PROVISIONING FOR BANDWIDTH SHARING AND EXCHANGE

Robert C. Hampshire
Princeton University, Department of Operations Research and Financial Engineering
Engineering Quadrangle, Princeton NJ 08544
rhampshi@princeton.edu

William A. Massey
Princeton University, Department of Operations Research and Financial Engineering
Engineering Quadrangle, Princeton NJ 08544
wmassey@princeton.edu

Debasis Mitra
Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974-0636
dmitra@lucent.com

Qiong Wang
Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974-0636
chiwang@lucent.com

Abstract Customers of bandwidth services can be divided into two distinct groups: those customers requesting bandwidth for the future and those desiring bandwidth immediately. We develop a dynamic network provisioning methodology that minimally satisfies the QoS (blocking probability) requirements for the 'on-demand' customers. Our method is sufficiently general and captures time varying trends in the demand for services as well as different bandwidth requests for the multiple classes of customers. This allows a network provider to be efficient in reserving excess bandwidth for forward contracts. Asymptotic results and bounds for the Erlang loss system are invoked to obtain simple approximate solutions to this bandwidth provisioning problem.

Keywords: Bandwidth exchanges, network economics, network provisioning, Erlang B formula, heavy traffic limits, loss systems.

Introduction

In this paper, we develop a bandwidth provisioning scheme for a service network that satisfies the "on demand" customers. This sets the stage for providing bandwidth to serve customers with long-term contracts. Consider two broad categories of demand:

1 Immediate Demands

2 Forward Demands

Immediate Demand (ID) is the traditional category where customers make requests for bandwidth and expect the resources immediately. One advantage to traditional service is that there are historical records and statistical techniques for forecasting demand, which is expected to be stable, and describing its statistical properties, such as distributional information on arrivals and holding periods. One disadvantage however, is that there are corresponding expectations on the part of customers for a high quality of service, i.e., low blocking rates.

Forward Demand (FD), on the other hand, is the service category that is expected to grow rapidly with the increased availability of bandwidth in the Internet's infrastructure and universal high-capacity access to the Internet. Consider the following examples of application services that will create FD. Schools that offer distance learning, such as MIT or U.C. Berkeley, want to have bandwidth available from the campus to each learning site commencing at 10 am every Monday and Thursday during the term. Large corporations want contracts for guaranteed bandwidth supply for carrying internal communication traffic. Other carriers lease capacity for an extended period of time to defer capital investment in infrastructure.

We model the ID requests as multi-class Poisson. Say there are n ID classes, with class i characterized by (λ_i, μ_i, b_i), where λ_i is the Poisson rate of arrivals, $1/\mu_i$ is the mean holding time of individual demands, and b_i is the bandwidth demand on individual requests. We leave open for the present the matter of the distributions of the holding periods. An example of bandwidths demanded by differing classes is $\{$64 kps, 128 kps, 256 kps, 384 kps$\}$.

FD requests are indexed by i, and the j-th request is characterized by the four-tuple (R_j, S_j, T_j, b_j), where R_j is the time that the request is made, S_j is the start time of the bandwidth demand, T_j is its termination time, and b_j is the bandwidth requested.

We do not propose any specific statistical model for FD, in part because it is in a nascent stage, data is unavailable and also, as with any new service, the demand rates are unstable and unpredictable. It is our expectation that the holding times $T_j - S_j$ with be typically longer than in ID, and that the requested bandwidths b_j will also be larger.

Indeed if the holding times $T_j - S_j$ last for several hours or days, then there are important consequences on the modelling of ID. It becomes necessary to incorporate time dependencies, particularly in the arrival rates λ_i. We propose to consider time inhomogeneous Poisson processes, i.e., $\lambda_i \equiv \lambda_i(t)$ for $i = 1, \ldots, n$.

This paper focuses on a strategy to satisfy the ID customers. A provisioning methodology is developed to allocate the least amount of bandwidth needed to accommodate the QoS requirements of the ID customers, so that more capacity can be made available to serve the forward demand.

This provisioning scheme is developed first for a single customer class. Each member of this class requests a unit amount of resources and has identical demand characteristics that only depend on the current price. An asymptotic provisioning solution is obtained for the steady-state single class case. Next, the demand function for this single class case is allowed to depend on time. In this time-varying single class case an approximation technique is employed to develop a provisioning solution. The results for the single class steady-state and time-varying cases are then generalized to a multiple class case. This generalization allows for multiple customer classes each requesting distinct amounts of bandwidth and each having unique demand characteristics. Armed with the single class results and techniques of reversible systems, a multi-class provisioning solution is realized.

1. Canonical Design Problems for the Erlang Loss Model

Let us first investigate the single customer class case ($n = 1$). It is assumed that all the customers in this class request a unit amount of bandwidth ($b_i = 1$) and are governed by the same demand function that only depends on the price. Let customers arrive according to a Poisson process, where λ equals the mean *arrival rate*. Moreover, let the *holding time* for the unit bandwidth resource be random and assume that different customers have i.i.d. holding times, where $1/\mu$ equals the mean holding time. The unit amount of bandwidth requested by a customer is called a *channel* and we define L to equals the total number of channels. The resulting queueing model for this single class case is the classical Erlang loss model. Assuming a homogeneous Poisson arrival rate, it is typically denoted as an $M/G/L/L$ queue. When all channels are in use, the system is called *blocked* and we define ϵ to equal the probability that the system is blocked.

If there is an infinite amount of bandwidth available, then every customer requesting a channel receives it. The total number of channels *requested* by customers at a given time is called the *offered load* and we define q to equal its mean. It is a function of the aggregate demand for bandwidth. The $M/G/\infty$ (infinite server queue) is viewed as the offered load process for bandwidth

requests. The steady state distribution for the $M/G/\infty$ queue length Q_∞ is Poisson where

$$\mathsf{Pr}(Q_\infty = i) = \frac{e^{-q}q^i}{i!} \tag{1}$$

for all $i = 0, 1, \ldots$ and $q = \lambda/\mu$. Since $\mathsf{E}[Q_\infty] = \mathsf{Var}[Q_\infty] = q$, it follows that q equals the mean of Q_∞ and $\sqrt{q}$ equals the standard deviation of Q_∞.

