2003

Capacity expansion for a loss system with exponential demand growth

Alexander Simampo
Sabritec, Inc.

Sarah M. Ryan
Iowa State University, smryan@iastate.edu

Capacity Expansion for a Loss System with Exponential Demand Growth

Alexander Simampo

Sarah M. Ryan*

Department of Industrial and Manufacturing Systems Engineering
Iowa State University
2019 Black Engineering
Ames, IA 50011-2164, USA

December, 2001

*Corresponding Author: email smryan@iastate.edu
Voice: 515-294-4347
Fax: 515-294-3524

This is a manuscript of an article that is published as, Simampo, A. and S. M. Ryan, "Capacity Expansion for a Loss System with Exponential Demand Growth," Computers and Operations Research, 30, 1525-1537 (2003). Available at: https://doi.org/10.1016/S0305-0548(02)00081-3. Posted with permission.

Scope and Purpose

Internet use has grown tremendously over the past few years. Businesses, educational institutions, government organizations as well as individuals have become heavily dependent on the Internet. One of the significant aspects of the Internet is access. This paper considers a potential "busy signal" problem in Internet access on a dial-up system due to insufficient capacity at the Internet Service Provider (ISP) company. We showed how to exploit a linearity property of the Erlang loss formula to determine the appropriate capacity in order to meet a specified level of customer service. We also applied a model to determine the optimal time points for the ISP company to expand its capacity in order to maintain a specified service level when confronted with exponential growth of subscribers.

Abstract

We study a loss system to forecast the demand for capacity based on the forecast demand for service and a specified service level. A little-used property of the Erlang loss formula allows the linear transformation of demand for service into demand for capacity. Next, given the forecast demand for capacity, we approximate a long run optimal capacity expansion policy by optimizing over successively longer finite time horizons. Analytical formulas together with regression analysis show the significance of the number of potential customers, frequency and duration of their requests for service, and the specified service level on the demand for capacity. Numerical sensitivity analysis exposes the effects of cost parameters, the demand growth rate and the required rate of return on the optimal time intervals between expansions.

Keywords: Capacity Expansion, Loss System, Erlang Loss Formula.

I. Introduction

In a competitive market, each firm's survival depends on its ability to attract and retain customers. Competition may initially appear as high price elasticity of demand. However, once prices have been driven as low as possible, the basis for competition shifts to quality of service. This service quality elasticity is particularly significant in industries in which each customer subscribes to one of many competing providers of a generic type of service. If the cost to switch to a competing provider is low, subscribers who repeatedly find the service unavailable will soon be lost. Therefore, each firm must ensure its own subscribers' satisfaction.

One example of low cost generic service is Internet access from a dial-up Internet Service Provider (ISP). Currently in the United States, dial-up access is available from a variety of large and small service providers, either for a fee of about $20 per month, or for no charge other than the willingness to view and, in some cases, propagate advertisements. In the latter case, the ISP's revenue comes from the advertisers who gain a captive audience of the ISP's subscribers as well as recipients of the subscribers' electronic mail. The ISP simply provides a bank of modems into which its own subscribers may dial. Switching to a competing provider involves no more than making a few phone calls and notifying one's correspondents of a new email address.

If no modem is available when a subscriber dials in, the subscriber's modem encounters a busy signal and is not able to connect. The modem may be programmed to repeatedly redial the access number until a connection is made, or the subscriber may have to manually redial later. Either way, the subscriber experiences an annoying delay in obtaining service. Repeated instances are likely to cause her to switch to a competing provider. The modem bank can be viewed as a system of parallel servers whose customers are limited to the pool of subscribers,

and for which no queue may form. A subscriber becomes a customer only when she attempts to dial in for service. Each server can serve only one customer at a time. Therefore, if all the servers are busy when a customer arrives, that *customer* is lost (however, the *subscriber* may not be lost to the ISP until much later, if ever). The service level can be measured as the probability that a modem is available when a customer arrives.

In this context, at any point in time the ISP must determine the size of its modem bank required to provide an acceptably high service level. This decision is critical because the company does not want to lose its subscribers due to unreliable service. At the same time, low access fees and limited advertising revenue mean that its profit margin is very narrow. Though some ISPs operate by buying access to other firms' modem banks, in this paper we assume that the firm owns all its own capacity. Therefore, managing its capital investment is critical for its survival.

The demand for service can be measured as the traffic intensity or *offered load* of the queuing system, which is defined as the product of the composite arrival rate of customers and the average service (connect) time. When confronted with increasing demand, the ISP must decide when to expand capacity by adding modems to its bank. Based on Bieler and Stevenson [1] and Dumortier [2], the Internet will grow exponentially over the foreseeable future. Therefore, the model used to minimize the present value of capacity expansion cost is based on an exponential growth assumption. For a given forecast of future demand, the size of the expansion at a given time point is simply the growth in demand for capacity over the interval until the next expansion time. We assume that there are both fixed and variable costs associated with each expansion.

This paper describes a new procedure to determine demand for capacity in a loss system and the application of a previously existing method to compute expansion time points that minimize the present value of expansion cost over an infinite horizon. We point out a useful characteristic of the Erlang loss function that appears to have hitherto gone largely unnoticed. An analysis of the sensitivity of the expansion policy to problem parameters reveals the effects of fixed and variable cost parameters, the demand growth rate, and the interest rate used to discount costs. Section II summarizes the relevant literature. In Section III we describe the methodology for determining demand for capacity and the optimal expansion time points. Section IV presents a numerical example and sensitivity analysis, and Section V concludes.

