Telecommunication Systems 19:1, 9–38, 2002

A Time-Dependent Queueing-Network Model to Describe the Life-Cycle Dynamics of Private-Line Telecommunication Services

CLEMENT McCALLA cmccalla@ems.att.com
AT&T, Room C5-2W03, 200 S. Laurel Avenue, Middletown, NJ 07748, USA

WARD WHITT wow@research.att.com
AT&T Labs, Room A117, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971, USA

Received 4 May 1999; Revised 22 February 2001

Abstract. We propose a model to help a service provider manage a family of private-line telecommunication services. The model is a time-dependent network of infinite-server queues. The queueing network is used to model switching from one service to another. The relevant time scale is quite long: service lifetimes are measured in years, while service life cycles are measured in decades. To capture changing technology and customer preferences over this extended period, the model includes time-dependent new-connection rates, time-dependent switching rates and general time-dependent service-lifetime distributions for the different services. Because of the long service lifetimes, there is typically a significant time lag between the peak arrival rate and the peak expected number of customers receiving service. Thus sales and revenue do not move together; instead sales is a leading indicator of revenue in a service life cycle. The network structure reveals how the life cycles of different services are related. We show that it is possible to reasonably fit a version of this relatively complex model to data and then analyze the model to obtain useful descriptions of system dynamics.

Keywords: private-line telecommunication services, life cycles, product management, marketing, queueing networks, networks of infinite-server queues, time-dependent queues

1. Introduction

In this paper we introduce a time-dependent queueing-network model that can be used to help a service provider manage a family of private-line telecommunication services. Private lines provide dedicated bandwidth between two or more locations, typically to support data communication applications. Service providers such as AT&T have families of private-line offerings at bandwidths varying from under 64 kilobits per second to 156 megabits per second (OC3), with even higher bandwidths anticipated in the near future. Customers lease private lines from a service provider, with orders ranging from single lines to hundreds of lines involved in a customer network. During the lifetime of the private line, the service provider collects revenue on a monthly basis. The global revenue derived from private line services was estimated to be $14 billion in 1996, with a growth rate of 8% [7]. We primarily want to assist product management in managing the revenue streams, coordinating the responsibilities for pricing and marketing the family of private-line services, and equipping the network platform with the requisite type and level of network resources.

We introduce a mathematical model to help understand and manage the private-line business. We want to be able to predict future service lifetimes and connection request rates for each private-line service, and thus predict future sales, costs and revenue. We want to quantify both the expected values and the level of uncertainty. We want to expose the life-cycle dynamics of the different private-line services.

Understanding the private-line business is complicated by the very long service lifetimes, which typically are measured in years. The long service lifetimes imply that we need historical data over an extended time period. At the same time, there are rapid changes in technology and customer preferences, as illustrated by the recent explosion of Internet use. Recent developments in the domestic marketplace for private lines have witnessed unprecedented high demand accompanied by a migration to the higher bandwidth services, along with competitive pressure to reduce price. The great rate of change suggests that the future cannot be accurately predicted simply by extrapolating from the past. Thus, an appropriate tool should be able to provide predictions exploiting both historical data and marketing and technology projections.

We propose a queueing network model for the private line services, with the queues or nodes representing different private-line services. The number of jobs or customers at each queue is the number of private lines of a particular service. Since our main concern is with product management, we do not focus on the location of these lines (but see section 10). The service times at each queue are the service lifetimes of the private lines. The external arrival processes are new connection requests for that private-line service. The flows from one queue to another are switching from one private-line service to another. We especially want to distinguish between new demand and switching.

In order for the tool to be useful, e.g., to be able to rapidly answer "what if" questions, we need to be able to rapidly calculate the desired performance measures, using approximations if necessary. We need the tractability of the Queueing Network Analyzer (QNA) and similar performance analysis tools; see [29, 31, and references therein]. However, the very long service times and rapid changes in technology and customer preferences suggest using a very different queueing network model. In particular, for private lines it is important to represent time-dependent connection rates, disconnect rates and switching rates. Hence, instead of the stationary model and steady-state analysis in QNA, we propose a time-dependent queueing network model and time-dependent analysis.

In order to achieve tractability with a time-dependent queueing network model, we also make a simplifying assumption. We assume that there is always ample capacity; i.e., we assume that each private-line service request can be met with negligible delay or blocking (loss). We represent the ample capacity by letting every queue in the network have infinitely many servers. The infinite-server assumption greatly simplifies the analysis, making the system a linear system. The infinite-server assumption is often reasonable when addressing product management concerns, but it should be recognized that this assumption can potentially limit the applicability of the model.

To analyze a time-dependent network of infinite-server queues, we apply methodology from [20], but we also address an important feature not considered there. In particular, arrivals often occur in batches, with the batch identity not necessarily being maintained while the lines are in service. For example, there might be new connections of four lines and ten lines at two different times, and then disconnects of three lines, six lines and one line at three different times. We are not concerned with describing these details exactly, but we want to provide a reasonably accurate description of the overall (macroscopic) behavior. Fortunately, we are able to make a relatively careful and accurate analysis of the expected number of lines in service over time, and develop a rougher approximation for the associated time-dependent probability distributions. Thus we regard the mean number of lines in service as the main prediction, and the standard deviation and probability distribution as refinements.

Here is how the rest of this paper is organized. We describe related literature in section 2. In section 3 we describe the business process for a family of private-line services, highlighting the role of product management. In sections 4–6 we focus on a single private-line service. In section 4 we develop the specific model and show how it can be used to calculate the time-dependent mean number of lines in service and describe the life-cycle behavior. In section 5 we consider the problem of fitting the model to data. We discuss parametric model simplifications dictated by the need to fit the model to data. We also show that the analysis of the model simplifies after making these additional assumptions. In section 6 we develop the method to approximately calculate the full distribution of the number of lines in service.

In sections 7–9 we extend the model to a time-dependent network of infinite-server queues to model a family of private-line services. We discuss an important decomposition property making it possible to analyze the individual queues separately after determining the internal arrival rates (combining exogenous input with input from switching from other services). As in section 5, the need to fit the model to data dictates making model simplifications. In sections 8 and 9 we show how the network can be analyzed under such model simplifications. We give illustrative numerical examples.

In section 10 we briefly discuss the application of the model to short-term forecasting and capacity management. We introduce a special approach for doing short-term forecasting that produces reliable variance estimates, allowing for very different arrival and departure batches. This approach can be used to set thresholds on the deviation between the forecast and the actual demand that trigger warning signals. Finally, we present our conclusions in section 11.

2. Related literature

Private-line service over the telephone network has a long history that goes back to the invention of the telephone and thereby precedes most other telecommunication services including switched voice. For an account of some of that history, see [21,23]. In the literature there are several papers on mathematical modelling of private-line services. Nucho [25] used a birth–death process model to capture the transient behavior. Smith [30] was the first to recognize the value of a time-dependent infinite-server queue to model private-line services. Specifically, he assumed that service orders arrive according to a nonhomogeneous Poisson process with exponential arrival rate, each service order is a random batch of private lines with a general distribution, and the lifetime of the order (entire batch) has an exponential distribution. Smith's model was used by Doverspike and Jha [9] in a study of routing methods of private lines in a network. An infinite-server model for private lines related to Smith's but allowing more general new-connection rates was studied by Jennings et al. [16]. We extend the models of Smith [30] and Jennings et al. [16] in several ways: by not requiring that the batch stay together while in service, by allowing more general time-dependent arrival processes, and by introducing a network of queues to represent a family of private-line services with switching from one service to another.

There has been some work on analyzing the more difficult problem of time-dependent queueing networks with significant delays; see [18,26,28]. A principal motivating example for that work is the airport network, where airplane delays clearly play an important role. There is also a substantial literature on infinite-server queues and networks of such queues, much of which is cited in [10,20]. Networks of infinite-server queues are also known as stochastic compartmental models; see [13,14,22].

We have indicated that our model should help understand marketing strategies. For related work in this direction, see [3,4,6,24]. Unlike much of that literature, we do not attempt to model customer response to advertising. There already is considerable understanding of product life cycles. For example, in a situation closely paralleling private line services, Potts [27] describes dual life cycles in the computer industry due to (i) revenue from product sales and (ii) from servicing the installed base. Thus, our contribution is not so much the general structure, but a way to quantify it.

3. The role of product management

In the introduction we indicated that our proposed methodology is intended to assist product management. In this section we describe the role of product management in the business process.

A simplified view of the business process for a family of private-line services is shown in figure 1. Negotiations are conducted in the sales process to fulfill a service request, which can range from one to hundreds of private lines. At the conclusion of a successful sale, the provisioning process is initiated to assemble the private line from the inventory of network resources maintained by the capacity management process. If successful, service is provided to the customer; otherwise it is not (or perhaps delayed). During the lifetime of a private line, the billing process collects revenue on a monthly basis, and the maintenance process resolves quality problems when necessary. Shortly before the end of the lifetime of the private line, negotiations are once more conducted in the sales process. The customer can elect to migrate to another private-line service or

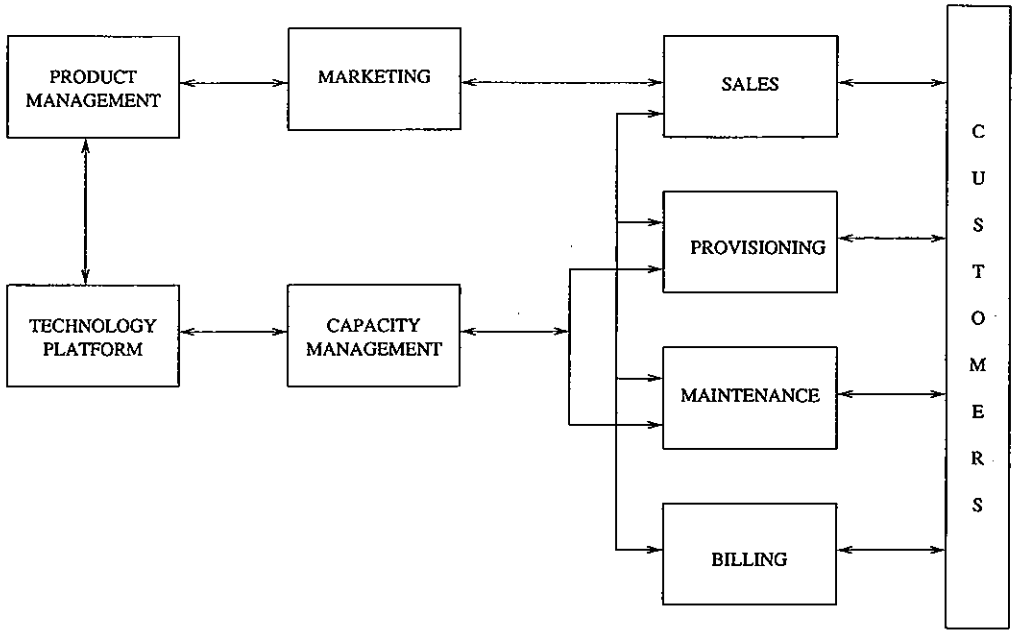

Figure 1. Private-line business process.

leave the system altogether, possibly to subscribe to another service or to another service provider. At the end of the lifetime of the private line service, the network resources used by the private line are returned to a common pool for reuse by other customers.

The marketing process focuses on the revenue side of the profit equation, while the technology platform addresses the cost side. The marketing process sets the price levels, initiates promotion efforts when necessary, establishes the service features (performance, reliability, etc.), and strives to maintain healthy relationships between the members of the family of private-line services. The technology platform process ensures that the network is outfitted with the appropriate technology in the appropriate configuration to support the evolving needs of family of private line services in a graceful and cost-effective manner.