In the context of this single class, unit bandwidth, classical Erlang loss model, we can discuss three canonical design problems:

1 The Quality of Service (QoS) Problem.

2 The Provisioning Problem.

3 The Pricing Problem.

In the next section, we generalize these basic problems to the case of a multi-class bandwidth model.

The first of three problems is the *quality of service (QoS) problem*. It can be described graphically by the following block diagram. Formally the problem

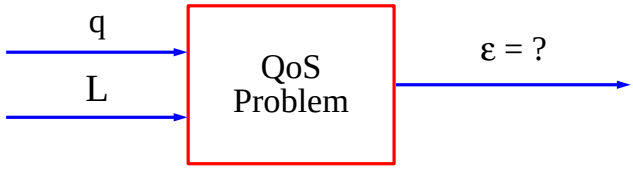

Figure 1. The quality of service (QoS) problem.

statement is as follows: Given the number of channels L and the mean of the offered load q, what is the resulting probability of blocking ϵ experienced by the single customer class?

An exact solution to the QoS problem was obtained by Erlang [2]. The solution is the classical *Erlang blocking formula*. It states that if L is the total number of channels available and q is the mean of the offered load then the blocking probability equals:

$$\beta_L(q) = \frac{q^L}{L!} \Big/ \sum_{i=0}^{L} \frac{q^i}{i!}. \tag{2}$$

We can rewrite this formula as a conditional probability of the offered load process and obtain:

$$\beta_L(q) = P(\, Q_\infty = L \,|\, Q_\infty \leq L \,) = P(\, L-1 < Q_\infty \leq L \,|\, Q_\infty \leq L \,). \tag{3}$$

What is *probabilistically* clear (using the theory of time reversible Markov chains, see Kelly [7]) but *physically* paradoxical is that the infinite server queue which experiences no congestion gives complete insight into the analysis of systems with blocking. Also this conditional form is quite useful in the heavy traffic analysis needed for the provisioning problem.

Now we relax the constraints on the arrival process and let customers arrive according to a non-homogeneous Poisson process where at time t, $\lambda(t)$ equals the mean rate of the non-homogenous Poisson process. The offered load process $\{ Q_\infty(t) \mid t \geq 0 \}$ for this time varying case is the $M_t/G/\infty$ queue. At time t, the $M_t/G/\infty$ queue has a Poisson distribution or

$$P(Q_\infty(t) = i) = \frac{e^{-q(t)} q(t)^i}{i!}, \tag{4}$$

whenever $Q_\infty(0)$ has a Poisson distribution, which includes $Q_\infty(0) = 0$. Moreover, assuming that the holding times are exponential, the mean of the time varying offered load process is then:

$$\frac{d}{dt} q(t) = \lambda(t) - \mu \cdot q(t). \tag{5}$$

To model more general service distributions, we can numerically solve a similar set of ordinary differential equations for a phase type service. The total number of equations used for such distributions equals the number of service phases.

Now that the distribution of the time varying offered load process is known, how does one find a solution to the QoS problem? The modified offered load (MOL) approximation is employed to give an approximate solution to the time-varying QoS problem. Given L channels, if $Q_L(t)$ equals the number of channels in use at time t, then

$$\mathsf{Pr}(Q_L(t) = L) \approx \beta_L(q(t)) = P(Q_\infty(t) = L | Q_\infty(t) \leq L). \tag{6}$$

where $q(t)$ solves the above differential equation. This result can be found in Jagerman [5]. Error bounds for this approximation are given by Massey and Whitt [11]. The MOL approximation is at its best during periods of small blocking probabilities, which in practice is when such approximations are most useful.

The second canonical problem is the *provisioning problem*, which is the main thrust of this paper. Formally the problem statement is as follows: Given a mean offered load q, what is the smallest number L of channels needed to guarantee a QoS probability of blocking less than ϵ?

We use the work on server staffing in Jennings, Mandelbaum, Massey and Whitt [3] as motivation to develop a provisioning solution. If L is the amount of provisioned bandwidth that satisfies the single class QoS constraint, then L

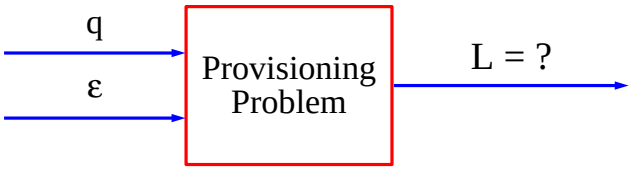

Figure 2. The provisioning problem.

should at least be as big as the mean of the offered load. It is also reasonable to add extra capacity to handle random demand fluctuations bigger than the mean. In this spirit we set the number of channels equal to the mean plus some multiple x of the standard deviation of the offered load or

$$L(q,x) = \lceil\, q + x\sqrt{q}\,\rceil, \tag{7}$$

where x is selected in [3] by computing the inverse of a Gaussian tail distribution. The inverse of the Gaussian tail distribution is useful for approximating solutions to provisioning problems for delay systems but not for loss systems. The more appropriate function to use in this paper is suggested by the work of Jagerman [6].

Recall that the probability of blocking ϵ equals the following conditional probability:

$$\beta_L(q) = \frac{P(Q_\infty = L)}{P(Q_\infty \leq L)} \tag{8}$$

where Q_∞ has a Poisson distribution. If we scale up the mean of the offered load, then we have the asymptotic result

$$\lim_{q\to\infty} \sqrt{q} \cdot \beta_{L(q,x)}(q) = \frac{\phi(x)}{\Phi(x)} = \text{“}P(N(0,1) = x \,|\, N(0,1) \leq x)\text{"} \tag{9}$$

where $N(0,1)$ has a normal distribution or formally

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2} \quad \text{and} \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{10}$$

This result can be found in Jagerman [6].