II. Background and Relevant Literature

Since the release in the mid 1990's of popular web browsers such as Netscape and Microsoft Internet Explorer, growth in the Internet has been very rapid. However, measuring the size of the Internet, much less predicting its future growth, is a daunting task. Measures of the Internet's scale include the number of Internet hosts, the numbers of dial-up and direct connections from homes and offices, and the numbers of residential and corporate subscribers. Rai, Ravichandran and Samaddar [3] compared the use of diffusion and nonlinear regression models to describe the growth in the number of Internet hosts, or nodes that serve as gateways to the rest of the network. Typically, several users are connected to each host. They estimated the models using data collected from 1981 to 1994 and then compared the fit of actual growth through 1997 with the model forecasts. They concluded that an exponential model with a growth rate of approximately 16% per year provided the best fit, though even this model slightly underpredicted growth during the mid-1990's.

Bieler and Stevenson [1] predicted the growth of the global Internet over 1998 to 2005 using several measures. Their forecasting model considered countries and regions separately and took into account technological, socio-economic and regulatory developments. Their forecasts for dial-up and permanent connections, residential and corporate subscribers and users, and private Internet hosts in each major region of the world all exhibited exponential growth. They predicted that most of the growth after 2002 will derive from developed Asia and Western Europe, not the United States. Dial-up connections will far exceed and outpace permanent connections. However, they predicted that revenue from providing Internet access to subscribers will decline over the forecast period and pointed out the need for ISPs to provide high quality and value-added services.

Since a modem bank provides no "waiting space" for customers when all the modems are occupied, it can be modeled as a loss system, i.e., a queuing system in which the maximum number of customers equals the number of servers. When arrivals follow a Poisson process, the steady state distribution for the number of customers in the system follows a truncated Poisson distribution. In particular, the well-known Erlang loss formula evaluates the probability $p_c(a)$ that an arriving customer finds all the servers busy and thus is lost to the system. This probability depends only on the number of servers, *c*, and the offered load, *a*, which equals the product of the arrival rate and the mean service time. It holds regardless of the service time distribution [4].

Because of the Erlang loss formula's widespread use in telecommunications and other service systems, many properties, approximations and bounds have been derived. Messerli [5] showed that $p_c(a)$ is convex in *c*, so that the load carried on the last server is monotonically decreasing in the number of servers. Jagerman [6] and Harel [7] derived bounds and

approximations that could be used in optimization problems such as server allocation. Krishnan [8] and Harel [9] focused on convexity and concavity properties of the formula with respect to arrival and service rates. As will be discussed further below, Krishnan, like Newell [4], also pointed out the asymptotic linearity of $p_c(a)$ in c for large values of a.

Capacity expansion models have a long history in operations research. Freidenfelds [10] and Luss [11] discuss an extensive set of mathematical models, ranging from the simple models such as linear deterministic demand to more complex models such as interaction of two different capacity types with both deterministic and stochastic demand. One of the earliest studies was by Sinden [12], who defined the concept of long run optimality and developed a general method for solving an infinite horizon problem by a sequence of finite horizon approximations. He showed that under certain conditions it is optimal to expand capacity at regular time intervals. Srinivasan [13] also proved this result for the particular case in which demand growth is geometric and equipment costs exhibit economies of scale following a power law. Expansion sizes must therefore increase geometrically at the same rate as demand. Ryan [14] generalized this model to stochastic demand growth with a geometric trend and lead times for adding capacity.

Snow [15] applied Srinivasan's model in a study of communications satellite capacity expansion and also explored the effects of complicating factors such as depreciation and technological change. Smith [16] studied capacity expansion for exponential demand growth with a choice of different facility types to install and developed an efficient method for solving for the optimal first facility by solving successively longer finite horizon problems. More recently, Berman and Ganz [17] developed capacity expansion models for service providers with many geographically dispersed facilities and pointed out the unique aspects of expanding capacity for a service that cannot be stored nor transported. Gaimon and Ho [18] used game theory to

study capacity acquisition by competing service providers in order to reduce prices and increase demand.

III. Methodology.

The analysis of the capacity expansion problem consists of two steps; first, determining the capacity required to provide a specified level of service, and second, deciding when to add capacity in order to maintain the same service level as demand increases.

1. Determining the Demand for Capacity

In this section we consider how to determine the demand for capacity based on demand for service. Each server (modem) can serve only one customer at a time. In contrast to many telephone systems, it is not possible to wait on hold for a server to become available. Therefore, when all servers are busy, the system cannot accommodate any other customers. Subsequent arrivals will simply get a busy signal. We first focus on the demand for capacity at a specific point in time.

Suppose there are n subscribers and, for $i = 1,...,n$, λ_i is the rate at which the i^{th} subscriber attempts to access an ISP modem. Based on the results of Çinlar [19], Denardo [20, pp.125-127] explained that if there is a large number of subscribers and each one is responsible for a small proportion of the number of access attempts, then the overall number of access attempts by subscribers is closely approximated by a Poisson process with rate $\lambda = \sum_{i=1}^{n} \lambda_i$. Let $1/\mu$ be the mean connect time for any subscriber once he or she succeeds in gaining access. We can model the modem pool as the queuing loss system originally devised by A. K. Erlang in 1917, in which the customers are subscribers who dial in, and a service consists of connection to a modem. The model assumes that incoming calls arrive as a Poisson stream; service times are mutually independent, identically distributed random variables; and all calls that arrive and find

every modem busy, that is, get a busy signal, are turned away. The possibility of retrials is addressed below. The steady-state number of customers in the system follows a truncated Poisson distribution with parameter $a = \lambda/\mu$, known as the offered load (see for example, Cooper [21]). If there are *c* servers, the probability that the system is full, or equivalently the probability that an arriving customer will be denied access, is given by

$$p_c(a) = \frac{a^c/c!}{\sum_{i=1}^{c} a^i/i!} \tag{1}$$

In this study, we focus on the number of servers or modems a provider needs in order to provide a specified service level, defined by the probability $1 - p_c(a)$. The offered load represents the demand for service. In view of the explosive growth of the Internet, we are interested in the required number of servers when the offered load is large and growing. As mentioned previously, Krishnan [8] and Newell [4] showed that if $a/c \gg 1$, then $p_c(a) \cong 1 - c/a$. Therefore, to achieve a specified service level $1 - p$, the required value of *c* is approximately equal to $(1-p)a$; that is, the required amount of capacity grows linearly with the offered load.