In order to meet corporate profitability objectives, product management develops and executes a business strategy plan that is coordinated across the various business functions. This plan includes a short-term (1-year horizon) aggregate forecast for each service that is used to set the budget and to monitor results, and a long term (5-year horizon) aggregate forecast to set long-term goals and manage the technology platform process. The formulation and execution of this business strategy depends on knowledge of the salient characteristics, dynamic behavior and coupling of the service life cycles. The acquisition and application of this knowledge is the objective of this paper.

In order to meet customer demand at any place and at any time with the right amount, the capacity management process needs a forecast disaggregated to the network nodes and links. This is discussed briefly in section 10. It is possible to do more reliable prediction for product management because it takes an aggregate view over all facilities.

4. A single private-line service

Before considering the full queueing-network model for a family of private-line services, we develop the queueing model for a single private-line service. Let customers arrive according to a stochastic point process $\{A(t)\colon t > -\infty\}$; i.e., $A(t)$ is the number of arrivals up to time t. There could be a definite starting time t_0, so that $A(t_0) = 0$, but we have not specified it. However, we assume that the distant past is not relevant. Let $A(t)$ be an orderly point process (i.e., have points occurring one at a time) with time-dependent arrival rate $\gamma(t)$, i.e.,

$$E\big[A(t)\big] = \int_{-\infty}^{t} \gamma(u)\,\mathrm{d}u, \quad t \geqslant 0, \tag{4.1}$$

where $E[A(t)] < \infty$. Having the mean finite implies that the distant past is indeed not relevant.

Given that a customer arrives at time t, let $B(t)$ be the (random) number of lines requested and let $S_j(t)$, $1 \leqslant j \leqslant B(t)$, be the (random) service lifetimes of these $B(t)$ lines. Let $Q(t)$ be the number of lines in service at time t. We allow the service lifetimes associated with a batch $B(u)$ arriving at time u to be dependent, e.g., we could have $S_1(u) = S_2(u)$, corresponding to identical lifetimes. We assume that the service lifetimes associated with a given batch have identical cdf's; i.e., for all j,

$$G_u(t) \equiv P\big(S_j(u) \leqslant t\big), \quad t \geqslant 0. \tag{4.2}$$

The common distribution assumption is without loss of generality because we can assume that the $B(t)$ indices are randomly permuted before being selected. Let $G_u^{\mathrm{c}}(t) \equiv 1 - G_u(t)$ be the complementary cdf (ccdf). Let $B(u)$ have a distribution depending on u, but otherwise let $B(u)$ be independent of the arrival process. Similarly, let $S_j(u)$ have a distribution depending on u but be otherwise independent of the batch size $B(u)$ and the arrival process.

With the framework above, $Q(t)$ can be expressed as the stochastic integral

$$Q(t) = \int_{-\infty}^{t} \sum_{j=1}^{B(u)} \mathbb{1}_{\{S_j(u) > t-u\}}\,\mathrm{d}A(u), \tag{4.3}$$

where $\mathbb{1}_E$ is the indicator function of the event E, i.e., $\mathbb{1}_E = 1$ when E holds and 0 otherwise; see [20, section 2] for background on the construction. Letting A_k be the kth arrival time, we can also express (4.3) as

$$Q(t) = \sum_{k=1}^{A(t)} \sum_{j=1}^{B(A_k)} \mathbb{1}_{\{S_j(A_k) > t-A_k\}}. \tag{4.4}$$

It is significant that the mean of $Q(t)$ and the departure (disconnect) rate $\lambda^-(t)$ can be expressed relatively simply. Because of the aggregation over all facilities associated with the product management view, we anticipate that the number of lines in service, $Q(t)$, will be quite close to its mean, so that the mean is the primary description of interest.

Theorem 4.1. Under the assumptions above, the mean number of lines in service at time t is

$$m(t) \equiv EQ(t) = \int_{-\infty}^{t} G_u^{\mathrm{c}}(t-u)\lambda^{+}(u)\,\mathrm{d}u \tag{4.5}$$

and the disconnect rate at time t is

$$\lambda^{-}(t) = \int_{-\infty}^{t} G_u(t-u)\lambda^{+}(u)\,\mathrm{d}u, \tag{4.6}$$

where $\lambda^{+}(t)$ is the total arrival rate, i.e.,

$$\lambda^{+}(t) = \gamma(t)\beta(t), \quad t \geqslant 0, \tag{4.7}$$

with $\beta(t) = EB(t)$.

Proof. See [20, (2.13)]. □

Formulas (4.5) and (4.6) coincide with corresponding formulas in the $M_t/GI_t/\infty$ model with nonhomogeneous Poisson arrival process having arrival-rate function $\lambda^{+}(t)$ and independent service times having time-dependent general service-time cdfs $G_u(t)$; see [10, remark 4]. As noted in [20, remark 2.3], the mean formula (4.5) is valid much more generally, in particular, in our setting without the Poisson property. Also, the service times within a batch may be dependent. However, we cannot conclude that $Q(t)$ has a Poisson distribution with the given mean or that the departure process is a Poisson process, as we can for the $M_t/GI_t/\infty$ model. We develop an approximation for the distribution of $Q(t)$ in section 6.

Thus, given the model primitives $\gamma(t)$, $\beta(t)$ and $G_u(t)$, it is possible to calculate first the total arrival-rate function $\lambda^{+}(t)$ by (4.7) and then the mean $m(t)$ and departure-rate function $\lambda^{-}(t)$ by (4.5) and (4.6). We can describe the service life cycle in terms of the three functions $\lambda^{+}(t)$, $m(t)$ and $\lambda^{-}(t)$. We assume that all three functions are unimodal, first increasing and then decreasing. We also assume that there are time lags in the peaks of $m(t)$ and $\lambda^{-}(t)$ behind the peak of $\lambda^{+}(t)$; i.e., the peaks of $m(t)$ and $\lambda^{-}(t)$ occur after the peak of $\lambda^{+}(t)$. (Later these properties will be deduced after making additional model assumptions.) It is thus natural to define three phases of a typical service life cycle:

(1) *growth* – when all three functions $\lambda^{+}(t)$, $m(t)$ and $\lambda^{-}(t)$ are increasing,

(2) *mature* – when $\lambda^{+}(t)$ is decreasing, but one or more of $m(t)$ and $\lambda^{-}(t)$ is still increasing, and

(3) *decline* – when all three functions $\lambda^{+}(t)$, $m(t)$ and $\lambda^{-}(t)$ are decreasing.

A typical service life cycle is depicted in figure 2. The three phases of a service life cycle are illustrated from actual add and disconnect data for three different AT&T private line services in the time period 1991–1997 in figures 3–5. In doing so, we ignore fluctuations in a short time scale, and focus on the main trend. Figure 3 illustrates

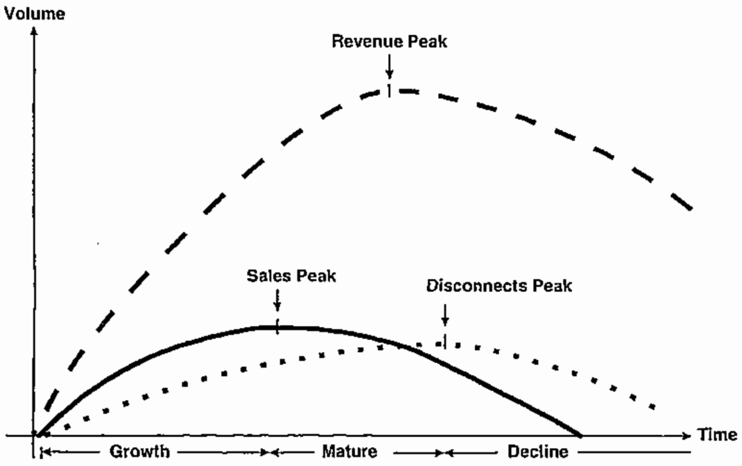

Figure 2. Private line service life cycle.

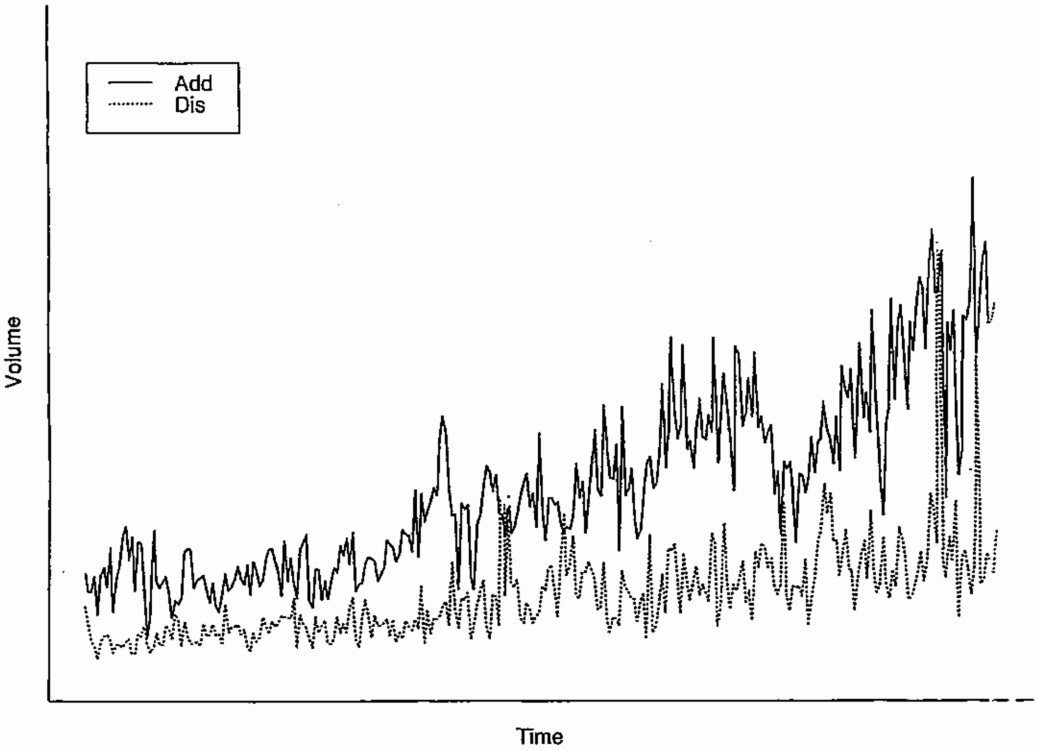

Figure 3. ADD/disconnect volumes for service 1.

the growth phase in which both adds and disconnects are increasing; figure 4 illustrates the decline phase in which both adds and disconnects are decreasing; and figure 5 illustrates the mature phase in which adds are decreasing but disconnects are still increasing. (There may be a transition to the decline phase toward the end of the time interval.)