Now we define an important special function. Let ψ be the inverse function to ϕ/Φ, where for all $x > 0$

$$\frac{\phi(\psi(x))}{\Phi(\psi(x))} = x. \tag{11}$$

The properties of the ψ function are of utmost importance to our analysis of the provisioning problem. We now explore several of the key properties for ψ.

Theorem 1 *If ψ is the inverse of ϕ/Φ, then it is strictly decreasing with*

$$\psi(y) + y > 0 \tag{12}$$

for all $y > 0$. Moreover, ψ is the unique solution to the nonlinear differential equation

$$\psi'(y) = \frac{-1}{(\psi(y) + y)y}, \tag{13}$$

with the initial condition $\psi(\sqrt{2/\pi}\,) = 0$.

Proof: We first show that ψ solves the above differential equation. Starting with the identity

$$\frac{\phi(x)}{\Phi(x)} = \frac{e^{-x^2/2}}{\int_{-\infty}^{x} e^{-t^2/2}\,dt} = \frac{1}{\int_0^{\infty} e^{-t^2/2+xt}\,dt}, \tag{14}$$

we obtain

$$\int_0^{\infty} e^{-t^2/2+\psi(y)t}\,dt = \frac{1}{y}. \tag{15}$$

Now we differentiate both sides by y and get

$$\psi'(y) \cdot \int_0^{\infty} te^{-t^2/2+\psi(y)t}\,dt = \frac{-1}{y^2}, \tag{16}$$

which gives us

$$\begin{aligned}
\frac{-1}{y^2} &= -\psi'(y) \cdot \int_0^{\infty} e^{\psi(y)t} \cdot \frac{d}{dt} e^{-t^2/2}\,dt \\
&= \psi'(y)\left(1 + \psi(y) \cdot \int_0^{\infty} e^{-t^2/2+\psi(y)t}\,dt\right) \\
&= \psi'(y)\left(1 + \frac{\psi(y)}{y}\right).
\end{aligned}$$

and the differential equation for ψ follows from this identity.

Using the above identity (15) and integration by parts, we have

$$y + \psi(y) = \frac{1}{\int_0^{\infty} e^{-t^2/2+\psi(y)}dt} + \psi(y) \tag{17}$$

$$= \frac{1 + \psi(y)\int_0^{\infty} e^{-t^2/2+\psi(y)}dt}{\int_0^{\infty} e^{-t^2/2+\psi(y)}dt} \tag{18}$$

$$= \frac{\int_0^{\infty} te^{-t^2/2+\psi(y)}dt}{\int_0^{\infty} e^{-t^2/2+\psi(y)}dt} \tag{19}$$

which shows that $y + \psi(y) > 0$ and completes the proof. ∎

The ψ function is the inverse of the hazard function. Because the ψ function solves a simple ordinary differential equation, we can easily compute it numerically. Moreover, ψ is a generic function so we can precompute a lookup table of values for $\psi(x)$ that can be used for all provisioning problems. We use a second order Runge-Kutta method to compute $\psi(x)$, based on the following approximation:

$$\psi(x+\Delta x) \approx \psi(x) - \frac{\Delta x}{(x+\Delta x/2)\left(x+\Delta x/2+\psi(x)-\Delta x/\Big(2x(x+\psi(x))\Big)\right)} \tag{20}$$

Given the ψ function, we can construct an *asymptotic channel provisioning solution.* If $\epsilon = \beta_L(q)$ and we set $L = \lceil\, q + x\sqrt{q}\,\rceil$, then

$$\epsilon \approx \frac{1}{\sqrt{q}} \cdot \frac{\phi(x)}{\Phi(x)} \text{ implies } x \approx \psi(\epsilon\sqrt{q}). \tag{21}$$

Making this approximation an equality gives us

$$L = \lceil\, q + \psi(\epsilon\sqrt{q})\sqrt{q}\,\rceil. \tag{22}$$

If we define $\ell(z) \equiv z + \psi(\,\epsilon\sqrt{z}\,)\sqrt{z}$. We can show from the properties for ψ that

$$\ell(0) = 0 \quad \text{and} \quad \ell(L/(1-\epsilon)) \geq L. \tag{23}$$

By the continuity of ℓ, there must exist some $0 < q \leq L/(1-\epsilon)$ where $\ell(q) = L$. Given the properties of ψ, we have

$$L = q + \psi(\,\epsilon\sqrt{q}\,)\sqrt{q} > q(1-\epsilon). \tag{24}$$

Define the carried load to be the mean number of customers that are admitted for service. If L is the actual number of channels that gives a steady state offered load of q and a QoS of ϵ, then the carried load is $q(1-\epsilon)$. This is consistent with the above inequality.

We now turn our focus to the time varying single class provisioning problem. An approximate provisioning solution can be realized via the modified offered load approximation combined with the ψ function. The solution takes the same form as above. The number of provisioned channels equals the mean of the offered load plus some multiple of the standard deviation of the offered load. The approximate time-varying provisioning solution is:

$$L(t) \approx q(t) + \psi\left(\epsilon\sqrt{q(t)}\right)\sqrt{q(t)} \tag{25}$$

where for the case of exponentially distributed service times, q solves the differential equation

$$\frac{dq}{dt}(t) = \lambda(t) - \mu \cdot q(t). \tag{26}$$

The provisioned number of channels, $L(t)$, is a continuous function of time due to the continuity of ψ and $q(t)$. Since $L(t)$ is set according to offered load $q(t)$, which is an expected value, it is possible that the actual number of users in the system exceeds the desired number of channels as specified by equation 25. This property is a unique by-product of the dynamic provisioning of network capacity. We define this scenario as a *ghost state*, and apply the following *non-preemptive service* discipline when the system reaches a ghost state:

- The excess channels process their last customers until their jobs are complete.
- During this period no new jobs are admitted.