Though this asymptotic relation captures the qualitative behavior, the slope $1 - p$ of the linear relation is not very accurate. Figures 1 and 2 show the values of *c* needed to achieve each of three specified service levels for different magnitudes of the offered load. These were found by solving Equation (1) for

$$c^*(p;a) = \min\{c : p_c(a) < p\}, \tag{2}$$

where *p* is fixed. The results of simple linear regression indicate that the demand for capacity is closely approximated by a linear function of the demand for service. In fact, most of the

deviation from the regression line is due to the requirement that *c* be an integer. The empirically determined slope depends on $1-p$ but differs significantly from equaling it, particularly for smaller values of *a*.

The assumption of Poisson arrivals may not be realistic when retrials are considered. However, under this paper's assumption of a high service level, the impact of retrials on the distribution of the number of customers in the system is negligible. Though we generally do not assume any particular service time distribution, a Markovian model for a system allowing retrials such as in [22] provides insight. In this model, customers who arrive to find all servers busy are assumed to attempt access again, with the time between an individual customer's retrials exponentially distributed with rate ν (for consistency, the symbols ν and μ are reversed from the reference paper). Then if $C(t)$ is the number of customers in the system and $N(t)$ is the number of customers in the process of redialing at time *t*, the process $\{C(t), N(t), t \geq 0\}$ is a continuous time Markov chain with infinitesimal transition rates as follows:

For $0 \leq i \leq c-1$ and $j \geq 0$,

$$q_{(i,j)(m,n)} = \begin{cases} \lambda, & \text{if } (n,m) = (i+1, j) \\ i\mu, & \text{if } (n,m) = (i-1, j) \\ j\nu, & \text{if } (n,m) = (i+1, j-1) \\ -(\lambda + i\mu + j\nu), & \text{if } (n,m) = (i, j) \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

For $i = c$ and $j \geq 0$,

$$q_{(c,j)(m,n)} = \begin{cases} \lambda, & \text{if } (n,m) = (c, j+1) \\ c\mu, & \text{if } (n,m) = (c-1, j) \\ -(\lambda + c\mu), & \text{if } (n,m) = (c, j) \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

If $P_{ij} \equiv \lim_{t\to\infty} \Pr\left[C(t) = i, N(t) = j\right]$ is the stationary probability of *i* customers in the system and *j* customers redialing then the loss probability is given by $p_c = \sum_{j=1}^{\infty} P_{cj}$. According to equations (3) and (4), transitions to retrial states with *j* > 0 take place only if a potential arrival occurs when *i* = *c*. Therefore, if p_c is small, these states have low probability of occurrence and their contribution to the overall rate of transition from *i* to *i* + 1 customers in the system is negligible. Since we are setting the capacity, *c*, to guarantee that p_c is small, we neglect retrials in the remainder of the analysis.

Based on the empirical observations illustrated in Figures 1 and 2, we henceforth assume that demand for capacity (required number of servers) grows linearly with the demand for service (offered load) and apply the slope obtained from regression. Growth in the offered load, defined above as $a = \sum_{i=1}^{n} \lambda_i / \mu$, could occur as a result of an increase in the number of subscribers (*n*), the individual subscribers' dial-in rates (λ_i), the length of connect times ($1/\mu$), or any combination of these three factors. While the demand for service typically fluctuates daily and weekly, the capacity requirement for a time period depends on the largest demand predicted to occur over the period. The input to the capacity expansion model is a forecast of the offered load, $a(t)$, which represents the peak demand in period t. In the capacity expansion study, we take growth in the offered load over time as given and make no attempt to isolate its causes. The important point is that demand for capacity is obtained by a linear transformation of demand for service. Exponential growth in the demand for service translates into exponential growth in the demand for capacity, with the same growth rate.

2. Optimal Capacity Expansion Policy

In order to maintain a specified service level at all times, the firm must keep up with the growth in demand. Therefore, the firm needs to expand its capacity from time to time as the demand for service increases. The reasonable questions are: when are the best times for the firm to expand its capacity and by how much at each time? Assuming that demand for capacity is deterministic and that expansions are instantaneous, the two questions are two sides of the same coin. Given a series of expansion sizes, the firm can wait to expand its capacity until the current capacity equals the minimum capacity required based on the Erlang loss formula above with a specified service level. Equivalently, once the expansion time points are determined, the size of an expansion is simply the forecast growth in demand for capacity between the current time and the next expansion time point.

Let $D(t)$ be the demand for capacity at time t, obtained according to the Erlang loss formula by a linear transformation of the demand for service at time t. We assume $D(t+s) > D(t)$ for all t and $s > 0$. Without loss of generality, assume that at time 0, current capacity equals $D(0)$ so that an expansion is needed immediately. The *policy*, or vector of decision variables, is $\mathbf{t} = (t_1, t_2, ...)$, the sequence of subsequent expansion time points over an indefinite time horizon. Given **t**, the size of the n^{th} expansion, which takes place at time t_{n-1}, equals $D(t_n) - D(t_{n-1})$, where $t_0 = 0$. Our objective is to minimize the sum of discounted costs of expansions. In general, the effect of discounting is to delay expenditures; however, economies of scale in the cost of capacity encourage larger and less frequent expansions.