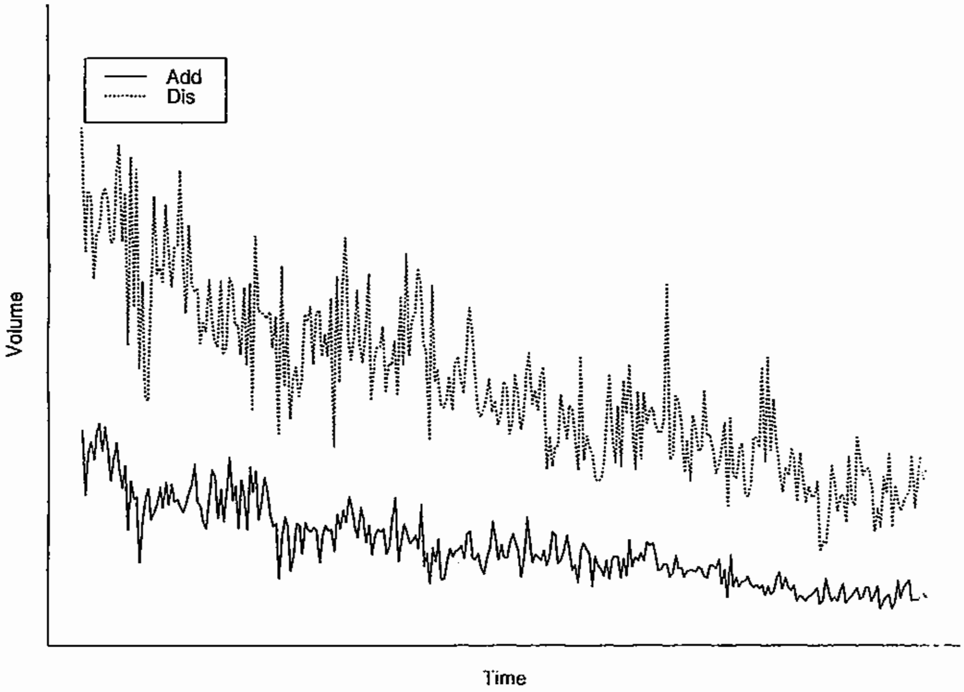

Figure 4. ADD/disconnect volumes for service 2.

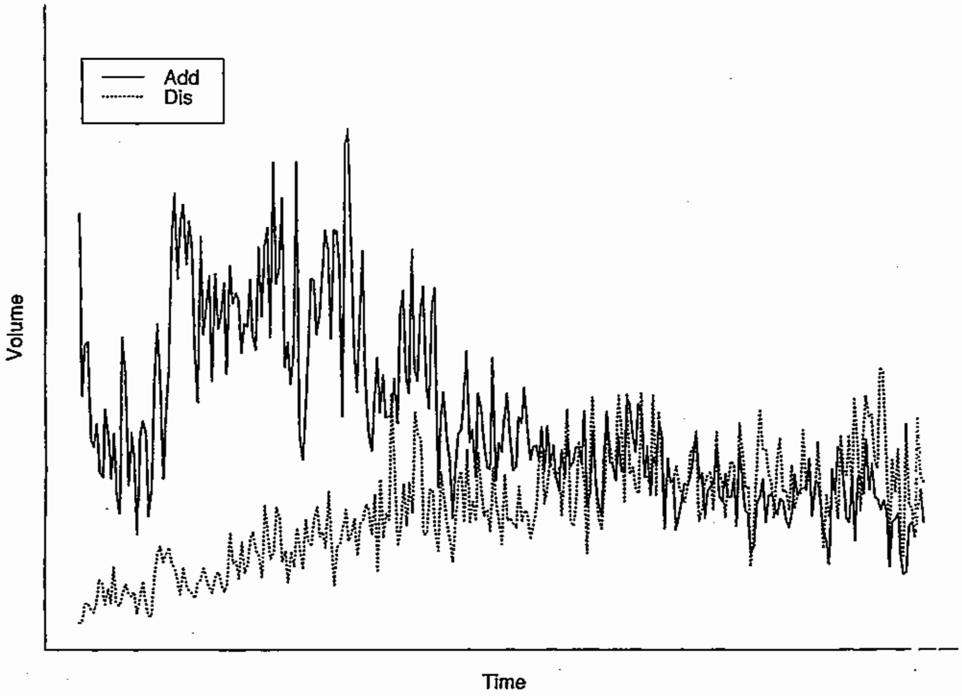

Figure 5. ADD/disconnect volumes for service 3.

5. Model simplification for fitting to data

Even though the single-service model as we have defined it in section 4 is not difficult to analyze using (4.5)–(4.7), it is difficult to apply because it requires specifying complex model primitives. It is natural to directly specify $\lambda^+(t)$ in terms of data on new connections (adds), so that it is not necessary to separately specify the component functions $\gamma(t)$ and $\beta(t)$, but specifying $\lambda^+(t)$ and $G_u(t)$ is difficult. We need to specify the *function* $\lambda^+(t)$ of the one variable t and the *function* $G_u(t)$ of the two variables t and u. It can be daunting to fit such functions to data. Over a long time period, there typically will be only a single observed sample path. We thus propose simplifications that drastically reduce the model fitting requirements.

We first consider the arrival rate function $\lambda^+(t)$. As can be seen from figures 3–5, it is often natural to fit $\lambda^+(t)$ to a linear function $c_0 + c_1 t$. We can divide the time interval into subintervals and count the number of new connections in each subinterval and then fit a linear function by least squares. As a refinement, it is natural to consider the quadratic function $\lambda^+(t) = c_0 + c_1 t + c_2 t^2$. Other alternative simple parametric models are exponential ($\lambda^+(t) = c_0 + c_1 e^{c_2 t}$) and logarithmic ($\lambda^+(t) = c_0 + c_1 \log(c_2 + c_3 t)$); e.g., an exponential form was used by Smith [30]. These specific forms require fitting only a few parameters. Methods for doing such fitting for nonhomogeneous Poisson processes have been studied by Massey et al. [19]. It is important to note that the linear and quadratic functions can make the function $\lambda^+(t)$ negative for some values of t. This usually causes no problem provided that $\lambda^+(t)$ is positive in the region of interest. However, it is good to check by doing numerical examples using (4.5) and (4.6) that no significant errors are introduced by the linear and quadratic approximations. Extensive experience with $M_t/GI/\infty$ queues indicates that this step is usually appropriate [10].

Next we propose two possible simplifications for the service times: (1) we can assume that the service-time cdf does not depend on the arrival time, giving us the function $G(t)$ of the one variable t, or (2) we can assume that there is a time-dependent service rate $\mu(t)$ operative at each time t. When the service rate $\mu(t)$ is constant, this second case is tantamount to having an exponential service-lifetime distribution. In the first case, we let $G(t)$ have a parametric form, such as gamma or Weibull, depending on only a few parameters. In the second case, just as with the arrival-rate function $\lambda^+(t)$, we can assume that $\mu(t)$ is linear or quadratic, so that it too depends on only a few parameters. The single cdf $G(t)$ is desirable if the time dependence is minor, but the non-exponential character of a service lifetime is strong. The time-dependent rate $\mu(t)$ is good if time dependence is strong. In summary, by the simplifications above, we can specify the arrival and service processes each by two or three parameters, which makes model fitting manageable.

The long service lifetimes make it difficult to estimate service-lifetime cdf's. First, provisioning records tend to be kept in the operational environment for only a few months, so that it is necessary to preserve the data in a data warehouse. Second, in order to determine the lifetime and life status of a specific private line, it is necessary to associate add and disconnect data for a specific private line over several years of provisioning records. Third, it is important to account for censoring (incomplete observations) in the statistical analysis, because a significant portion of service lifetimes will have begun before the measurement interval started, and a significant portion of service lifetimes will still be in progress at the end of the measurement interval. Ways to do statistical analysis that properly account for censoring are described in [17].

To quickly see the importance of properly treating censoring, suppose that we have data over 4 years for exponential lifetimes with mean $ES = 5$ years. A simple (incorrect) approach would be to restrict attention to the lifetimes that both started and ended in the 4-year measurement interval. Without doing any calculations, we see that this procedure forces the estimate of the mean to be less than 4, thus underestimating the true mean 5. However, the situation is actually much worse than that, because many lifetimes will start toward the end of the measurement interval. The expected value of this estimated mean is actually

$$\frac{\int_0^4 E[S \mid S \leqslant t]\lambda^+(t)\,\mathrm{d}t}{\int_0^4 \lambda^+(t)\,\mathrm{d}t}, \tag{5.1}$$

where $\lambda^+(t)$ is the arrival-rate function. Since

$$E[S \mid S \leqslant t] = \frac{\int_0^t u\mathrm{e}^{-u/5}\,\mathrm{d}u}{1 - \mathrm{e}^{-t/5}} < \frac{t}{2}, \tag{5.2}$$

$E[S \mid S \leqslant t] \approx t/2$ for $0 \leqslant t \leqslant 4$. Hence, if arrival rate $\lambda^+(t)$ is constant over the time interval $[0, 4]$, then the estimated mean is approximately 1, which greatly underestimates the true mean 5.

However, we can reasonably estimate the complementary cdf (ccdf or survival function) $G^{\mathrm{c}}(t) \equiv 1 - G(t)$ for $0 \leqslant t \leqslant 4$ as the proportion of those service times starting in $[0, 4 - t]$ whose lifetimes exceed t. This procedure yields an unbiased estimate for the cdf $G(t)$ over $0 \leqslant t \leqslant 4$, but no estimate for larger t. (The estimate will also tend not to be so good when t is near 4, because then there are relatively few observations.) Moreover, the estimate for each t is the estimate of the mean of a Bernoulli random variable from an i.i.d. sample, so that we know the variance. In particular, with n data points, the estimate of $G^{\mathrm{c}}(t)$ has variance $G^{\mathrm{c}}(t)G(t)/n$.

We can obtain an estimate for larger t by fitting a convenient parametric cdf, such as a two-parameter Weibull cdf, to the fitted cdf over $[0, 4]$, with the understanding that the tail ($G^{\mathrm{c}}(t)$ for $t > 4$) is obtained as an extrapolation with less justification.

In addition to making model fitting manageable, the model simplification also make it easier to analyze the model and deduce important managerial insights. First, if we use the time-dependent service rate $\mu(t)$, then we can use a single ordinary differential equation (ODE) to solve for $m(t)$ and $\lambda^-(t)$, i.e., letting $\dot{m}(t)$ denote the derivative with respect to time, we have

$$\dot{m}(t) = \lambda^+(t) - m(t)\mu(t) \tag{5.3}$$

and

$$\lambda^{-}(t) = m(t)\mu(t); \tag{5.4}$$

e.g., see [10, corollary 4; 20, section 8]. We can initialize the ODE (5.3) by letting $m(0)$ be the known initial number of lines in service.

Formula (5.4) also helps develop a good way to estimate the service rate $\mu(t)$. It is customary to actually have data on the disconnect rate $\lambda^{-}(t)$ and $m(t)$. We can think of $\mu(t)$ as $\lambda^{-}(t)/m(t)$. We can thus divide the total time interval into subintervals and look at the ratios of the number of disconnects to the average number of lines in service. We then can fit a linear or quadratic function to those counts to estimate $\mu(t)$.