Figure 3 is the state transition diagram for the single class customer case. It defines three distinct type of states: nonblocking states, blocking state and ghost states. If the system is in a nonblocking state then a transition to and from that state due to an arrival or service is allowed. While in a blocking state, any transition from this state due to an arrival is not permitted. In the ghost states a transition due to an arrival into a ghost state is forbidden. Only a transition due to a service from a ghost state is allowed.

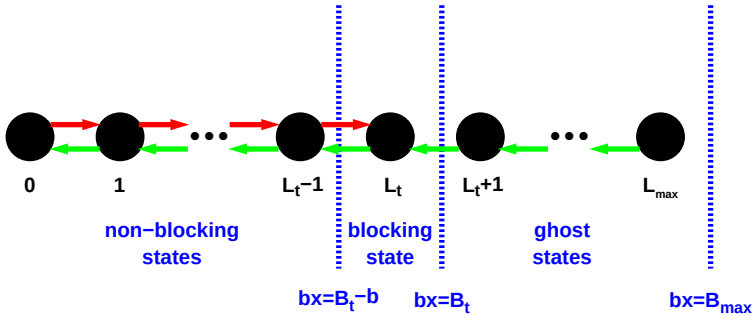

Figure 3. State transition diagram for the single class case.

Before concluding this section, should point out that there is a third design problem, called the *pricing problem*. Viewing price as a mechanism to control the offered load, this reduces to finding an offered load q that yields a QoS blocking probability ϵ given a total of L channels. This problem was addressed by Keon and Anandalingam [8] and for the case of a constant arrival rate, Courcoubetis and Reiman [1]. Also, a "Gaussian-distribution approximation based" approach is proposed by Lanning, Massey, Rider and Wang [9] for single-service models, and a "hazard function approximation" based approach is introduced for multi-service models in Hampshire, Massey and Wang [4].

2. Generalization to the Multi-Class Bandwidth Model

The single class results can be generalized to a multiple customer class setting. Suppose that we have a heterogeneous set of customers, where each class requests differing amounts of bandwidth. Let $\lambda_1, \ldots, \lambda_n$, $1/\mu_1, \ldots, 1/\mu_n$, and $b_1, \ldots, b_n$ be respectively, the call arrival rate functions, mean call holding times, and the amount of bandwidth requested for the n different classes of customers indexed by i. If there is an unlimited amount of available bandwidth, then all the classes behave like a collection of n-independent infinite server queues. We can then define an offered load model, where $Q_\infty^{(i)}(t)$ denotes the random number of customers simultaneously using b_i units of bandwidth. It follows that each $\left\{ Q_\infty^{(i)}(t) \mid t \geq 0 \right\}$ is an $M/G/\infty$ queueing process where each $Q_\infty^{(i)}(t)$ has a Poisson distribution whenever $Q_\infty^{(i)}(0)$ does. If we let R equal the offered load of the total requested bandwidth, then

$$R = \sum_{i=1}^{n} b_i Q_\infty^{(i)} \tag{27}$$

where in steady state $\mathsf{E}[Q_\infty^{(i)}] = \mathsf{Var}[Q_\infty^{(i)}] = q_i = \lambda_i/\mu_i$. Consequently,

$$\mathsf{E}[R] = \sum_{i=1}^{n} b_i q_i \;\text{ and }\; \mathsf{Var}[R] = \sum_{i=1}^{n} b_i^2 q_i. \tag{28}$$

Let B be the total amount of available bandwidth. We can then formulate a carried load model where $Q_B^{(i)}(t)$ equals the random number of customers simultaneously using b_i units of bandwidth at time t, given an admission control policy that rejects any arriving customer requesting more bandwidth than is available.

We now reconsider the QoS problem for multiple customer classes. The

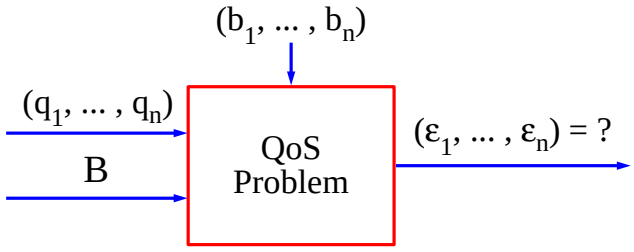

Figure 4. The multi-class bandwidth quality of service (QoS) problem.

blocking for class j customers equals the probability of the event that $\sum_{i=1}^{n} b_i Q_B^{(i)}$ is greater than $B - b_j$.

Since the $Q_\infty^{(i)}(t)$'s are mutually independent Poisson random variables, we know that the probability given above is some generic function $\beta_B^{(i)} : \Re^n \rightarrow \Re$ of the $q_i(t)$'s where $q_i(t) = E[Q_\infty^{(i)}(t)]$. Let $\mathbf{q} = (q_1, ..., q_n)$ and $\mathbf{b} = (b_1, ..., b_n)$. In general, if $Q_1, \ldots, Q_n$ are a collection of mutually independent Poisson random variables with $q_i \equiv E[Q_i]$, if we define $\beta_B^{(i)}$ to be

$$\begin{aligned}
\beta_B^{(i)}(\mathbf{q}, \mathbf{b}) &= \mathsf{Pr}\left(B - b_i < \sum_{j=1}^{n} b_j Q_B^{(j)} \right) \\
&= \mathsf{Pr}\left(B - b_i < \sum_{j=1}^{n} b_j Q_\infty^{(j)} \leq B \,\middle|\, \sum_{j=1}^{n} b_j Q_\infty^{(j)} \leq B \right) \\
&= \frac{\mathsf{Pr}\left(B - b_i < \sum_{j=1}^{n} b_j Q_\infty^{(j)} \leq B \right)}{\mathsf{Pr}\left(\sum_{j=1}^{n} b_j Q_\infty^{(j)} \leq B \right)},
\end{aligned}$$

and $\mathbf{q} = (q_1, \ldots, q_n)$. Then this equals the steady state blocking probability for class i. This result follows from time reversibility as discussed in Kelly [7].