Sinden [12] originally defined the concept of a *long run optimal policy* as one to which no other policy is preferable in the long run. Let $Y(t, \mathbf{t})$ be the cumulative (discounted) cost of

expansions up to time *t* given the policy **t**. The policy $\mathbf{t}'$ is preferable to **t** if there exists a finite *T* such that for all *t* > *T*, $Y(t,\mathbf{t}') < Y(t,\mathbf{t})$. Sinden also outlined the following procedure to approximate one of possibly several long run optimal policies by optimizing over increasingly long time horizons.

Let $\mathbf{t}^* = (t_1^*, t_2^*, ...)$ be a long run optimal sequence of expansion time points. For some fixed future time point, *t*, and an integer, *k*, we can find the optimal expansion time points given that there are *k* expansions up to time *t*. Let

$$Y_k(t) = \min\{Y(t,\mathbf{t}) : \mathbf{t} \text{ such that } t_k = t\} \tag{5}$$

and let t_k' be the value of *t* such that $Y_k(t) = Y_{k+1}(t)$. Clearly, for small values of *t*, a single expansion is optimal, and as *t* increases, the optimal number of expansions grows as well. That is, for time horizons *t* between t_{k-1}' and t_k', $Y_k(t)$ is the smallest of $\{Y_1(t), Y_2(t), ...\}$. Sinden illustrates the typical shapes of the functions $\{Y_k(t), t \geq 0\}$ and the increasing values of $\{t_k', k \geq 1\}$. Furthermore, when optimizing up to time t_k', the policies that attain $Y_k(t_k') = Y_{k+1}(t_k')$ provide upper and lower bounds on t_n^* for $n \leq k$. Let

$$\overline{t}_n^{(k)} = t_n \text{ in } \operatorname{argmin}\{Y(t_k', \mathbf{t}) : \mathbf{t} \text{ s.t. } t_k = t_k'\} \tag{6}$$

and

$$\underline{t}_n^{(k)} = t_n \text{ in } \operatorname{argmin}\{Y(t_k', \mathbf{t}) : \mathbf{t} \text{ s.t. } t_{k+1} = t_k'\}, \tag{7}$$

where $\arg\min\{\cdot\}$ denotes the argument, i.e., the vector **t** in equations (6) and (7), that minimizes the quantity in brackets. Then $\lim_{k\to\infty} \underline{t}_n^{(k)} = t_n^* = \lim_{k\to\infty} \overline{t}_n^{(k)}$. As will be demonstrated in the numerical

example, for each *n*, the gap between the lower and upper bounds on the optimal n^{th} expansion time point decreases as the finite time horizon is lengthened.

IV. Numerical Illustration

Bieler and Stevenson [1] provided a forecast of global growth in dial-up subscriptions over a seven-year time horizon. We experimented with several types of regression models and found that the closest fit to the forecast was achieved by an exponential function, as illustrated in Figure 3. The exponential growth rate is approximately 18% per year, with a coefficient of determination of 99.27%. Though the raw numbers obviously do not apply to a single ISP, for the sake of illustration we will assume that demand for service could increase at this rate.

Assume that the peak demand for service (offered load) in year *t* is given by $a(t) = a_0 e^{gt}$. Then based on the analysis in Section III, the demand for capacity is given by $D(t) = ma(t) + b$, where the slope, *m*, and intercept, *b*, are found from linear regression as in Figure 1 or 2. Given a sequence of expansion times **t**, the size of the n^{th} expansion is given by $m\left(a(t_n) - a(t_{n-1})\right)$. (The intercepts in Figures 1 and 2 are irrelevant.) Finally, we assume that the cost of capacity is composed of a fixed cost, *c*, plus a variable cost of *d* per unit of capacity. Then the discounted cost of the n^{th} expansion is given by $c + dma_0\left(e^{gt_n} - e^{gt_{n-1}}\right)$. Therefore,

$$Y_k(t) = \min_{t_1,\ldots,t_{k-1}} \left\{ \sum_{n=1}^{k} e^{-rt_{n-1}} \left[c + dma_0\left(e^{gt_n} - e^{gt_{n-1}}\right) \right] \right\}, \tag{8}$$

where $t_0 = 0$ and $t_k = t$. Note that the parameters *d*, *m*, and a_0 combine multiplicatively. An increase in the variable cost of capacity, the required service level (which determines *m*), or the initial demand for service would each have the same effect on the optimal expansion times.

A numerical example illustrates the model and allows analysis of the sensitivity of the optimal policy to the model's parameters. We initially set g = 0.18, r = 0.20, c = 100, d = 25,

and the initial offered load, a_0 = 400. We used a study horizon of seven years to match the forecast period in Figure 3. The high initial demand for service combined with the large growth rate implies that the offered load over the study horizon would be in the range covered in Figure 2. Assuming a required service level of 99.99%, we applied the corresponding slope of the regression line for demand for capacity, m = 1.05. We used *Mathematica* [23] to optimize Equation (5) simultaneously over an increasing number of expansion time points over successively longer study horizons. Figure 4 shows how the lower and upper bounds on the expansion time points from Equations (3) and (4) converge to their optimal values as the time horizon is increased. By the end of the seven-year forecast period, the difference between the upper and lower bounds on the optimal first expansion time is less than 5%. If the forecast can be extended out to ten years, this difference drops to 2%. The gap between the bounds for subsequent expansion times is wider. However, in practice, the optimal expansion times could be obtained in a rolling fashion. Using the initial forecast at time 0, one could closely approximate t_1^* and install capacity equal to $D\left(t_1^*\right) - D(0)$. At the time, *t*, when demand had risen to equal capacity again, the manager could adjust the forecast and use the same procedure to approximate t_2^*, install capacity $D\left(t_2^*\right) - D(t)$, and so forth.