Given a service-lifetime cdf $G(t)$ independent of the arrival time, there are convenient simple forms for the mean $m(t)$ and departure rate $\lambda^{-}(t)$; see [10, section 3]. Let S be a random variable with cdf $G(t)$ and let S_{e} be a random variable with the stationary-excess cdf $G_{\mathrm{e}}(t)$ associated with the service lifetime cdf $G(t)$, i.e.,

$$G_{\mathrm{e}}(t) = P(S_{\mathrm{e}} \leqslant t) = \frac{1}{ES}\int_0^t G^{\mathrm{c}}(u)\,\mathrm{d}u. \tag{5.5}$$

The moments of S_{e} are simply related to the moments of S, i.e.,

$$E\bigl[S_{\mathrm{e}}^k\bigr] = \frac{E[S^{k+1}]}{(k+1)E[S]}, \quad k \geqslant 1. \tag{5.6}$$

Given a time-independent service lifetime S, we can conveniently express $m(t)$ and $\lambda^{-}(t)$ as

$$m(t) = E\bigl[\lambda^{+}(t - S_{\mathrm{e}})\bigr]E[S] \tag{5.7}$$

and

$$\lambda^{-}(t) = E\bigl[\lambda^{+}(t - S)\bigr]; \tag{5.8}$$

see [10]. From (5.7) and (5.8), we obtain convenient simple approximations by moving the expectation inside λ^{+}, i.e.,

$$m(t) \approx \lambda^{+}(t - ES_{\mathrm{e}})E[S] \tag{5.9}$$

and

$$\lambda^{-}(t) \approx \lambda^{+}(t - ES). \tag{5.10}$$

If, in addition, $\lambda^{+}(t)$ is quadratic, then we obtain simple exact formulas. If $\lambda^{+}(t) = c_0 + c_1 t + c_2 t^2$, then

$$m(t) = \lambda^{+}(t - ES_{\mathrm{e}})ES + c_2\,\mathrm{Var}(S_{\mathrm{e}})ES, \tag{5.11}$$

$$\lambda^{-}(t) = \lambda^{+}(t - ES) + c_2\,\mathrm{Var}(S). \tag{5.12}$$

Formula (5.11) implies that there is a *time lag* ES_{e} in $m(t)$ behind $\lambda^{+}(t)ES$ and a *space shift* of $c_2\,\mathrm{Var}(S_{\mathrm{e}})ES$. Similarly, formula (5.12) implies that there is *time lag* of ES in

$\lambda^-(t)$ behind $\lambda^+(t)$ and a *space shift* of $c_2\,\mathrm{Var}(S)$. Note that the third moment of S enters in only through the space shift in $m(t)$, i.e., from (5.6),

$$\mathrm{Var}(S_e) = \frac{4E(S^3)E(S) - 3E(S^2)^2}{12E(S)^2}. \tag{5.13}$$

We emphasize that the simple quadratic formulas (5.11) and (5.12) must be regarded as approximations, because they may depend on arrival-rate functions that are negative for some times. However, these approximations are convenient because they reveal the essential structural form, and they are often sufficiently accurate. Accuracy can be checked by comparing with the exact formulas in section 4.

Using equations (5.7)–(5.12), we would not initialize with the known present number of lines in service. Instead, we would assume that we started in the indefinite past and apply the model with the historical arrival-rate function and service-lifetime cdf. Assuming that there is relatively low variability due to the aggregation over facilities associated with the product-management view, we anticipate that the mean $m(t)$ will agree closely with the observed number of lines in service for past t. When this agreement is close, we will have more faith in the predictions about the future.

The three phases of a service life cycle are conveniently approximated quantitatively when we use the quadratic approximation in (5.11) and (5.12). First, from (5.11) and (5.12), it is evident that $m(t)$ and $\lambda^-(t)$ inherit the quadratic form of $\lambda^+(t)$, and are unimodal with a unique maximum. Then the peak of $m(t)$ lags behind the peak of $\lambda^+(t)$ by ES_e, while the peak of $\lambda^-(t)$ lags behind the peak of $\lambda^+(t)$ by ES.

Thus, using the quadratic approximation, the length of the mature period is

$$L_m = \max\{ES, ES_e\} = ES\max\left\{1, \frac{1+c_s^2}{2}\right\} = \begin{cases} \dfrac{ES(1+c_s^2)}{2}, & c_s^2 \geqslant 1, \\ ES, & c_s^2 \leqslant 1, \end{cases} \tag{5.14}$$

where c_s^2 is the squared coefficient of variation (SCV, variance divided by the square of the mean) of a service lifetime.

An important implication is that sales (represented by $\lambda^+(t)$) is a leading indicator of revenue (represented by $m(t)$). Thus, on observing a peak in sales, the product manager can predict a peak in revenue after a time increment of $ES_e = ES(1 + c_s^2)$. During this interval, the revenue and the supporting network resources continue to grow, although sales declines. This time lag is significant (e.g., of order 5 years) because of the long service lifetimes of private-line services. This is in contrast to most economic endeavors in which sales and revenue move in lockstep.

Formula (5.11) shows the sensitivity of the revenue (again represented by $m(t)$) to the mean ES as well as to the distribution of S beyond the mean. The fact that the mean ES appears as a multiplicative factor in (5.11) shows that revenue tends to be directly proportional to the mean, as one would expect. However, (5.11) shows that there is an additional *time shift* by ES_e and *space shift* by $\mathrm{Var}(S_e)$. In general (without assuming $\lambda^+(t)$ is quadratic), [10, theorem 3] implies that when the variability of the service-lifetime cdf decreases (as measured by convex stochastic order), the expected revenue

$m(t)$ increases when sales is increasing (in the growth phase). The revenue is more responsive to changes in sales when the variability of the service-lifetime distribution is lower; see [8].

The dependence of revenue upon the service-lifetime cdf can help product managers decide how to influence the service-lifetime cdf. Currently all leading interexchange service providers (AT&T, MCI and Sprint) influence service-lifetime cdfs by offering discounts that increase with the length of the committed lifetime; see [5].

6. Approximating the time-dependent distribution

In formulas (4.5), (5.3) and (5.7), we have provided expressions for $m(t) \equiv EQ(t)$, the time-dependent mean number of lines in service. We anticipate that the time-dependent probability distribution will cluster closely about the mean because of the aggregation over facilities associated with the product-management view, but now we want to provide a way to estimate the variability. In this section we develop approximations for the variance and the full time-dependent probability distribution of $Q(t)$. Our main idea is to act as if the connecting lines in an arriving batch stay in service together and depart (disconnect) together. This allows us to focus on an arrival process of batches with a certain batch-size distribution. We also must make more explicit probabilistic assumptions.

As approximations, we assume, first, that the connection-request (of batches) stochastic process $\{A(t)\colon t > -\infty\}$ is a nonhomogeneous *Poisson process* with rate $\gamma(t)$ and, second, that each request is for a random number of lines, where successive requests are i.i.d. and distributed as B with a cdf F having mean β independent of the request arrival time. Under these assumptions, the number of batches in service behaves as an $M_t/GI/\infty$ queue. If we use a time-dependent service-lifetime cdf $G_u(t)$ for each batch, then the number of batches in service at time t has a Poisson distribution with mean

$$\nu(t) = \int_{-\infty}^{t} G_u^{c}(t-u)\gamma(u)\,\mathrm{d}u. \tag{6.1}$$

Formula (6.1) is of the same form as (4.5). If we use a service-rate function $\mu(t)$, then the number of batches in service at time t has a Poisson distribution with mean $\nu(t)$, where $\nu(t)$ satisfies the ODE

$$\dot{\nu}(t) = \gamma(t) - \nu(t)\mu(t), \tag{6.2}$$

just as in (5.3). Then the number of lines in service at time t has a batch-Poisson (or compound-Poisson) distribution with Poisson mean $\nu(t)$ and batch-size cdf F. With this framework, we need to fit $\gamma(t)$ and F to data instead of just $\lambda^{+}(t)$. We can fit $\gamma(t)$ to connection (batch) request data the same way we fit $\lambda^{+}(t)$. We can fit F by looking at the empirical distribution of actual arriving batch sizes.

These fitting procedures are straightforward, but a complication comes from the fact that the approximation assumptions are not likely to be closely satisfied, so that we may find it desirable to adjust the estimate of the batch-size distribution. To do so, we can exploit our more precise knowledge of the time-dependent mean number of lines in service, $m(t)$, obtained by the methods of sections 4 and 5. Within our batch-Poisson model, we can express $m(t)$ as

$$m(t) = \nu(t)\beta, \tag{6.3}$$

where $\beta \equiv EB$ is the mean batch size. If the three separate estimates of $m(t)$, $\nu(t)$ and β do not satisfy (6.3), then we can *redefine* β to be $\beta(t) \equiv m(t)/\nu(t)$. In general, this will cause β to be replaced by the time-dependent $\beta(t)$, but that presents no difficulty.

We next apply the adjustment to the entire batch-size distribution. We let the batch-size distribution of the active batches at time t be distributed as $B_t \equiv m(t)B/\beta\nu(t)$, i.e., with time-dependent cdf

$$F_t(x) = F\big(x\beta\nu(t)/m(t)\big), \quad x \geqslant 0. \tag{6.4}$$

This adjustment simply multiplies B by a time-dependent function so that the time-dependent mean is $\beta(t) = m(t)/\nu(t)$. Thus the SCV of the cdf F_t is the SCV of F for all t, say c_b^2.

After the adjustment, the number of lines in service at time t is approximated by a batch-Poisson distribution, where the Poisson mean is $\nu(t)$ and the batch-size cdf is $F_t(x)$ in (6.4) with mean $\beta(t)$. The overall mean number of lines in service at time t is then $m(t)$, as previously determined in sections 4 and 5. We can then proceed to calculate the variance and cdf of $Q(t)$, based on this approximation, using standard properties of batch-Poisson distributions, see [11, chapter XII]. First, the variance is

$$\begin{aligned}\sigma^2(t) \equiv \operatorname{Var} Q(t) = \nu(t)E\big[B_t^2\big] &= \nu(t)E\Big[\big(m(t)B/\beta\nu(t)\big)^2\Big]\\ &= \frac{m(t)^2E[B^2]}{\beta^2\nu(t)} = \frac{m(t)^2(c_b^2+1)}{\nu(t)},\end{aligned} \tag{6.5}$$

where c_b^2 is the SCV of B and thus also B_t for all t.

Next, suppose that the batch size B_t with cdf F_t has Laplace–Stieltjes transform (LST)

$$\hat{f}_t(s) \equiv E\mathrm{e}^{-sB_t} \equiv \int_0^\infty \mathrm{e}^{-sx}\,\mathrm{d}F_t(x). \tag{6.6}$$

Then $Q(t)$ has the associated LST

$$\hat{q}_t(s) \equiv E\mathrm{e}^{-sQ(t)} = \mathrm{e}^{-\nu(t)(1-\hat{f}_t(s))}. \tag{6.7}$$

The associated complementary cdf $H_t^c(x) \equiv P(Q(t) > x)$ has Laplace transform

$$\widehat{H}_t^c(s) \equiv \int_0^\infty \mathrm{e}^{-sx}H_t^c(x)\,\mathrm{d}x = \frac{1-\hat{q}_t(s)}{s}. \tag{6.8}$$

Similarly, if B_t is integer-valued with generating function

$$f_t^*(z) \equiv Ez^{B_t} \equiv \sum_{n=0}^{\infty} z^n P(B_t = n), \tag{6.9}$$

then $Q(t)$ is also integer-valued and its probability mass function $P(Q(t) = n)$ has generating function

$$h^*(z) \equiv Ez^{Q(t)} = \mathrm{e}^{-\nu(t)(1-f^*(z))}. \tag{6.10}$$

Since B_t will typically be integer-valued, so will $Q(t)$. Then its probability mass function can be computed easily from its generating function (6.10) by numerical inversion; e.g., see [1]. In this way we obtain an estimate of the full time-dependent distribution of $Q(t)$.

If the Poisson mean $\nu(t)$ is not too small and the batch-size cdf $F_t(x)$ is not too irregular, it is natural to approximate the distribution of $Q(t)$ by a normal distribution with mean $m(t)$ and variance $\sigma^2(t)$ in (6.5). The normal approximation can be checked by making comparisons with the numerically calculated distribution.

With these approximations, we can check that $Q(t)$ should indeed be close to its mean $m(t)$. Since typical values of $Q(t)$ are 10^5, there is no problem when the batches are suitably small. To quickly illustrate, if $m(t) = 10^5$ and all batch sizes are 10, then $SD(Q(t))/EQ(t) = 0.01$. However reliable prediction can be difficult if there are very rare, very large batches.