We now reconsider the capacity provisioning problem with time-varying arrival rates for multiple services. In this case, the blocking at time t for class j customers equals the probability of the event that $\sum_{i=1}^{n} b_i Q_B^{(i)}(t)$ is greater than $B(t) - b_j$. The modified offered load approximation for this probability is defined to be

$$\mathsf{Pr}\left(B - b_i < \sum_{j=1}^{n} b_j Q_B^{(j)}(t) \right) \approx \mathsf{Pr}\left(B - b_i < \sum_{j=1}^{n} b_j Q_\infty^{(j)}(t) \,\middle|\, \sum_{j=1}^{n} b_j Q_\infty^{(j)}(t) \leq B \right). \tag{29}$$

One justification for this approximation is that it gives the exact answer when the arrival rates are constant and the system is in steady state. Thus an approximate QoS solution is :

$$\begin{aligned}
\beta_B^{(i)}(\mathbf{q}(\mathbf{t}), \mathbf{b}) &= \mathsf{Pr}\left(B - b_i < \sum_{j=1}^{n} b_j Q_\infty^{(j)}(t) \leq B \,\middle|\, \sum_{j=1}^{n} b_j Q_\infty^{(j)}(t) \leq B \right) \\
&= \frac{\mathsf{Pr}\left(B - b_i < \sum_{j=1}^{n} b_j Q_\infty^{(j)}(t) \leq B \right)}{\mathsf{Pr}\left(\sum_{j=1}^{n} b_j Q_\infty^{(j)}(t) \leq B \right)}.
\end{aligned}$$

We now reconsider the provisioning problem for multiple customer classes. If q_i is the mean offered load for customers requesting b_i units of bandwidth, then the multiple class provisioning problem is to answer the question: What is the smallest amount B of bandwidth needed to guarantee a probability of blocking less than ϵ_i for each class i?

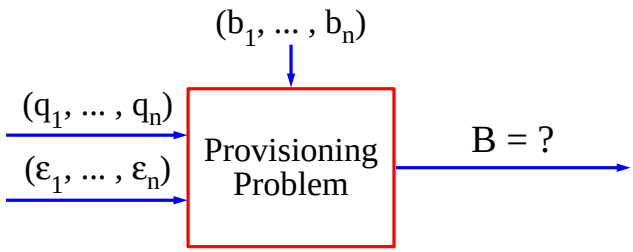

Figure 5. The multi-class bandwidth provisioning problem.

Recall that R is the offered load of the total requested bandwidth. If B is the amount of provisioned bandwidth that satisfies the multi-class QoS constraints, then B should be at least as big as the mean of the offered load R. It is also reasonable to add extra capacity to handle random demand fluctuations bigger than the mean. In this spirit we set the amount of bandwidth equal to the mean plus some multiple x of the standard deviation of the offered load. As in the single class case, we scale up the offered load of each class. In this limiting regime an asymptotic provisioning solution is found. If

$$B(\eta, x) \equiv \eta \cdot \sum_{i=1}^{n} b_i q_i + x \sqrt{\eta \cdot \sum_{i=1}^{n} b_i^2 q_i} \tag{30}$$

where η is a scaling factor for the offered loads, then we have the limiting result:

$$\lim_{\eta \to \infty} \sqrt{\eta} \beta^{(i)}_{B(\eta,x)}(\mathbf{q}, \mathbf{b}) = \frac{b_i}{\sqrt{\sum_{i=1}^{n} b_i^2 q_i}} \cdot \frac{\phi(x)}{\Phi(x)}, \tag{31}$$

where ϕ and Φ are defined the same as for the single class case. This limiting result can be found in the papers of Reiman [13] as well as Mitra and Morrison [12]. Since ψ is a decreasing function, then the constraint $\beta^{(i)}_B(\mathbf{q}, \mathbf{b}) \leq \epsilon_i$ asymptotically (using the value of $\sqrt{\eta}\beta^{(i)}_{B(\eta,x)}(\mathbf{q}, \mathbf{b})$ as $\eta \to \infty$ to approximate its value at $\eta = 1$) implies

$$\frac{b_i}{\sqrt{\sum_{i=1}^{n} b_i^2 q_i}} \cdot \frac{\phi(x)}{\Phi(x)} \leq \epsilon_i \quad \Rightarrow \quad x \geq \psi\left(\frac{\epsilon_i}{b_i} \sqrt{\sum_{i=1}^{n} b_i^2 q_i} \right). \tag{32}$$

Our provisioned amount of bandwidth must satisfy the QoS conditions for all of the classes. Thus if x satisfies all the QoS conditions, then

$$x \geq \max_{1 \leq i \leq n} \psi\left(\frac{\epsilon_i}{b_i} \cdot \sqrt{\sum_{i=1}^{n} b_i^2 q_i} \right) \tag{33}$$

which is equivalent to

$$x \geq \psi \left(\min_{1 \leq i \leq n} \frac{\epsilon_i}{b_i} \cdot \sqrt{\sum_{i=1}^{n} b_i^2 q_i} \right). \tag{34}$$

Making this inequality an equality, we have the provisioning solution:

$$B = \sum_{i=1}^{n} b_i q_i + \psi \left(\min_{1 \leq i \leq n} \frac{\epsilon_i}{b_i} \cdot \sqrt{\sum_{i=1}^{n} b_i^2 q_i} \right) \sqrt{\sum_{i=1}^{n} b_i^2 q_i}. \tag{35}$$

This result leads to an asymptotic rule of thumb which states:

> **Asymptotic Rule of Thumb:** The dominant QoS classes are the ones with the smallest ϵ_i / b_i ratio.

Satisfying their requirements provides more than enough bandwidth for all the other classes.