Note that the intervals between the optimal expansion times narrow as demand increases more rapidly. The equal intervals between expansions obtained by Sinden [12] and Srinivasan [13] depend on a power function for capacity cost economies of scale and do not apply to the fixed-plus-variable cost studied in this paper.

Gauging the sensitivity of the optimal policy to problem parameters is perhaps more valuable than the results for any particular numerical example. The demand growth rate, the rate of return, and the fixed and variable costs are all quantities that are difficult to estimate or

forecast. Though the optimal policy consists of an indefinitely long sequence of expansion times, the first one is the most important since it determines how much capacity must be installed at time 0. In the following we look at the sensitivity of t_1^* to changes in *g*, *r*, *c*, and *d*, changing one parameter at a time with all else held constant.

One would expect that increasing the fixed or overhead cost of an expansion, *c*, would result in larger and less frequent expansions. Conversely, increasing the cost of each unit of capacity, *d*, would prompt more frequent, smaller expansions. Figures 5 and 6 confirm these intuitions and show the second order effects as well. The value of t_1^* is increasing and concave in *c* and decreasing convex in *d*. Also, since the effect of changing *m* or a_0 is identical to that of changing *d*, either a larger initial demand for service or a more stringent service level requirement dictate that the next expansion should occur earlier. However, these effects diminish as the changes grow larger.

An increase in the demand growth rate, *g*, requires more frequent expansions, as shown in Figure 7. The value of t_1^* is also convex in *g*. However, the effect of increasing the interest rate, *r*, or the rate of return demanded of investments, is more difficult to predict. On one hand, delaying expansions means that the accompanying costs will be pushed farther in the future where they are more heavily discounted. On the other hand, if expansions are less frequent then they and their costs must be larger. In particular, the cost of the initial expansion at time 0, which is not discounted at all, increases with t_1^*. Figure 8 shows how t_1^* changes with *r*, in this case using a demand growth rate of $g = 0.10$. It demonstrates that the second effect of increased discounting dominates the first, so that a higher rate of return means that costs should be spread more evenly over time by small, frequent expansions.

V. Conclusions

This paper presented a model to determine the number of servers needed to accommodate growing subscriber demand in a loss system. The number of servers needed increases approximately linearly as the offered load increases. Therefore, the growth rate in the demand for service translates directly into the same growth rate for the demand for capacity. Moreover, in a loss system, growth in peak demand for service could result from an increase in the number of subscribers, an increase in the average rate at which each one attempts to access the system, or a longer average connect time. These three influences could combine in various ways to increase the offered load.

In order for the company to stay competitive, it must determine the best time to install additional capacity. The study found that an increase in the demand growth rate will reduce the optimal time intervals between expansions. An increase in the rate of return required of investments has the same effect. As for the cost parameters, both fixed cost and equipment cost influence the optimal time interval. An increase in the fixed cost will extend the length of the optimal time intervals. On the other hand, an increase of equipment cost or the required service level will reduce the optimal time intervals.

Given a forecast for demand, the long run optimal expansion time points can be approximated by solving successively longer finite horizon optimization problems. The procedure for approximating the optimal expansion time points can be carried out easily with current numerical optimization software. The ability to uncover one expansion point at a time in a rolling fashion allows demand forecasts to be updated as more information becomes available.

Further research in this area could address multiple layers of equipment. For example, an ISP uses not only modems, but routers and dedicated phone line are essential components of its

service provision. The number of components of equipment at the lower end is typically a multiple of the number of components at the upper end. Therefore, the capacity expansion problem for the system is really a set of interrelated problems at each equipment level.

It may be possible through the firm's pricing structure to manipulate the number of subscribers and their frequency and duration of use of the system. Together with the specified service level, these determine the demand for capacity, which in turn determines the capacity cost. The results of this paper's model could be used as part of an overall effort to maximize the firm's profitability with the pricing structure and service level as decision variables.

Finally, technological innovation from the newcomers in the industry can cause an established firm to lose subscribers. Current examples in the ISP industry are the Digital Subscriber Line (DSL), cable modems, and free Internet. Each firm must determine the set of technologies to invest in as well as when to invest in additional capacity. Furthermore, technological change typically results in decreasing capacity costs over time. In continuing research we are examining the ways technological change affects the optimal expansion policy.

Acknowledgment

This work was supported in part by the National Science Foundation under grant DMI-9996373.

References

1. Bieler D, Stevenson I. Internet Market Forecasts: Global Internet Growth 1998-2005: Ovum; 1998 December, 1998. Report No.: Ovum Report.

2. Dumortier P. Shortcut techniques to boost Internet throughput. Alcatel Telecommunications Review 1997(4th Quarter):300-306.

3. Rai A, Ravichandran T, Samaddar S. How to anticipate the Internet's global diffusion. Communications of the ACM 1998;41(10):97-106.

4. Newell GF. Applications of Queueing Theory. 2nd ed. New York: Chapman and Hall; 1982.

5. Messerli EJ. Proof of a convexity property of the Erlang B formula. Bell System Technical Journal 1972;51:951-953.

6. Jagerman DL. Some properties of the Erlang loss function. Bell System Technical Journal 1974;53(3):525-551.

7. Harel A. Sharp bounds and simple approximations for the Erlang delay and loss formulas. Management Science 1988;34(8):959-972.

8. Krishnan KR. The convexity of loss rate in an Erlang loss system and sojourn in an Erland delay system with respect to arrival and service rates. IEEE Transactions on Communications 1990;38(9):1314-1316.