7. A family of private-line services

The methods in sections 4–6 can be applied to any single private-line service. Now we want to extend the analysis to a family of private-line services. The main phenomenon we are trying to capture is switching from one service to another. Suppose that there are N services and let i index the service. For each service i, there still is the connection arrival-rate function $\lambda_i^+(t)$, but now we want to distinguish between exogenous requests and switching from one service to another. Let $\alpha_i(t)$ denote the exogenous arrival rate of new connection requests for service i and let $p_{ij}(t)$ be the proportion of disconnects from service i that are followed by connects to service j (switching from i to j), which we also interpret as the probability of switching, assuming Markovian switching. Paralleling section 4, we also assume that the successive batch-size and service-lifetimes depend upon time and the service, but otherwise are mutually independent and independent of the system history. Then the total arrival rate (line connection requests) for service can be expressed as

$$\lambda_i^+(t) = \alpha_i(t) + \sum_{j=1}^{N} \lambda_j^-(t) p_{ji}(t), \tag{7.1}$$

where $\lambda_j^-(t)$ is the line disconnect-rate function for service j. The disconnect flow rate from service i (without switching to another service) is

$$\delta_i(t) = \lambda_i^-(t)\left(1 - \sum_{j=1}^{N} p_{ij}(t)\right). \tag{7.2}$$

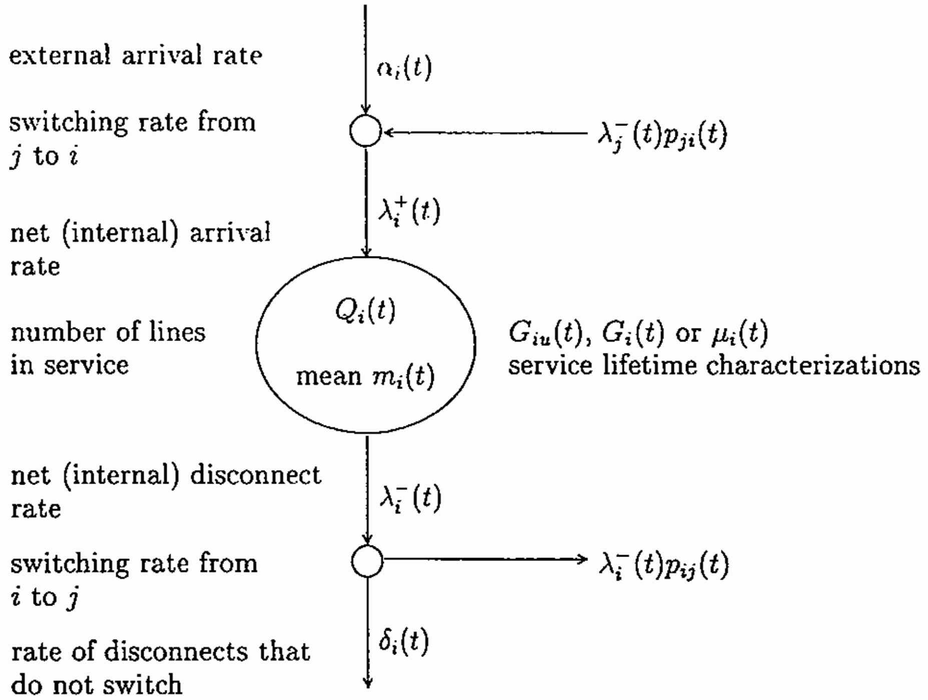
external arrival rate
$\alpha_i(t)$
switching rate from j to i
$\lambda_j^-(t)p_{ji}(t)$
$\lambda_i^+(t)$
net (internal) arrival rate
$Q_i(t)$
number of lines in service
mean $m_i(t)$
$G_{iu}(t)$, $G_i(t)$ or $\mu_i(t)$ service lifetime characterizations
net (internal) disconnect rate
$\lambda_i^-(t)$
switching rate from i to j
$\lambda_i^-(t)p_{ij}(t)$
$\delta_i(t)$
rate of disconnects that do not switch

Figure 6. Flow through service i for a family of private-line services.

We will show how to determine $\lambda_i^+(t)$ and $\lambda_i^-(t)$ below. The behavior of service i within the family is described by the flow diagram in figure 6.

The sales process should have information to distinguish between new connections (to be used for estimating $\alpha_i(t)$) and switched connections (to be using for estimating $p_{ij}(t)$). However, this information may not be recorded in an operational database. In that case, it is possible to deduce this information from the provisioning (adds/disconnects) records. New connections of one service within a suitably short time interval (e.g., one month) of disconnects of another service can be designated as switches. Marketing information may be used to predict switching behavior in the future. The analysis can be useful to answer "what if" questions about the life-cycle behavior under various possible switching scenarios. Then the future rates $\alpha_i(t)$ and $p_{ij}(t)$ may be postulated as part of the scenario description.

Just as in section 5, there is a need for model simplification in order to fit the functions $\alpha_i(t)$ and $p_{ij}(t)$ to data. Just as before, we suggest simple parametric forms that reduce estimation to two or three parameters for each function. In particular, we propose letting $\alpha_i(t)$ be linear or quadratic and $p_{ij}(t)$ be either constant or linear, but other possibilities can be analyzed.

Given that $\alpha_i(t)$ and $p_{ij}(t)$ have been estimated, there is an important *exact-decomposition property* that simplifies the overall analysis. Each queue can be analyzed separately after the internal arrival rates $\lambda_i^+(t)$ are determined; i.e., we can separate the analysis into two steps:

(1) determining the net "internal" arrival rates $\lambda_i^+(t)$ and

(2) analyzing the behavior of each service i given $\lambda_i^+(t)$.

The important point is that once we determine $\lambda_i^+(t)$ we can analyze each service i by the methods of sections 4–6.

We can calculate $\lambda_i^+(t)$ for all i by the methods [20, section 7] for the $(M_t/GI/\infty)^N/M_t$ network, which has independent nonhomogeneous Poisson processes as external arrival processes for the services and time-dependent Markovian routing. For the purpose of calculating the internal arrival rates $\lambda_i^+(t)$, the Poisson-arrival-process assumption is not needed.

First, suppose that each line of service i has a time-dependent service-lifetime cdf $G_{iu}(t)$. Paralleling (4.6), the disconnect-rate function should satisfy

$$\lambda_j^-(t) = \int_{-\infty}^{t} G_{ju}(t-u)\lambda_j^+(u)\,\mathrm{d}u. \tag{7.3}$$

Moreover, [20, theorem 7.2] shows that $\lambda_i^+(t)$ should be regarded as the minimal nonnegative solution to (7.1) and (7.3); i.e., in general the solution is not necessarily unique. The combination of equations (7.1) and (7.3) constitute the *input equations*.

We can conveniently calculate $\lambda_i^+(t)$ recursively by keeping count of transitions. For that purpose, let $\lambda_i^n(t)$ be the arrival rate that occurs on the nth transition. Then $\lambda_i^+(t)$ can be calculated as the infinite sum

$$\lambda_i^+(t) = \sum_{n=1}^{\infty} \lambda_i^n(t), \tag{7.4}$$

where λ_i^n is calculated recursively by setting $\lambda_i^1(t) = \alpha_i(t)$, and for all $n \geqslant 1$

$$\lambda_i^{n+1}(t) = \sum_{j=1}^{N} \int_{-\infty}^{t} \lambda_j^n(u)\,\mathrm{d}G_{ju}(t-u)p_{ji}(t). \tag{7.5}$$

If customers tend not to switch too often before terminating service altogether and leaving the system, which is often the case, then the series in (7.4) will converge rapidly and can be approximated by truncating to a finite sum with a few terms.

If instead of service-lifetime cdfs $G_{iu}(t)$, we postulate time-dependent service rates $\mu_i(t)$, then we can solve for the means $m_i(t)$ by solving the system of N ODEs

$$\dot{m}_i(t) = \alpha_i(t) + \sum_{j=1}^{N} m_j(t)\mu_j(t)p_{ji}(t) - m_i(t)\mu_i(t), \quad 1 \leqslant i \leqslant N. \tag{7.6}$$

Then

$$\lambda_i^-(t) = m_i(t)\mu_i(t) \tag{7.7}$$

and $\lambda_i^+(t)$ can be computed via (7.1). (In this case, we do not afterwards have to calculate $m_i(t)$.)

In order to calculate the distribution of $Q_i(t)$ for each i, as in section 6, we need to determine the internal batch arrival rates $\gamma_i^+(t)$, so that we can calculate the mean number of batches in service $\nu_i(t)$ via (6.1). We can obtain $\gamma_i^+(t)$ just as we obtained $\lambda_i^+(t)$ via (7.1)–(7.5) if we replace $\alpha_i(t)$ there by the exogenous batch arrival rate. Similarly, if we work with service-rate functions $\mu_i(t)$, we can obtain $\nu_i(t)$, $\gamma_i^+(t)$ and $\gamma_i^-(t)$ via (7.1), (7.6) and (7.7) if we replace $\alpha_i(t)$ by the exogenous batch arrival rate. The remaining steps are just as in section 6.

The exact-decomposition property has important implications for the organizational structure of product management. Given that the internal new-connection rates $\lambda_i^+(t)$ for each service can be estimated, it is possible to manage the different services separately. The equations determining the internal rates $\lambda_i^+(t)$ serve a coordinating role. However, it should be noted that there remains some interdependence; e.g., if the service-lifetime cdf of one service is altered, then the internal arrival rates of other services will be altered as well.

8. Time-independent service lifetime CDFs and transition probabilities

In sections 5 and 7 we saw that model fitting to data requires us to introduce simplifying parametric model assumptions. As a consequence, we assume that the external arrival-rate functions $\alpha_i(t)$ and the switching proportions (or probabilities) $p_{ij}(t)$ be low-order polynomials. In this section and the following one, we investigate the consequence of this assumption on the internal arrival rate functions $\lambda_i^+(t)$ and the time-dependent means $m_i(t)$. For this analysis, we assume that the service-lifetime cdfs are independent of the arrival time. (No special analysis is needed for the service-rate case, because then we can directly apply (7.6).)

In this section we consider the simplest case, in which $\alpha_i(t)$ is a polynomial for all i but $p_{ij}(t) \equiv p_{ij}$ independent of t for all i and j. First, by a minor extension of theorems 6.2 and 7.4 of [20], if $\alpha_i(t)$ is a polynomial of degree at most m for all i and $p_{ij}(t)$ is independent of t, then $\lambda_i^+(t)$ itself is a polynomial of degree at most m. The additional extension is to allow the external arrival processes to be different polynomials of degree at most m, instead of constant multiples of an single polynomial. It is easy to see that the same proof applies to this more general case.

Moreover, it is then easy to solve for the coefficients. If $\alpha_i(t) = a_{i0} + a_{i1}t + \cdots + a_{im}t^m$, $1 \leqslant i \leqslant N$, then $\lambda_i^+(t) = c_{i0} + c_{i1}t + \cdots + c_{im}t^m$, $1 \leqslant i \leqslant N$, where the coefficients c_{ij} can be calculated directly. In particular, let P be the $N \times N$ matrix with elements p_{ij} and let C_k be the $1 \times N$ row vector $(c_{1k}, \ldots, c_{Nk})$. Then

$$C_k = B_k(I - P)^{-1}, \tag{8.1}$$

where $B_k = (b_{1k}, \ldots, b_{Nk})$ and

$$b_{ik} = a_{ik} + \sum_{l=k+1}^{m} \sum_{j=1}^{N} c_{jl}(-1)^{l-k} \binom{l}{k} E\left[S_j^{l-k}\right] p_{ji}. \tag{8.2}$$

We solve for b_{ik} and c_{ik} recursively backwards in k, starting with $k = m$. From (8.1), we see that $b_{im} = a_{im}$, $1 \leqslant i \leqslant N$. When we calculate b_{ik} by (8.2), we will have calculated c_{jl} for $l > k$ by using (8.1) and (8.2) in previous steps of the recursion.