These results can be generalized to the time varying arrival case. The approximate time-varying provisioning solution at time t is

$$B(t) = \sum_{i=1}^{n} b_i q_i(t) + \psi \left(\min_{1 \leq i \leq n} \frac{\epsilon_i}{b_i} \cdot \sqrt{\sum_{i=1}^{n} b_i^2 q_i(t)} \right) \sqrt{\sum_{i=1}^{n} b_i^2 q_i(t)} \tag{36}$$

where if we assume that the service time for each class is exponentially distributed, then each $q_i(t)$ solves the differential equation

$$\frac{d}{dt} q_i(t) = \lambda_i(t) - \mu_i \cdot q_i(t). \tag{37}$$

These results are due to the modified offered load approximation. The bandwidth function $B(t)$ is a continuous function of time. Service discipline assumptions need to be made as in the single class case. During times of capacity reduction customers hold their resources until their job is complete. Also during this period no new customers of that class are admitted for service. Figure 6 is the state space transition diagram for a system with two classes of customers. It is assumed that class 2 customers request more bandwidth, b_i, than the first class. This figure defines four distinct type of states: nonblocking states, class 2 blocking states, class 1 and 2 blocking states and ghost states. If the system is in a nonblocking state then a transition to and from a state due to an arrival or a service is allowed for both classes. While in a class 2 blocking state, transitions from these states due to an arrival of a class 2 customer is not permitted. In the class 1 and 2 blocking states transitions from these states due to an arrival of a class 1 or class 2 customer is not permitted. In the ghost states a transition due to an arrival of either class into the ghost state is forbidden. Only a transition due to a service is allowed in ghost states.

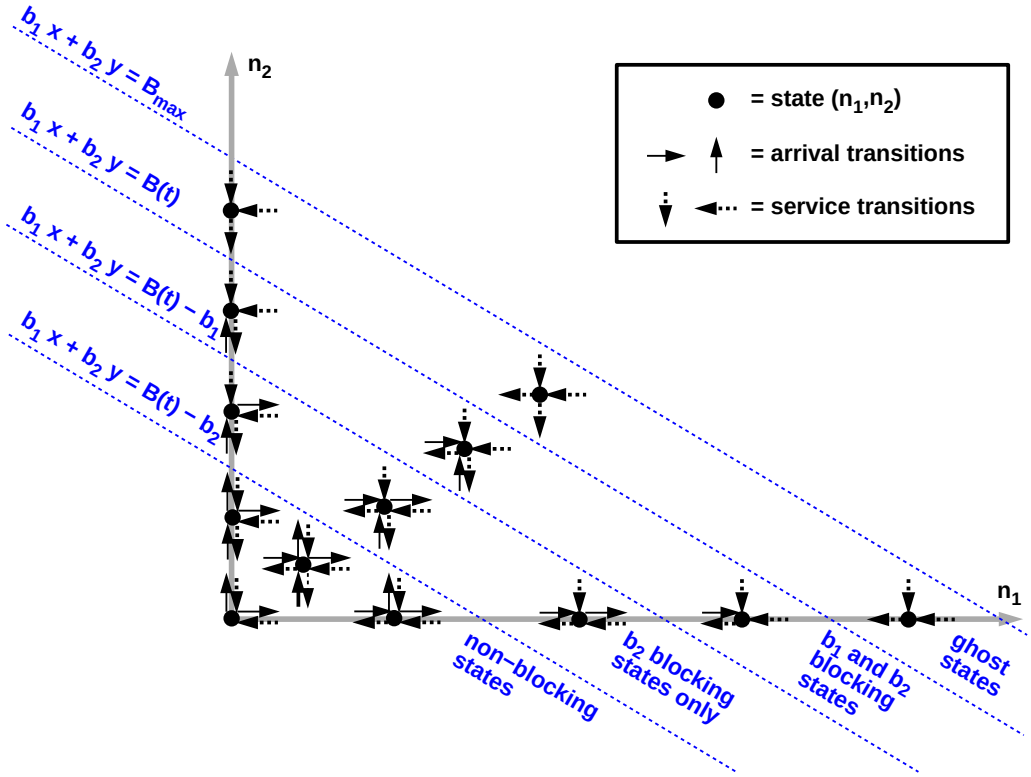

Figure 6. State transition diagram for the two-class case.

Before we conclude this section, we state for completeness the general multiple class bandwidth version of the pricing problem. Given the desired QoS probability of blocking ϵ_i for each class i requesting b_i units of bandwidth and given the amount of bandwidth B, what is the largest offered load q_i that yields a QoS blocking probability less than ϵ_i? An approximate algorithm for solving this problem is explored in the paper Hampshire, Massey and Wang [4].

3. Numerical Results

Numerical results are given for the provisioning problem with two customer classes. These two classes may have time varying arrival functions. The provisioning problem is solved to determine the amount of bandwidth $B(t)$ needed at any given time. Next we use this prescribed bandwidth at time t to formulate the "exact" Markovian loss model. Then at each time step numerically integrate the forward equations for this model and compute the transient blocking

probabilities. Once the blocking probabilities are computed we compare them to their respective QoS bounds.

The numerical example consists of two heterogenous customer classes. Let customers of the first class arrival according to a Poisson process with mean rate $\lambda_1(t) = 30$, requesting 20 units of bandwidth and desiring no more then 4 percent blocking . Customers of second class arrive according to a nonhomogeneous Poisson process with mean rate $\lambda_2(t) = 40 + 10\sin(2\pi t/80)$, requesting 5 units of bandwidth and desiring no more than 1 percent blocking.

For the numerical results presented, the planning horizon is 80 time units. It is assumed that the customer holding times are mutually independent and exponentially distributed with the mean of a single time unit.