9. Harel A. Convexity properties of the Erlang loss formula. Operations Research 1990;38(3):499-505.

10. Freidenfelds J. Capacity Expansion: Analysis of Simple Models with Applications. New York: North-Holland; 1981.

11. Luss H. Operations research and capacity expansion problems: a survey. Operations research 1982;30(5):907-947.

12. Sinden FX. The replacement and expansion of durable equipment. J.Soc.Indust,Appl.Math, 1960;8(3):466-480.

13. Srinivasan TN. Geometric rate of growth of demand. In: Manne AS, editor. Investments for Capacity Expansion: Size, Location, and Time-Phasing. Cambridge: MIT Press; 1967. p. 150-156.

14. Ryan SM. Capacity expansion for random exponential demand growth. Ames, IA: Industrial & Manufacturing Systems Engineering, Iowa State University; 2000 August. Working Paper No. 00-109.

15. Snow MS. Investment cost minimization for communications satellite capacity: refinement and application of the Chenery-Manne-Srinivasan model. Bell Journal of Economics 1975;6(2):621-643.

16. Smith RL. Turnpike results for single location capacity expansion. Management Science 1979;25(5):474-484.

17. Berman O, Ganz Z. The capacity expansion problem in the service industry. Computers & Operations Research 1994;21(5):557-572.

18. Gaimon C, Ho JC. Uncertainty and the acquisition of capacity: a competitive analysis. Computers & Operations Research 1994;21(10):1073-1088.

19. Çinlar E. Superposition of point processes. In: Lewis P, A. W., editor. Stochastic Point Processes: Statistical Analysis, Theory, and Applications. New York: Wiley-Interscience; 1972. p. 549-606.

20. Denardo EV. Dynamic Programming: Models and Applications. Englewood Cliffs, NJ: Prentice-Hall; 1982.

21. Cooper RB. Introduction to Queuing Theory. 2nd ed. New York: North-Holland; 1981.

22. Falin G. A survey of retrial queues. Queueing Systems 1990;7:127-168.

23. Wolfram S. The Mathematica Book. 4th ed. Cambridge: Cambridge University Press; 1999.

Vitae

Alexander Simampo received his B.S and M.S. in Industrial Engineering from Iowa State University. He has worked as an Industrial Engineer Consultant and has several years working experience in a shoe manufacturing plant and a foundry plant.

Sarah M. Ryan is an Associate Professor of Industrial and Manufacturing Systems Engineering at Iowa State University. She holds a B.S. in Systems Engineering from the University of Virginia and M.S.E. and Ph.D. degrees in Industrial and Operations Engineering from the University of Michigan. Her research interests are in stochastic modeling and dynamic optimization. She is a senior member of IIE and a member of INFORMS and ASEE.

Figure Captions

Figure 1. Number of servers required to achieve a specified service level as a function of a moderate offered load.

Figure 2. Number of servers required to achieve a specified service level as a function of a large offered load.

Figure 3. Forecast of the number of dial-up Internet connections.

Figure 4. Convergence to optimal expansion times (at left); and illustration of optimal expansion policy (at right).

Figure 5. Effect of variable cost of capacity on the optimal time to the first expansion.

Figure 6. Effect of fixed expansion cost on the optimal time to the next expansion.

Figure 7. Effect of demand growth rate on the time to the next expansion.

Figure 8. Effect of the rate of return on the time to the next expansion.

Figure 1.

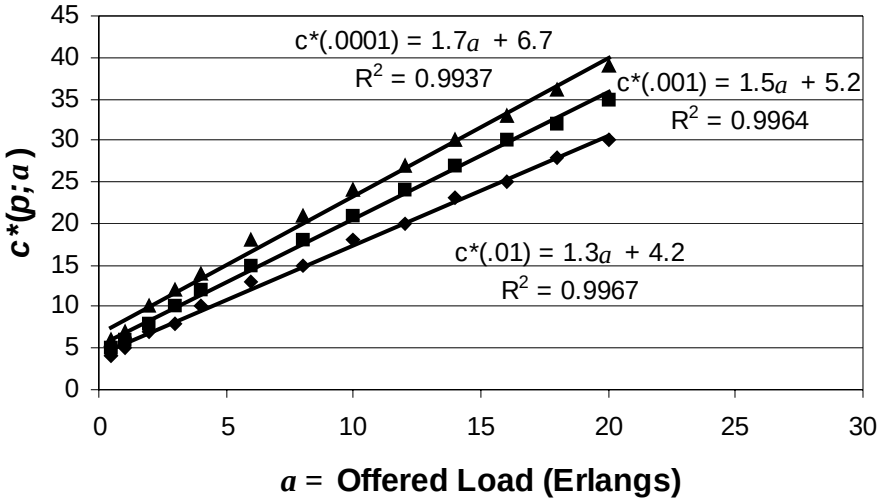
c*(.0001) = 1.7a + 6.7
R2 = 0.9937
c*(.001) = 1.5a + 5.2
R2 = 0.9964
c*(.01) = 1.3a + 4.2
R2 = 0.9967
c*(p;a)
a = Offered Load (Erlangs)
0
5
10
15
20
25
30
35
40
45

Figure 2.

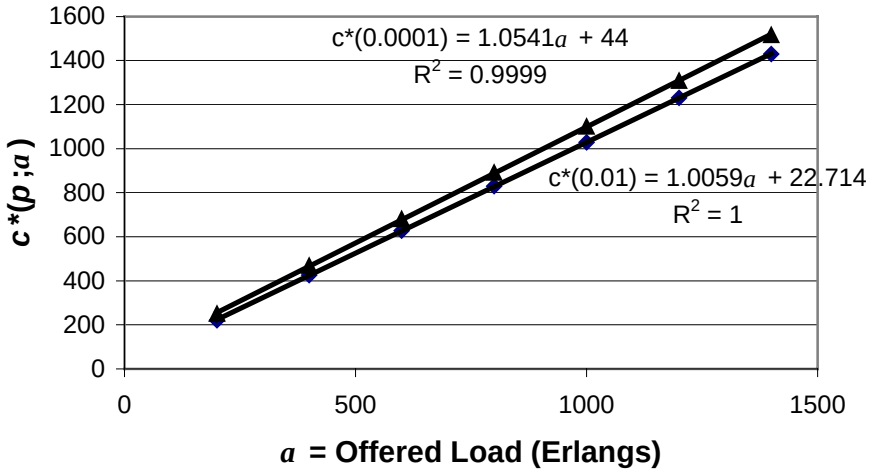
c*(0.0001) = 1.0541a + 44
R2 = 0.9999
c*(0.01) = 1.0059a + 22.714
R2 = 1
c*(p;a)
a = Offered Load (Erlangs)
0
200
400
600
800
1000
1200
1400
1600
0
500
1000
1500

Figure 3.