Note from (8.2) that the general service-time cdfs G_j influence the net arrival rates $\lambda_i^+(t)$ only through their first m moments. From (4.1), we see that the distributions G_j can thus influence the means $m_i(t)$ only through the first m moments of S_i and S_{ie}, i.e., through the first $m+1$ moments of S_i, see (5.6). Similarly, since $\lambda_i^-(t) = E[\lambda_i^+(t - S_i)]$, the service-time distributions G_j can influence the behavior of the net departure-rate functions $\lambda_i^-(t)$ only through the first m moments of G_j. Thus, if the external arrival-rate functions are quadratic, only the first three service-time moments matter.

Since the service-lifetime cdf is time-independent, the basic formulas for $m(t)$ in (4.5) and $\lambda^-(t)$ in (4.6) are convolution integrals. Hence we can exploit Laplace transforms to advantage. Moreover, since the switching probabilities are also time-independent we extend the Laplace transform analysis to the network setting. We are then able to compute the desired quantities by numerically inverting the Laplace transforms, e.g., by applying [2]. To give the details, let $\hat{h}(s)$ denote the Laplace transform of $h(t)$ and the Laplace–Stieltjes transform of $H(t)$ when $h(t)$ is the density of a cdf H, i.e.,

$$\hat{h}(s) = \int_0^\infty \mathrm{e}^{-st} h(t)\,\mathrm{d}t = \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}H(t). \tag{8.3}$$

Then, by (4.5) and (4.6),

$$\widehat{m}_i(s) = \hat{\lambda}_i^+(s)\widehat{G}_i^{\mathrm{c}}(s) \quad \text{and} \quad \hat{\lambda}_i^-(s) = \hat{\lambda}_i^+(s)\hat{g}_i(s). \tag{8.4}$$

Moreover, if $p_{ij}(t)$ is independent of time for all i and j, then taking Laplace transforms in (7.1) and (7.3) yields

$$\hat{\lambda}_j^+(s) = \hat{\alpha}_j(s) + \sum_{i=1}^{N} \hat{\lambda}_i^+(s)\hat{g}_i(s)p_{ij} \tag{8.5}$$

or, in matrix notation,

$$\hat{\Lambda}^+(s) = \hat{A}(s)\big(I - B(s)\big)^{-1} = \frac{\hat{A}(s)\mathrm{Adj}(I - B(s))}{\det(I - B(s))}, \tag{8.6}$$

where $\hat{\Lambda}^+(s) = (\lambda_1^+(s), \ldots, \lambda_N^+(s))$, $\hat{A}(s) = (\hat{\alpha}_1(s), \ldots, \hat{\alpha}_N(s))$ and $\widehat{B}(s) = (\hat{b}_{ij}(s))$ with $\hat{b}_{ij}(s) = \hat{g}_i(s)p_{ij}$. We can then combine (8.4) and (8.6) to obtain expressions for the Laplace transforms of $m_i(t)$ and $\lambda_i^-(t)$ in the network setting.

Example 8.1. We illustrate the results by considering a two-service example, in which both services have quadratic external arrival-rate functions with finite positive maxima. The external arrival rate for product i is $\alpha_i(t) = a_{i0} + a_{i1}t + a_{i2}t^2$ for some specified constants a_{ij}, $0 \leqslant j \leqslant 2$. The general theory then tells us that the net arrival rate for product i is also a quadratic function of the form $\lambda_i^+ = c_{i0} + c_{i1}t + c_{i2}t^2$, where the parameters c_{ij} are to be determined. In particular, we have $C_k = B_k(I - P)^{-1}$ for $C_k = (c_{1k}, c_{2k}), k = 0, 1, 2$.

With switching matrix

$$P = \begin{pmatrix} 0 & p_{12} \\ p_{21} & 0 \end{pmatrix}, \qquad (I - P)^{-1} = \frac{1}{1 - p_{12}p_{21}} \begin{pmatrix} 1 & p_{12} \\ p_{21} & 1 \end{pmatrix}. \tag{8.7}$$

We solve for the vectors B_k and C_k recursively, in decreasing order. First, $b_{i2} = a_{i2}$, so that

$$(c_{12}, c_{22}) = (a_{12}, a_{22}) \frac{1}{(1 - p_{12}p_{21})} \begin{pmatrix} 1 & p_{12} \\ p_{21} & 1 \end{pmatrix}, \tag{8.8}$$

i.e.,

$$c_{12} = \frac{a_{12} + a_{22}p_{21}}{1 - p_{12}p_{21}}, \qquad c_{22} = \frac{a_{12}p_{12} + a_{22}}{1 - p_{12}p_{21}}. \tag{8.9}$$

Note that c_{12} and c_{22} are always negative when a_{12} and a_{22} are negative, so that the net arrival-rate functions $\lambda_1^+(t)$ and $\lambda_2^+(t)$ inherit the finite-positive-maximum property of the external arrival-rate functions $\alpha_1(t)$ and $\alpha_2(t)$.

Next, given c_{12} and c_{22},

$$b_{i1} = a_{i1} - 2E(S_1)c_{12}p_{1i} - 2E(S_2)c_{22}p_{2i}. \tag{8.10}$$

Finally, given c_{11}, c_{12}, c_{21} and c_{22},

$$b_{i0} = a_{i0} + \big(E\big(S_2^2\big)c_{22} - 2E(S_2)c_{21}\big)p_{2i} + \big(E\big(S_1^2\big)c_{12} - 2E(S_1)c_{11}\big)p_{1i}. \tag{8.11}$$

Finally, given $\lambda_i^+(t)$ for $i = 1$ and 2, we can calculate $m_i(t)$ and $\lambda_i^-(t)$ by (5.11) and (5.12). With the new parametric representation, it is easy to see whether we are initially in a growth phase or a declining phase; i.e., $\alpha'(0) = d_0(d_2 - d_1)$, so that $\alpha'(0) > 0$ if and only if $d_2 \geqslant d_1$.

As a concrete example, we consider the two extremal arrival-rate functions

$$\begin{aligned} \alpha_1(t) &= 150 - (t - 5)^2 = 125 + 10t - t^2, \\ \alpha_2(t) &= 100 - 0.5(t + 5)^2 = 87.5 - 5t - 0.5t^2. \end{aligned} \tag{8.12}$$

The first function $\alpha_1(t)$ has its peak at $t_{\max}^1 = 5$ and so is initially in a growth period ($\alpha_1'(0) > 0$), while the second function $\alpha_2(t)$ has its peak at $t_{\max}^2 = -5$ and so is initially declining ($\alpha_2'(0) < 0$). The first external arrival-rate function $\alpha_1(t)$ is positive on the interval $(-7.25, 17.2)$, while the second $\alpha_2(t)$ is positive on the interval $(-19.14, 9.14)$.

We set the eight remaining parameters as follows: $p_{12} = 3/4$, $p_{21} = 1/2$, $E(S_1) = 2$, $E(S_1^2) = 8$, $E(S_1^3) = 48$, $E(S_2) = 1$, $E(S_2^2) = 2$, $E(S_2^3) = 6$. The lifetime moments are consistent with exponential distributions having means 2 and 1, respectively. (Hence, this case could also be solved by ODEs as in (7.6).)

From (8.7)–(8.11), we obtain the solution:

$$(I-P)^{-1} = \begin{pmatrix} \frac{8}{5} & \frac{6}{5} \\ \frac{4}{5} & \frac{8}{5} \end{pmatrix}, \qquad (c_{12}, c_{22}) = (-2.0, -2.0),$$

$$(b_{11}, b_{21}) = (8.0, 1.0), \qquad (c_{11}, c_{21}) = (13.60, 11.20),$$

$$(b_{10}, b_{20}) = (111.60, 35.25), \qquad (c_{10}, c_{20}) = (206.76, 190.32),$$

so that

$$\begin{aligned} \lambda_1^+(t) &= 206.76 + 13.6t - 2.0t^2, \qquad t_{\max}^{1+} = 3.4, \\ \lambda_2^+(t) &= 190.32 + 11.2t - 2.0t^2, \qquad t_{\max}^{2+} = 2.8. \end{aligned} \tag{8.13}$$

The switching makes the net arrival rate substantially larger than the external arrival rate, i.e., $c_{i0} > a_{i0}$ for each i. In this example the switching also has made the net arrival rate at service 2, $\lambda_2^+(t)$, be initially in a growth phase ($\lambda_2^{+'}(0) > 0$), even though the external arrival rate is declining ($\alpha_2'(0) < 0$).

The means $m_i(t)$ and internal departure rates $\lambda_i^-(t)$ are now easily computed by combining (5.7), (5.8) and (8.13). Note that the time lags in $m_i(t)$ behind $\lambda_i^+(t)$ are $E(S_{1e}) = 2$ and $E(S_{2e}) = 1$, respectively. Since $S_e \overset{d}{=} S$ for exponential variables, $\mathrm{Var}(S_{1e}) = 4$ and $\mathrm{Var}(S_{2e}) = 1$. Hence

$$m_1(t) = 2\lambda_1^+(t-2) - 16 \quad \text{and} \quad m_2(t) = \lambda_2^+(t-1) - 2 \tag{8.14}$$

for $\lambda_1^+(t)$ and $\lambda_2^+(t)$ in (8.13). The peak for $m_1(t)$ occurs at $t_{\max}^{1+} + 2 = 5.4$, while the peak for $m_2(t)$ occur at $t_{\max}^{2+} + 1 = 3.8$. The six functions $\alpha_i(t)$, $\lambda_i^+(t)$ and $m_i(t)$ for $i = 1, 2$ are displayed in figure 7.

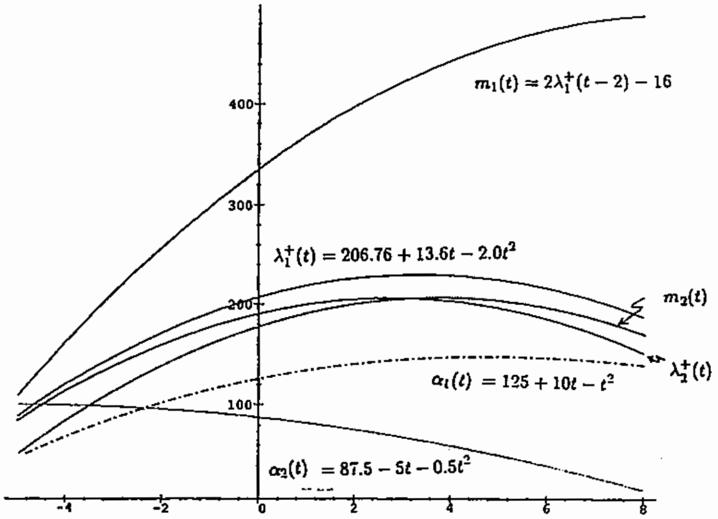

Figure 7. The six quadratic functions.

In this context we can examine various marketing strategies. If we can estimate how these strategies will affect the basic model elements $\alpha_i(t)$, $p_{ij}(t)$ and S_i, then we can apply the model to show the resulting impact upon $\lambda_i^+(t)$, $m_i(t)$ and $\lambda_i^-(t)$. Finally, numerical integration can be performed to show that the explicit formulas here are close to the exact numerical formulas obtained when $\alpha_i(t)$ is replaced by 0 when it is negative.