The bandwidth function, $B(t)$, is a continuous function of time. In practice, a service provider changes the size of the network only at discrete times. The intervals on which the size of the network is held constant are called provisioning periods. The amount of bandwidth allocated over a provisioning period is the maximum of $B(t)$ over that provisioning interval. The provisioning periods can be made to be finer and finer. Thus as the provisioning period becomes infinitesimally small, the continuous provisioning solution is obtained.

The two period provisioning scenario is considered first. The top graph in Figure 7, is a plot of the transient blocking probabilities computed by numerically integrating the forward equations for the Markovian loss model with ghost states. The lower graph is a plot of the provisioning solution $B(t)$ which we use to compute the discrete approximation of $B(t)$ for exactly two provisioning periods. Notice at time 40 the apparent discontinuity in the blocking probabilities is reality a discontinuity of the *derivative* of the blocking probabilities, which are actually continuous functions of time. This phenomena is due to the generation of ghost states. At time 40, the amount of provisioned resources decreases instantaneously. This activates the non-preemptive service assumptions, thus blocking arrivals of new requests. Now compare the transient blocking probabilities to the QoS targets. It is seen that the transient blocking probabilities are in reasonable range of the targets. As the number of provisioning periods is increased, the transient blocking probabilities are closer to the QoS targets. In Figure 8, we consider the case of eight provisioning periods. The derivative discontinuities in the blocking probabilities are caused by the generation of ghost states. The reasoning follows from above. Finally, turning to the continuously provisioned system, the transient blocking probabilities approach the desired QoS requirement for each class.

4. Summary

We have presented three canonical problems that arise from the Erlang loss model. These problems have a natural interpretation for a network service

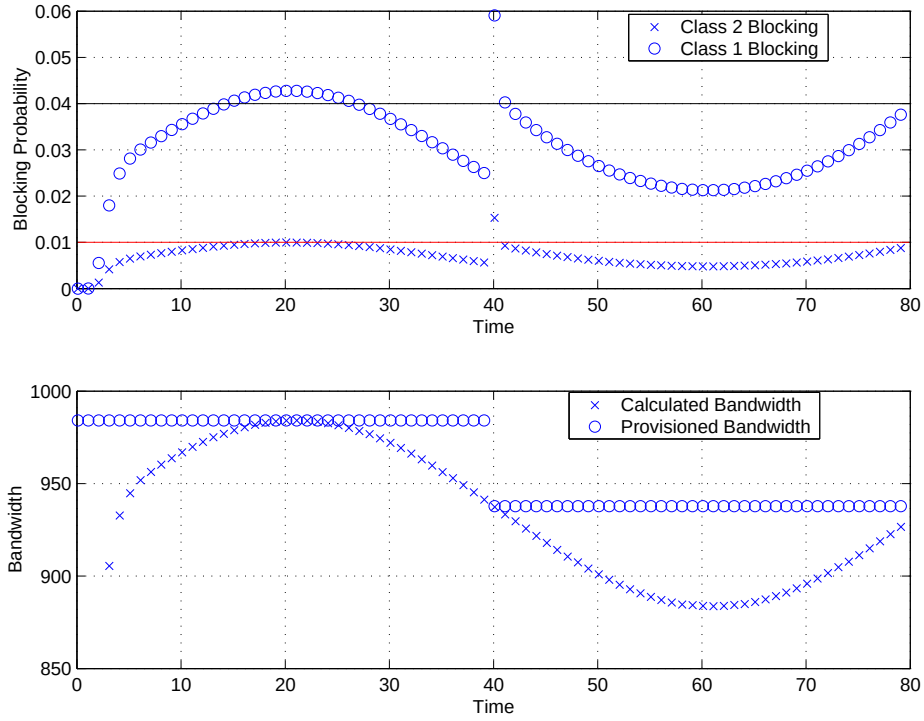

Figure 7. Two period provisioning example.

provider. The QoS problem is a classical problem that Erlang addressed in 1917. The pricing problem is the topic of another paper [4]. Much of this paper was dedicated to solving the provisioning problem. An asymptotic provisioning solution for a system offering multiple services was presented. A numerical example was also given in which there were two types of services and non-stationary demand for the services. It was observed that this provisioning methodology performs as desired. The provisioning solution is a result of an asymptotic scaling of the offered load. Therefore, we expect more desirable results as the demand for services increases. In the numerical example we assumed that the service time distributions were exponential. We should note that our provisioning solution is also valid for phase-type service distributions, where the mean offered load satisfies a system of n differential equations where n is the number of phases.

The provisioning solution is a planning tool for a network service provider that is offering multiple differentiated services that each have unique QoS guarantees. The bandwidth function, $B(t)$, can be used as schedule for capacity management. Our methodology for computing the provisioned bandwidth