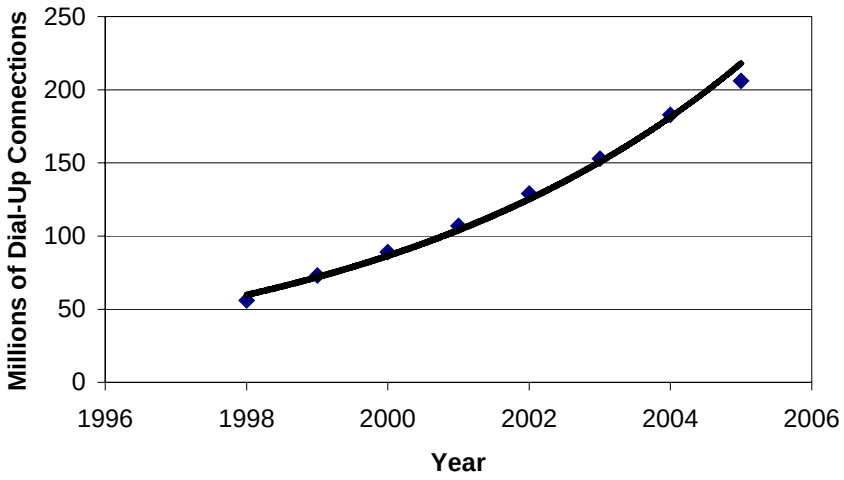
Millions of Dial-Up Connections
250
200
150
100
50
0
1996
1998
2000
2002
2004
2006
Year

Figure 4.

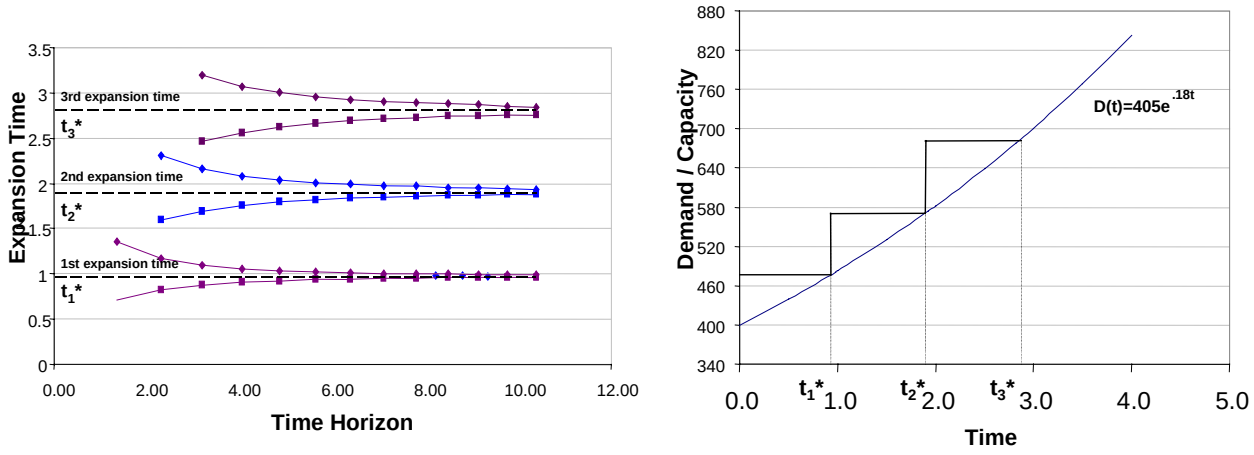
Expansion Time
Time Horizon
3rd expansion time
2nd expansion time
1st expansion time
t3*
t2*
t1*
0.00
2.00
4.00
6.00
8.00
10.00
12.00
0
0.5
1
1.5
2
2.5
3
3.5
Demand / Capacity
Time
D(t)=405e.18t
340
400
460
520
580
640
700
760
820
880
0.0
1.0
2.0
3.0
4.0
5.0

Figure 5.