9. Polynomial arrival rates and switching probabilities

In this section we allow the switching probabilities $p_{ij}(t)$ as well as the external arrival rates $\alpha_i(t)$ to be polynomial functions of t. This case is considerably more complicated than the case of constant switching probabilities considered in section 8, because the net arrival-rate functions $\lambda_i^+(t)$ no longer can be represented exactly as polynomials. However, we can still represent $\lambda_i^+(t)$ as the series $\sum_{n=1}^{\infty} \hat{\lambda}_i^n(t)$ as in (7.4) for $\lambda_i^{n+1}(t)$ in (7.5). The nice property now is that the terms $\hat{\lambda}_i^n(t)$ in this series are polynomial for each n. Unfortunately, though, the degree of the polynomial $\hat{\lambda}_i^n(t)$ increases with n. In particular, by induction with (7.2), if $p_{ij}(t)$ is a polynomial of degree at most r for all i and j, and if $\alpha_i(t)$ is a polynomial of degree at most m for all i, then $\hat{\lambda}_i^n(t)$ is a polynomial of degree at most $m + (n-1)r$ for all i and n.

However, if the series converges rapidly, so that we can consider only a few terms in the series, then $\lambda_i^+(t)$ will be approximated by a polynomial of low degree. For example, an insightful case is when $m = r = 1$ and only $n = 2$ terms are considered. In this case $\alpha_i(t)$ and $p_{ij}(t)$ are linear, while the approximation for $\lambda_i^+(t)$ is quadratic. We consider a numerical example with that structure below.

Now we indicate how to recursively calculate the coefficients of

$$\hat{\lambda}_i^n(t) = \sum_{k=0}^{m+(n-1)r} c_{ik}^{(n)} t^k,$$

where $\alpha_i(t)$ is a polynomial of degree at most m and $p_{ij}(t)$ is a polynomial of degree at most r. To express the result, let

$$\bar{\lambda}_i^n(t) = E\big[\hat{\lambda}_i^n(t - S_i)\big] = \sum_{k=0}^{\infty} d_{jk}^{(n)} t^k \quad \text{and} \quad p_{ij}(t) = \sum_{k=0}^{r} p_{ijk} t^k. \tag{9.1}$$

Then, by (7.2), we can solve for $c_{ik}^{(n)}$ and $d_{jk}^{(n)}$ recursively by setting $c_{ik}^{(1)} = a_{ik}$,

$$d_{jk}^{(n)} = \sum_{l=k}^{m+(n-1)r} c_{jl}^{(n)} \binom{l}{k} (-1)^{l-k} E\big(S_j^{l-k}\big) \tag{9.2}$$

and

$$c_{ik}^{(n+1)} = \sum_{j=1}^{N} \sum_{l=0}^{k} d_{jl}^{(n)} p_{ji(k-l)}, \tag{9.3}$$

where $p_{ijk} = 0$ for $k > r$.

Example 9.1. We illustrate by considering a two-service example, in which the external arrival rates $\alpha_i(t)$ and switching probabilities are all linear. We assume that the switching probabilities are sufficiently small that it suffices to consider only two or three terms $\hat{\lambda}_i^n(t)$, i.e., we assume that $\lambda_i^+(t) \approx \sum_{n=1}^{3} \hat{\lambda}_i^n(t)$.

We now show what the recursion in (9.2) and (9.3) becomes in this simple case. Here are the formulas for the relevant coefficients:

$$c_{i0}^{(1)} = a_{i0}, \qquad c_{i1}^{(1)} = a_{i1}, \qquad d_{j0}^{(1)} = c_{j0}^{(1)} - c_{j1}^{(1)} E(S_j), \qquad d_{j1}^{(1)} = c_{j1}^{(1)},$$

$$c_{i0}^{(2)} = d_{i0}^{(1)} p_{1i0} + d_{20}^{(1)} p_{2i0}, \qquad c_{i1}^{(2)} = d_{10}^{(1)} p_{1i1} + d_{11}^{(1)} p_{1i0} + d_{20}^{(1)} p_{2i1} + d_{21}^{(1)} p_{2i0},$$

$$c_{i2}^{(2)} = d_{11}^{(1)} p_{1i1} + d_{21}^{(1)} p_{2i1}, \qquad d_{j0}^{(2)} = c_{j0}^{(2)} - c_{j1}^{(2)} E(S_j) + 2c_{j2}^{(2)} E(S_j^2),$$

$$d_{j1}^{(2)} = c_{j1}^{(2)} - 2c_{j2}^{(2)} E(S_j), \qquad d_{j2}^{(2)} = c_{j2}^{(2)},$$

$$c_{i0}^{(3)} = d_{10}^{(2)} p_{1i0} + d_{20}^{(2)} p_{2i0}, \qquad c_{i1}^{(3)} = d_{10}^{(2)} p_{1i1} + d_{11}^{(2)} p_{1i0} + d_{20}^{(2)} p_{2i1} + d_{21}^{(2)} p_{2i0},$$

$$c_{i2}^{(3)} = d_{11}^{(2)} p_{1i1} + d_{12}^{(2)} p_{1i0} + d_{21}^{(2)} p_{2i1} + d_{22}^{(2)} p_{2i0}, \qquad c_{i3}^{(3)} = d_{12}^{(2)} p_{1i1} + d_{22}^{(2)} p_{2i1}.$$

Note that, just as in example 8.1, only the first two moments of S_j affect $\hat{\lambda}_i^1$, $\hat{\lambda}_i^2$ and $\hat{\lambda}_i^3$. However, higher moments play a role in the $\hat{\lambda}_i^n$ for $n > 3$. To consider a concrete example, let

$$p_{12}(t) = p_{21}(t) = 0.3 - 0.05t, \qquad \alpha_1(t) = 100 + 10t; \qquad \alpha_2(t) = 100 - 10t,$$

$E(S_i) = 1$, $E(S_i^2) = 2$, $E(S_i^3) = 6$ for $i = 1, 2$. Then

$$c_{10}^{(1)} = a_{10} = c_{20}^{(1)} = a_{20} = 100, \quad c_{11}^{(1)} = a_{11} = 10, \quad c_{21}^{(1)} = a_{21} = -10,$$
$$d_{10}^{(1)} = 90, \quad d_{20}^{(1)} = 110, \quad d_{11}^{(1)} = 10, \quad d_{21}^{(1)} = -10,$$
$$c_{10}^{(2)} = 33, \quad c_{20}^{(2)} = 27, \quad c_{11}^{(2)} = -25, \quad c_{21}^{(2)} = -15,$$
$$c_{12}^{(2)} = 0.5, \quad c_{22}^{(2)} = -0.5,$$
$$d_{10}^{(2)} = 60, \quad d_{20}^{(2)} = 40, \quad d_{11}^{(2)} = -26, \quad d_{21}^{(2)} = -14,$$
$$d_{12}^{(2)} = 0.5, \quad d_{22}^{(2)} = -0.5, \quad c_{10}^{(3)} = 12, \quad c_{20}^{(3)} = 18, \quad c_{11}^{(3)} = -6.2,$$
$$c_{21}^{(3)} = -10.8, \quad c_{12}^{(3)} = 0.55, \quad c_{22}^{(3)} = 1.45, \quad c_{13}^{(3)} = 0.025, \quad c_{23}^{(3)} = -0.025.$$

Thus,

$$\lambda_1^+(t) \approx \sum_{n=1}^{3} \hat{\lambda}_1^n(t) = 145 - 21.2t + 1.05t^2 + 0.025t^3 \approx 145 - 21.2t + 1.05t^2, \quad (9.4)$$

$$\lambda_2^+(t) \approx \sum_{n=1}^{3} \hat{\lambda}_2^n(t) = 145 - 35.8t + 0.95t^2 - 0.025t^3 \approx 145 - 35.8t + 0.95t^2. \quad (9.5)$$

From the coefficients c_{i3}, we can see that the cubic term is not too important when t is suitably small, e.g., for $t \leqslant 3$.

From the results displayed above, we see that, just as for example 8.1, the switching causes the net arrival rates $\lambda_i^+(t)$ to differ significantly from the external arrival rates $\alpha_i(t)$. Note that $\lambda_1^+(t)$ and $\lambda_2^+(t)$ both are decreasing at time 0, while only $\alpha_2(t)$ is. Using the quadratic approximations for $\lambda_i^+(t)$ in (9.4) and (9.5), we obtain associated approximations for the mean $m_i(t)$ from (5.3):

$$m_1(t) \approx \lambda_1^+(t-1) + 1.05 \quad \text{and} \quad m_2(t) \approx \lambda_2^+(t-1) + 0.95. \quad (9.6)$$

Since the service times are exponential, we could also calculate the means by solving the system of two ODEs. From (7.6), the two ODEs are:

$$\begin{aligned} \dot{m}_1(t) &= 100 + 10t + m_2(t)(0.3 - 0.05t) - m_1(t), \\ \dot{m}_2(t) &= 100 - 10t + m_1(t)(0.3 - 0.05t) - m_2(t) \end{aligned} \quad (9.7)$$

with an initial condition $m_1(t) = m_2(t) = 0$ for t in the suitably distant past, e.g., for $t = -10$. (We avoid initial conditions in (9.3) by assuming that we start empty at $t = -\infty$.)

10. Short-term forecasting and capacity management

In the previous sections, we have addressed the issues of fitting model parameters to available data from which we can obtain a long-term aggregate forecast, describing the mean as well as some indicator of the variability, such as the variance. For short-term forecasting, where short term may be a few months, we propose a more elementary approach. This approach gives the mean and variance of the forecast and requires the fitting of fewer parameters. A goal here is to better treat batches of private lines, which we have noted may not stay together in service after arrival. Our idea is that in a short time scale the adds and disconnects for any service should tend to be approximately *independent batch Poisson processes*. Let t_0 be the time at which the forecast is made. We exploit historical data on adds and disconnects to estimate a linear arrival rate $\lambda_i^+(t) = c_{i0} + c_{i1}(t - t_0)$ and a linear departure rate $\lambda_i^-(t) = d_{i0} + d_{i1}(t - t_0)$ of orders (batches), $t \geqslant t_0$. For example, we might estimate these linear rates by performing least squares fits (of batches). Similarly, we would use historical data to estimate the order-size (batch-size) distributions associated with adds and disconnects, treating adds and disconnects separately. For simplicity, we might assume that the batch-size distribution does not depend on time. With this method, it is not necessary to assume (and in practice may be important not to assume) that the order size remains fixed over the service lifetime, as we have assumed in our basic model in section 6. Instead, we can directly estimate the arrival and departure order-size distributions from historical data.