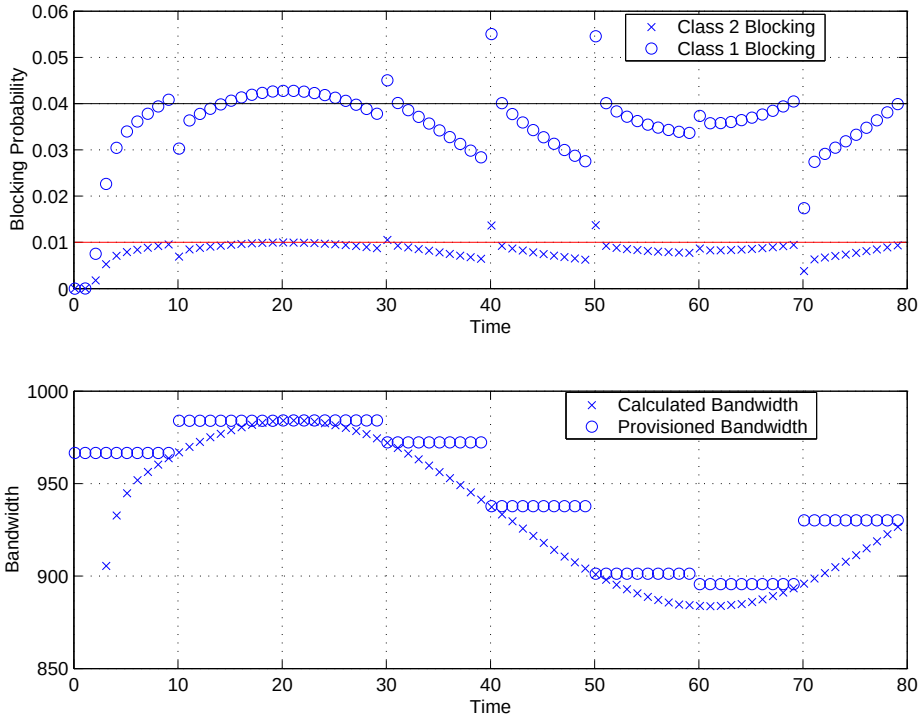

Figure 8. Eight period provisioning example.

schedule is lightweight and computationally inexpensive. This is because the function ψ can be simply computed from a lookup table. Therefore we can compute the provisioning schedule in realtime given forecasted demand for the services. The ability to compute the provisioning solution in realtime is a valuable property of our methodology.

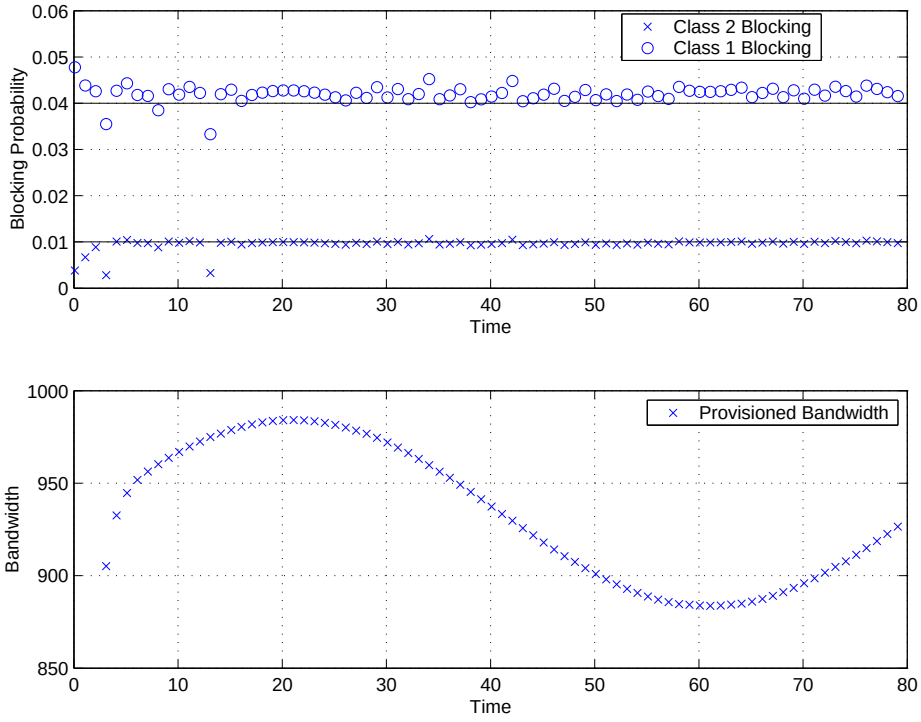

Figure 9. Continuous provisioning example.

References

[1] Courcoubetis, C.A. and Reiman, M.I."Pricing in a Large Single Link Loss System," *Teletraffic Engineering in a Competitive World, P. Key and D. Smith (editors)*, Elsevier, pp. 737-746, 1999.

[2] Erlang, A. K. "Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges," *The Post Office Electrical Engineers' Journal; (from the 1917 article in Danish in Elektroteknikeren vol. 13)*, pp. 189–197, 1918.

[3] Jennings, O. B., Mandelbaum A., Massey, W. A., and Whitt, W. "Server Staffing to Meet Time-Varying Demand," *Management Science*, pp. 1383–1394, 1996.

[4] Hampshire, R. C., Massey, W. A., and Wang, Q. "Dynamic Pricing for On-Demand Bandwidth Services," *Bell Laboratories Technical Report, 2002*.

[5] Jagerman, D. L. "Nonstationary Blocking in Telephone Traffic," *Bell System Technical Journal*, pp. 625–661, 1975.

[6] Jagerman, D. L. "Some Properties of the Erlang Loss Function," *The Bell System Technical Journal*, pp. 525–551, 1974.

[7] Kelly, F. P. "Reversibility and Stochastic Networks," John Wiley & Sons Ltd., 1979.

[8] Keon, N. and Anandalingam, G. "Real Time Pricing of Multiple Telecommunication Services Under Uncertain Demand," *Proceedings of the 7th International Conference on Telecommunication Systems, Modelling and Analysis*, pp. 28–47, 1999.

[9] Lanning, S., Massey, W. A., Rider, B. and Wang, Q. "Optimal Pricing in Queueing Systems with Quality of Service Constraints," *Proceedings of the 16th International Teletraffic Congress - ITC 16*, pp. 747–756, 1999.

[10] Massey, W. A. and Wallace, R. B. "An Asymptotic Optimal Design of the M/M/C/K Queue for Call Centers," submitted in 2002 to the *Selected Proceedings of the Council for African American Researchers in the Mathematical Sciences.*

[11] Massey, W. A. and Whitt, W. "An Analysis of the Modified Offered Load Approximation for the Nonstationary Erlang Loss Model," *Annals of Applied Probability*, pp. 1145–1160, 1994.

[12] Mitra, D. and Morrison, J. "Erlang Capacity and Uniform Approximations for Shared Unbuffered Resources," *IEEE/ACM Trans. Networking*, pp. 558-570, 1994.

[13] Reiman, M. I. "A Critically Loaded Multiclass Erlang Loss System," *Queueing Systems*, pp. 65–82, 1991.