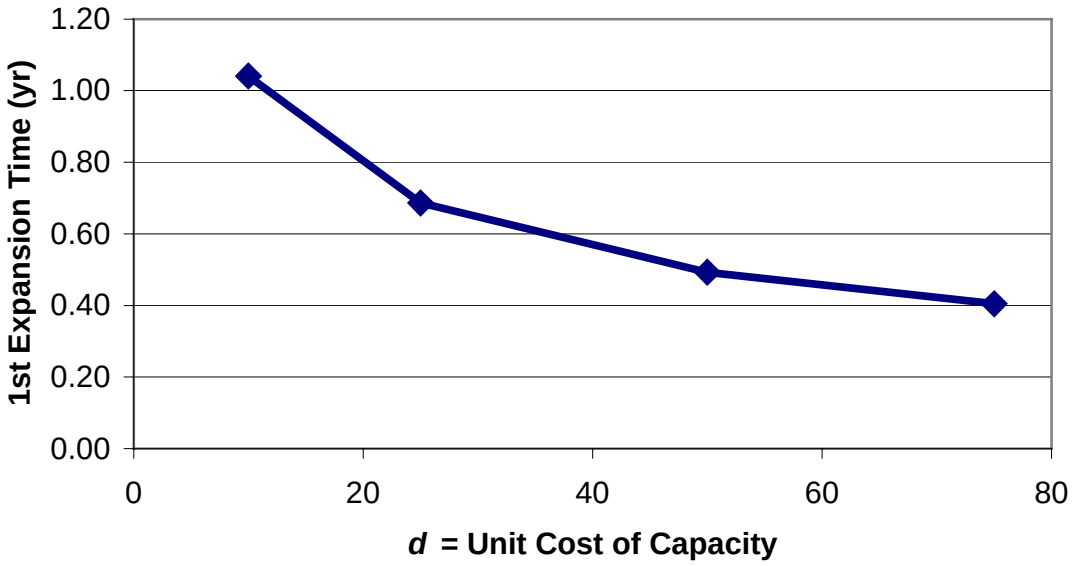
1st Expansion Time (yr)
1.20
1.00
0.80
0.60
0.40
0.20
0.00
0
20
40
60
80
d = Unit Cost of Capacity

Figure 6.

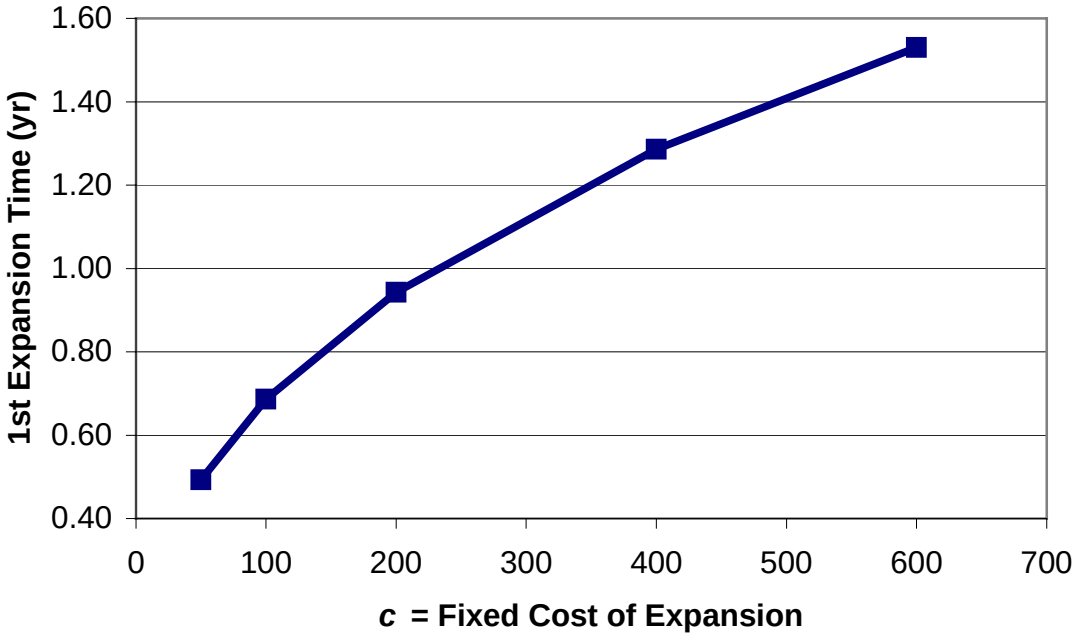
1.60
1.40
1.20
1.00
0.80
0.60
0.40
1st Expansion Time (yr)
0
100
200
300
400
500
600
700
c = Fixed Cost of Expansion

Figure 7.

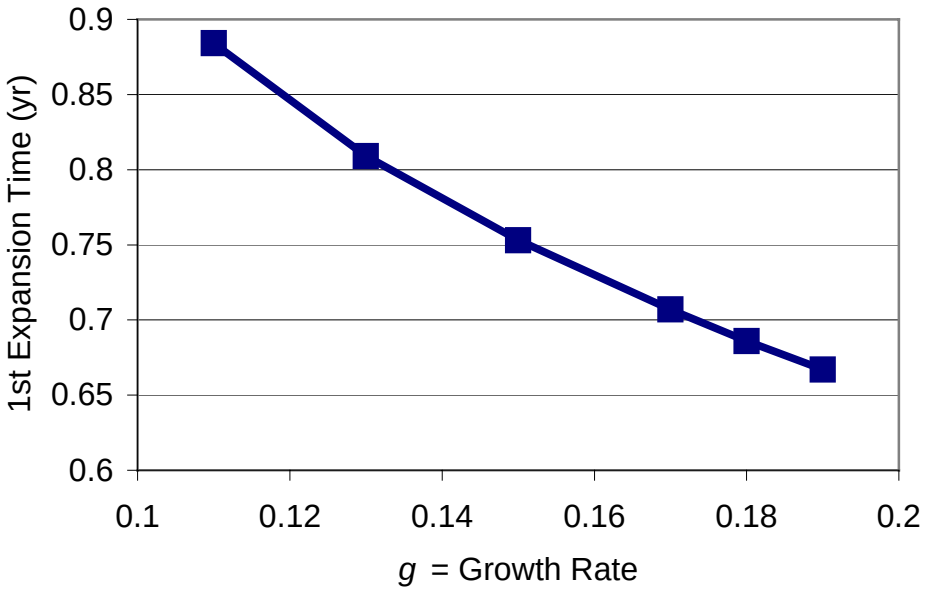
0.9
0.85
0.8
0.75
0.7
0.65
0.6
1st Expansion Time (yr)
0.1
0.12
0.14
0.16
0.18
0.2
g = Growth Rate

Figure 8.