Let X_i and Y_i represent the random arrival and departure order sizes (whose distributions have been estimated). Then, assuming that the arrival and departure processes are independent batch-Poisson processes over the time interval $[t_0, t]$, the mean and variance of the number $Q_i(t)$ of lines in service at time t are predicted to be

$$\begin{aligned} m_i(t) \equiv E\big[Q_i(t)\big] \approx m_i(t_0) &+ \left[c_{i0}(t-t_0) + \frac{c_{i1}(t-t_0)^2}{2}\right]E[X_i] \\ &- \left[d_{i0}(t-t_0) + \frac{d_{i1}(t-t_0)^2}{2}\right]E[Y_i] \qquad (10.1) \end{aligned}$$

and

$$\begin{aligned} \sigma_i^2(t) \equiv \operatorname{Var} Q_i(t) \approx &\left[c_{i0}(t-t_0) + \frac{c_{i1}(t-t_0)^2}{2}\right]E\big[X_i^2\big] \\ &+ \left[d_{i0}(t-t_0) + \frac{d_{i1}(t-t_0)^2}{2}\right]E\big[Y_i^2\big]. \qquad (10.2) \end{aligned}$$

The batch-Poisson property is important to avoid underestimating the variance. If we assume ordinary Poisson processes, then the mean estimate $m(t)$ in (10.1) would be unchanged. Then the order-size means $E[X_i]$ and $E[Y_i]$ would be incorporated in the arrival and departure rates. With the Poisson assumption, contributions to the mean and variance would both correspond to $[c_{i0}(t-t_0) + c_{i1}(t-t_0)^2/2]E[X_i]$ and $[d_{i0}(t-t_0) + d_{i1}(t-t_0)^2/2]E[Y_i]$. However, with the compound Poisson assumptions, the variance is larger, because

$$\begin{aligned} &\left[c_{i0}(t-t_0) + \frac{c_{i1}(t-t_0)^2}{2}\right]E\big[X_i^2\big] \\ &\quad = \left[c_{i0}(t-t_0) + \frac{c_{i1}(t-t_0)^2}{2}\right]E[X_i]\big(E[X_i]\big(1 + c_{X_i}^2\big)\big), \qquad (10.3) \end{aligned}$$

where $c_{X_i}^2$ is the SCV of X_i. If $c_{X_i}^2 > 0$, then $E[X_i](1 + c_{X_i}^2) > 1$ and the variance prediction is larger than in the Poisson case.

Given that we can estimate the forecast variance in equation (10.2), this forecast approach provides a basis for tracking the actual usage with reference to the mean forecast. A 2σ deviation can trigger a warning, while a 3σ deviation can indicate that immediate action is necessary to determine the cause for the deviation from the forecast.

It is often possible to do more accurate short-term forecasting than long-term forecasting, because the very large projects leading to very large batch sizes tend to be known and thus can be treated separately. This leads to an adjustment in the mean after analyzing the smaller batches without any increase in the variance.

So far we have focused on the total number of leased lines for each service, which is appropriate from a product manager's view, where the primary concerns are sales, revenue and overall service life cycles. However, the stochastic network model can also be applied for capacity management. When the concern is capacity management, our focus shifts to the number of lines in service at a given node or link (facility). An important simplification when we focus on a single facility is the reduction of the batch effect. When a customer installs a full private-line network, typically only a few lines are used on each facility. Since different services may use common facilities, we may add over different private lines services. Thus we can use the same model to describe the evolution of capacity in use on specified facilities, focusing on one service at a time. Further discussion of capacity management using time-dependent infinite-server queues appears in [15].

When we consider capacity management, we may use either long-term forecasting formulas developed in previous sections or the short-term forecasting formulas in equations (10.1) and (10.2). However, it is important to account for the fact that capacity is often modular; i.e., it often comes in fixed sizes. Assuming that capacity is fungible, as is likely to be the case in an asynchronous transfer mode (ATM) environment, we can apply formulas in [15] to determine required capacities. With modularity, we could increase to the next available size. In particular, the normal approximation for the required capacity at time t from [15] is

$$s(t) = \lceil m(t) + 0.5 + z_a \sigma(t) \rceil, \tag{10.4}$$

where $\lceil x \rceil$ is the least integer greater than x, $N(0, 1)$ is a standard (mean 0, variance 1) normal random variable, $P(N(0; 1) > z_\alpha) = \alpha$ and the "blocking" requirement is that

$$P\bigl(Q_i(t) \geqslant s(t)\bigr) \leqslant \alpha, \qquad P\bigl(Q(t) \geqslant s(t) - 1\bigr) > \alpha. \tag{10.5}$$

11. Conclusions

We have shown how a time-dependent network of infinite-server queues can be used to describe the time-dependent use of private-line telecommunication services with relatively long service lifetimes. We have proposed this methodology primarily to assist product management, whose role in the business process is described in section 3. Sections 4–6 developed the model for a single private-line service. In section 4 we provided a remarkably general expression for the time-dependent mean number of lines in service. In section 6 we developed an approximation for the full distribution based on the assumption that batches of lines connected together remain in service together and then depart together. We suggested keeping the original mean computed via sections 4, 5, and using the approximation only to develop an approximating distribution beyond the mean. In section 5 we pointed out that the need to fit the model to data leads us to assume that the basic model elements are each characterized by only a few parameters. A benefit of these simplifying assumptions is that we obtain more elementary descriptions of system behavior. Examples are the ODE in (5.3) and the quadratic approximation in (5.7).

In sections 7–9 we showed how the model in sections 4–6 can be extended to a stochastic network model for a family of private-line services, the primary objective being to distinguish between new connection requests and switching from one service to another. A decomposition principle makes it possible to first solve for the net arrival-rate functions $\lambda_i^+(t)$ and then afterwards analyze the behavior of each service separately, using the methods of sections 4–6. In section 7, as in section 5, we pointed out that the need to fit the model to data leads us to introduce relatively simple parametric models for the external arrival-rate functions $\alpha_i(t)$ and the switching probabilities $p_{ij}(t)$. In sections 8 and 9 we showed how the special structure assumed for $\alpha_i(t)$ and $p_{ij}(t)$ yields corresponding simple structure in the net arrival-rate functions $\lambda_i^+(t)$. In section 8 we discussed the special case in which the external arrival-rate functions are polynomials (e.g., linear or quadratic), while the switching probabilities are not time-dependent. In section 9 we discussed the more general special case in which both the external arrival-rate functions and the switching probabilities are polynomials, showing the combined effect of both forms of time dependence upon the net arrival-rate functions and the time-dependent mean number of customers in each service. Concrete examples in sections 8 and 9 illustrated both tractability and the insights that can be gained.

We showed that the model structure supports defining a three-phase life cycle for each service, depending upon the behavior of the three functions $\lambda_i^+(t)$, $m_i(t)$ and $\lambda_i^-(t)$, assuming that all three first increase and then decrease. In the growth period, all three functions are increasing, while in the decline period, all three functions are decreasing. Since the peaks of $m_i(t)$ and $\lambda_i^-(t)$ lag behind the peak of $\lambda_i^+(t)$, there is also a mature period, where $\lambda_i^+(t)$ is declining, but one of $m_i(t)$ and $\lambda_i^-(t)$ is still increasing, as in figures 2 and 5. This structure is well illustrated by the case in which $\lambda_i^+(t)$ is a quadratic function, which makes $m_i(t)$ and $\lambda_i^-(t)$ quadratic functions as well. Then the relations among the functions $\lambda_i^+(t)$, $m_i(t)$ and $\lambda_i^-(t)$ are clearly quantified in (5.11) and (5.12). These formulas show that the time lags in the peaks of $m_i(t)$ and $\lambda_i^-(t)$ behind the peak of λ_i^+ are typically different.

In section 10 we showed how the model can be used for short-term forecasting and capacity management. Short-term forecasting can be done by assuming that the adds and disconnects can be modeled by independent batch-Poisson processes.

Throughout we focused on private-line telecommunication services, but the methods should also be applicable to other leased services with long service lifetimes, such as Internet connectivity, cable television and real estate rental.

References

[1] J. Abate and W. Whitt, Numerical inversion of probability generating functions, Operations Research Letters 12 (1992) 245–251.

[2] J. Abate and W. Whitt, Numerical inversion of Laplace transforms of probability distributions, ORSA Journal on Computing 7 (1995) 36–43.

[3] F.M. Bass, A new product growth model for consumer durables, Management Science 15 (1969) 215–227.

[4] F.M. Bass, T.V. Krishnan and D.C. Jain, Why the Bass model fits without decision variables, Marketing Science 13 (1994) 203–223.

[5] *CCMI Guide to Networking Services*, Interlata Private Line Services, Vol. 3 (Center for Communications Management Information, 1995).

[6] R. Chatterjee and J. Eliashberg, The innovation diffusion process in a heterogeneous population: A micromodeling approach, Management Science 36 (1990) 1057–1079.

[7] Data Communications Staff, The 1997 data communications market forecast, Data Communications Magazine 25(17) (1996) 54–64.

[8] J.L. Davis, W.A. Massey and W. Whitt, Sensitivity to the service-time distribution in the nonstationary Erlang loss model, Management Science 41 (1995) 1107–1116.

[9] R.D. Doverspike and V. Jha, Comparison of routing methods for DCS-switched networks, Interfaces 23 (1993) 21–34.

[10] S.G. Eick, W.A. Massey and W. Whitt, The physics of the $M_t/G/\infty$ queue, Operations Research 41 (1993) 731–742.

[11] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd ed. (Wiley, New York, 1968).

[12] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd ed. (Wiley, New York, 1971).

[13] M.R. Garzia and C.M. Lockhart, Nonhierarchical communications networks: An application of compartmental modeling, IEEE Transactions on Communications 37 (1989) 555–564.

[14] J.A. Jacques, *Compartmental Analysis in Biology and Medicine*, 2nd ed. (University of Michigan Press, Ann Arbor, 1985).

[15] O.B. Jennings, A. Mandelbaum, W.A. Massey and W. Whitt, Server staffing to meet time-varying demand, Management Science 42 (1996) 1381–1394.

[16] O.B. Jennings, W.A. Massey and C. McCalla, Optimal profit for leased lines services, in: *Teletraffic Contributions for the Information Age, Proceedings of ITC 15*, eds. V. Ramaswami and P.E. Wirth (Elsevier, Amsterdam, 1997) pp. 803–814.

[17] J.F. Lawless, *Statistical Models and Methods for Lifetime Data* (Wiley, New York, 1982).

[18] K.M. Malone, *Dynamic Queueing Systems: Behavior and Approximations for Individual Queues*, Ph.D. dissertation, Operations Research Center (Cambridge University, Cambridge, MA, 1995).

[19] W.A. Massey, G.A. Parker and W. Whitt, Estimating the parameters of a nonhomogeneous Poisson process with linear rate, Telecommunication Systems 5 (1996) 361–388.

[20] W.A. Massey and W. Whitt, Networks of infinite-server queues with nonstationary Poisson input, Queueing Systems 13 (1993) 183–250.

[21] C.R. Master and L.R. Pamm, The digital data system launches a new era in data communications, Bell Laboratories Record 53 (1975) 420–426.

[22] J.H. Matis and T.E. Wehrly, Generalized stochastic compartmental models with Erlang transit times, J. Pharmacokin. Biopharm. 18 (1990) 589–607.

[23] P. Mertz and D. Mitchell, Aspects of data transmission using private line voice telephone channels, Bell System Technical Journal 36 (1957) 1451–1486.

[24] J.A. Norton and F.M. Bass, A diffusion theory model of adoption and substitution for successive generations of high-technology products, Management Science 33 (1987) 1069–1086.

[25] R. Nucho, Transient behavior of the Kendall birth–death process-applications to capacity expansion for special services, Bell System Technical Journal 60 (1981) 57–87.

[26] M.D. Peterson, D.J. Bertsimas and A.R. Odoni, Decomposition algorithms for analyzing transient phenomena in multiclass queueing networks in air transportation, Operations Research 43 (1995) 995–1011.

[27] G.W. Potts, Exploit your product's service life cycle, Harvard Business Review (September/October 1988) 4–7.

[28] B.W. Schmeiser and M.R. Taaffe, Time-dependent queueing network approximations as simulation external control variates, Operations Research Letters 16 (1994) 1–9.

[29] M. Segal and W. Whitt, A queueing network analyzer for manufacturing, in: *Teletraffic Science for Cost-Effective Systems, Networks and Services, Proceedings of ITC 12*, ed. M. Bonatti (North-Holland, Amsterdam, 1989) pp. 1146–1152.

[30] D.R. Smith, A model for special-service circuit activity, Bell System Technical Journal 62 (1983) 2911–2934.

[31] W. Whitt, The queueing network analyzer, Bell System Technical Journal 62 (1983) 2779–2815.