda \frac{\rho^l}{l!} \int_{[0,\infty]^{l+1}} h(\mathbf{b}_l)1(\mathbf{b}_{l+1})\mathrm{d}F(\mathbf{b}_{l+1})$$
$$= \sum_{l=0}^{\infty} (\mu\rho - \lambda) \frac{\rho^l}{l!} \int_{[0,\infty]^{l+1}} h(\mathbf{b}_l)1(\mathbf{b}_{l+1})\mathrm{d}F(\mathbf{b}_{l+1}) = 0,$$

where the last equality follows from $\mu\rho - \lambda = 0$. Similarly, making the same change in variables in (2.29) and adding it to (2.26) plus (2.30) gives zero. Hence $E[\mathbf{Q}(h)] = 0$. □

We now show how to use Theorem 2.9 to obtain explicit formulas for blocking probabilities for several interesting distribution functions $F(\cdot)$.

Uniform Distribution

Suppose that the object weights are uniformly distributed over $(0, C)$, that is,

$$F(b) = \frac{c}{C}, \qquad 0 < b < C.$$

Then

$$\sigma_l(c) = \frac{1}{C^l} \int_{[0,\infty]^l} 1(b_1 + \cdots + b_l < c)\mathrm{d}b_1 \cdots \mathrm{d}b_l$$
$$= \frac{(c/C)^l}{l!}$$

and, consequently,

$$\sum_{l=0}^{\infty} \frac{\rho^l}{l!}\sigma_l(c) = \sum_{l=0}^{\infty} \frac{(\rho c/C)^l}{(l!)^2} = J_0(2\sqrt{\rho c/C}),$$

where $J_0(\cdot)$ is the modified Bessel function of the first kind of order 0. Thus from (2.22) the distribution function for the knapsack utilization takes on an explicit form:

$$P(U \le c) = \frac{J_0(2\sqrt{\rho c/C})}{J_0(2\sqrt{\rho})}.$$

Hence the probability of blocking an object of size b is

$$1 - P(U \le C - b) = 1 - \frac{J_0(2\sqrt{\rho(C-b)/C})}{J_0(2\sqrt{\rho})}.$$

Gamma Distribution

For $\nu > 0$, let

$$\Gamma(\nu) = \int_0^{\infty} x^{\nu-1}e^{-x}\mathrm{d}x$$

be the gamma function. For $\nu = m$, with m an integer, we have $\Gamma(m+1) = m!$. The gamma density with parameters $\nu > 0$, $\alpha > 0$ is

$$f_{\alpha,\nu}(x) = \frac{1}{\Gamma(\nu)}\alpha^{\nu}x^{\nu-1}e^{-\alpha x}, \quad x > 0.$$

Many interesting shapes can be obtained with gamma densities (see Kleinrock [97], p. 124).

Now suppose that the object weights have a gamma density. Specifically, let U_l, $l \geq 0$, be independent random variables, each having the gamma density with parameters ν and α. It is well known that

$$E[U_i] = \frac{\nu}{\alpha} \qquad \text{var}[U_i] = \frac{\nu}{\alpha^2}$$

and that $U_1 + \cdots + U_l$ also has the gamma density with parameters $l\nu$ and α. Thus

$$\begin{aligned} P\left(\sum_{i=1}^{l} U_i \leq c\right) &= \int_0^c f_{\alpha,l\nu}(x)\mathrm{d}x \\ &= \frac{\alpha^{l\nu}}{\Gamma(l\nu)} \int_0^c x^{l\nu-1} e^{-\alpha x}\mathrm{d}x, \end{aligned}$$

from which we obtain

$$P\left(\sum_{l=1}^{L} U_i \leq c\right) = e^{-\rho}\left[1 + \sum_{l=1}^{\infty} \frac{\rho^l}{l!} \frac{1}{\Gamma(l\nu)} \alpha^{l\nu} \int_0^c x^{l\nu-1} e^{-\alpha x}\mathrm{d}x\right],$$

where L is an independent Poisson random variable with parameter ρ. Hence from (2.23) the distribution of the knapsack utilization is given by

$$P(U \leq c) = \frac{1 + \sum_{l=1}^{\infty} \frac{(\alpha^\nu \rho)^l}{l!\Gamma(l\nu)} \int_0^c x^{l\nu-1} e^{-\alpha x}\mathrm{d}x}{1 + \sum_{l=1}^{\infty} \frac{(\alpha^\nu \rho)^l}{l!\Gamma(l\nu)} \int_0^C x^{l\nu-1} e^{-\alpha x}\mathrm{d}x}. \tag{2.32}$$

Because the probability of blocking an object of size b is $P(U > C - b)$, the above formula can calculate it.

When $\nu = m$, with m an integer, the gamma density becomes the Erlang density:

$$f_{\alpha,m}(x) = \frac{\alpha(\alpha x)^{m-1}}{(m-1)!} e^{-\alpha x}, \quad x > 0.$$

In this case the integrals in (2.32) have a closed-form solution. After some calculation, we obtain

$$P(U \leq c) = \frac{e^\rho + e^{-\alpha c} \sum_{l=1}^{\infty} \frac{\rho^l}{l!} \sum_{j=0}^{lm-1} \frac{(\alpha c)^j}{j!}}{e^\rho + e^{-\alpha C} \sum_{l=1}^{\infty} \frac{\rho^l}{l!} \sum_{j=0}^{lm-1} \frac{(\alpha C)^j}{j!}}$$

We shall generalize this continuous-size model in Chapter 3 to knapsacks with state-dependent arrival and departure rates and in Chapter 5 to networks of knapsacks.

2.9 Bibliographical Notes

It is not clear to whom the product-form result for the stochastic knapsack (Theorem 2.1) should be attributed. It may have been known to Jensen or even Erlang [21]. It is an obvious consequence of the theory developed in Kelly's book [89], but is not stated explicitly there. Apparently the result was not widely known until 1981 when two papers devoted to the stochastic knapsack were published: one by Kaufman [86] and the other by Roberts [126]. In fact, in 1965 Gimpelson [59] appealed to discrete-event simulation to study the monotonicity behavior of the stochastic knapsack — and he made no reference to the product-form result.

The Recursive Algorithm 2.1 was independently published by Kaufman [86] and Roberts [126]. Hui [72] is often credited for the time-scale decomposition for ATM, discussed in Section 2.3. The Markov process model for contiguous slot assignment (Section 2.4) is due to Ramaswami and Rao [117]. The approximation procedure for contiguous slot assignment is due to Reiman and Schmitt [121].

The monotonicity results described in Section 2.6 are due to Ross and Yao [140]. These results offer theoretical explanations for the bizarre non-monotonicity behavior observed by Gimpelson. See also Nain [112] who obtained, through different techniques, monotonicity results for the case $K = 2$. Applying likelihood ratio and other ordering techniques to multiple server queueing systems, Smith and Whitt [148] give an excellent introduction to stochastic comparisons. Shanthikumar and Yao [145] [146] use likelihood ratio ordering to obtain numerous insightful results for closed queueing networks.

The asymptotic analysis of Section 2.7 is adapted from Reiman [120]. Corollary 2.7 is due to Reiman, although we have modified his proof. As already mentioned, Reiman's paper provides additional computational studies which compare the exact blocking probabilities to their asymptotic approximations.

Theorem 2.9 of Section 2.8 was originally stated and proved by Robert [124]; it is similar to a result for queueing systems obtained by Kipnis and Robert [96]. The results that follow Theorem 2.9 are new. We would also like to bring to the reader's attention two recent reports on multiservice networks: COST 224 Final Report, Performance eval-

uation and design of multiservice networks [128]; COST 242 Interim Report, Multi-rate models for dimensioning and performance evaluation of ATM networks [123].

2.10 Summary of Notation

Notation for Basic Knapsack

$\mathcal{I}$	non-negative integers
C	knapsack capacity
$ER[\rho, C]$	blocking probability for Erlang loss system
K	number of classes
$\mathcal{K}$	set of all classes
b_k	size of class-k objects
λ_k	arrival rate for class k
$1/\mu_k$	mean holding time for class k
$\rho_k = \lambda_k/\mu_k$	offered load for class k
B_k	blocking probability for class k
TH_k	throughput for class k
n_k	number of class-k objects in knapsack
$\mathbf{b} = (b_1, \ldots, b_K)$	size vector
$\mathbf{n} = (n_1, \ldots, n_K)$	state vector
$\mathbf{e}_k$	K-dimensional vector of all 0s except for a 1 in the kth place
$\mathcal{S}$	state space
$\mathcal{S}_k$	admittance region for class-k objects
$\mathcal{S}(c)$	set of states with occupancy c
$\pi(\mathbf{n})$	equilibrium state probability
$q(c)$	equilibrium occupancy probability
G	normalization constant
$g(c) = q(c)G$	unnormalized equilibrium state probability
X_k	random variable denoting the number of class-k objects in the system
$\mathbf{X} = (X_1, \ldots, X_K)$	state vector
U	knapsack utilization
UTIL	average utilization
Y_k	unconstrained cousin of X_k

V	unconstrained cousin of U
$r_X(n)$	ratio function

Notation for ATM

$\mathbf{f}$	admission policy
$\mathcal{S}(\mathbf{f})$	set of recurrent states under policy $\mathbf{f}$
Λ	set of allowable VC profiles
m_k	number of class-k bursts in progress
b_k^e	effective bandwidth for service k VCs
$\mathbf{b}^e$	effective bandwidth vector

Notation for Contiguous Slot Assignment

$\mathbf{n} = (n_w, n_0, \ldots, n_6)$	state vector
$\mathbf{Q}$	infinitesimal generator
$\mathbf{Q}_{ll}$	infinitesimal generator for system with l wideband calls
λ_n, λ_w	arrival rates for narrowband and wideband calls
μ_n, μ_w	departure rates for narrowband and wideband calls
$\tilde{B}_n$, $\tilde{B}_w$	limiting blocking probabilities for narrowband and wideband calls

Notation for Asymptotic Analysis

ρ_k^*	asymptotic normalized offered load for class k
ρ^*	asymptotic normalized aggregate offered load
$\hat{X}_k(C)$	normalized random variable for the number of class-k objects in the knapsack
$\hat{\mathbf{X}}(C)$	normalized vector
$\hat{\mathbf{X}}$	asymptotic normalized vector
$B_k(C)$	blocking probability for class k
$\hat{B}_k$	asymptotic blocking
α, δ, σ^2, β_k	asymptotic parameters
Z	standard normal random variable

Notation for Continuous Weights

λ	arrival rate for objects
$1/\mu$	average holding time for object
$\rho = \lambda/\mu$	
C	knapsack capacity
$F(\cdot)$	distribution function for object sizes
l	number of objects in knapsack
$b_1, \ldots, b_l$	weights of l objects in the knapsack
$\xi = \{b_1, \ldots, b_l\}$	state of knapsack
$\mathcal{S}$	set of all states
h	function from $\mathcal{S}$ to reals (random variable)
U	random variable for knapsack utilization
U_l	random variable with distribution function $F(\cdot)$
L	Poisson random variable with parameter ρ
$\mathbf{Q}$	infinitesimal generator