Which Arrival Law Parameters Are Decisive for Queueing System Performance

Reto Grünenfelder*) and Stephan Robert**)

*) Alcatel STR, Friesenbergstr. 75, CH-8055 Zurich, e_mail: gruenenfelder@alcatel.ch
**) Swiss Federal Institute of Technology (EPFL), Telecommunications Laboratory, CH-1015 Lausanne, e_mail: robert@tcomhp20.epfl.ch

Abstract This paper addresses the problem of finding the decisive parameters, the so-called key parameters, of a queueing system which most significantly influence expected occupation and loss of a finite capacity queue. It is shown that the key parameters of observational data, which are the expected arrival rate, the expected service rate and the spectral density at the frequency zero of the difference between arrivals and service time most significantly influence the performance of the queueing system. An algorithm is developed which shows that it is mostly possible to fit an MMPP(2) to the key parameters. A numerical example illustrates the importance of the key parameters and also shows the accuracy of the proposed fitting procedure.

I. Introduction

For the last few years the demand for multimedia services has continuously been increasing. To satisfy the provision of the Broadband Integrated Services Digital Network (B-ISDN) with high bandwidth becomes necessary. The Telecommunications Standardization Sector (ITU-T), the former CCITT, has selected the Asynchronous Transfer Mode (ATM) as the switching technology for the B-ISDN. ATM ensures the integration and the switching of all the required different services.

The multimedia traffic arriving to an ATM network is the superposition of various cell streams generated by individual input sources, such as voice, data and video. Basically, all the sources can be split into four service classes (A, B, C and D) depending on the requirements for Constant (CBR) or Variable Bit Rate (VBR) and timing relationships [12]. To provide these four service classes, four ATM Adaptation Layers (AALs) are prepared which are also based on bit rate (constant or variable) and timing relationships.

The basic parameter of the traffic descriptor of class A sources, in other words AAL1 compatible sources, is the peak cell rate. When considering, e.g. the Usage Parameter Control (UPC) at the User Network Interface (UNI) or the Quality of Service (QoS) parameter delay jitter, the maximally allowed Cell Delay Variation (CDV) must be taken into account [11,13].

For all the other service classes the traffic descriptors characterize a VBR profile. Connection-oriented traffic is mostly delay sensitive and connectionless traffic is loss sensitive. To summarize, from the performance point of view delay jitter and cell loss due to UPC or congestion are the most important network performance parameters.

It is well known that traffic streams do not form a renewal process and they are mostly bursty and correlated, e.g. [10]. An accurate stochastic process characterizing such single
traffic streams is the Semi Markov Process (SMP) [14,18]. The SMP is a Markov modulated doubly stochastic process. The Special Semi Markov Process (SSMP), a special class of the SMP, is an equivalent to the Discrete time Markovian Arrival Process (DMAP) [3]. It is very well suited to input traffic modelling [4]. Every service, or in other words traffic stream, of a multimedia communication can accurately be fitted to an SSMP. The aggregation of all these particular services yields the traffic stream of the multimedia communication. The statistical multiplexer is a well known and accurate model for superimposing different traffic streams. Its output process, the traffic descriptor of the multimedia communication, is an SSMP as well. The superimposed input traffic stream is an SSMP and its transition matrix is given by the Kronecker product of the transition matrices of the single traffic streams. The number of arriving cells in each phase is given by the Kronecker sum [17]. Obviously, the state space of the multimedia traffic descriptor becomes enormous.

When focusing only on the most relevant network parameters, typically delay jitter and cell loss, the state space can drastically be reduced. In an ATM system the mean cell delay jitter is given by the mean waiting time in the ATM switches and multiplexers. The cell loss ratio is given by the cell loss probability. When designing or measuring ATM systems, it is important to know which traffic parameters are generic and influence most significantly the system. In section II we show under very general conditions that the mean cell delay jitter and the cell loss strongly depend on the arrival law's key parameters, namely the load and the variance of each service and the spectral density at frequency zero, which is the sum over all lags of the arrival law’s autocovariance function. In section III we give an algorithm for fitting observational data to an MMPP(2), while matching the key parameters. Furthermore, we show that it is not always possible to fit these three most important arrival law parameters of the aggregated traffic streams to an SSMP with only two phases while each phase describes batch arrivals. We mainly focus on an SSMP with two states since the MMPP with two states, a particular case of the SSMP, is the best known and most used traffic descriptor for VBR sources, e.g. [1,4,10]. In section IV, we give one numerical example for demonstrating the accuracy of the fitting method, other examples [9] to the method developed in this paper demonstrate its accuracy as well.

II. Important Queueing Parameters

In this section we consider the mean cell delay and the cell loss during a busy period of a queueing system. A busy period is defined to begin with the arrival of a cell to an idle queueing system (time \( t \)) and to end when it next becomes idle (time \( \tau \)). Let \( N(t) \) denote the number of cells in the queueing system at time \( t \), where \(-\tau < t < \tau\), \( N(t) > 0 \) and \( N(-\tau) = N(\tau) = 0 \) and \( U(t) \) the number of arriving minus the number of served cells at time \( t \). Furthermore, we assume the queueing system's capacity to be limited to \( c \) places. The contents \( N(t) \) is then given by

\[
N(t) = \min\left\{ c, \int_{-\tau}^{\tau} U(s) \, ds \right\}, \quad -\tau \leq t \leq \tau. \tag{1}
\]

Let \( M(t) \) denote the cell loss at time \( t \), it is given by
\[ M(t) = \max \left\{ 0, -c + \int_{-\tau}^{t} U(s) ds \right\}, \quad -\tau \leq t \leq \tau. \quad (2) \]

We now focus only on the queueing system occupation \( N(t) \). Later on, it will be shown that similar arguments can also be applied to the loss \( M(t) \). Define the function \( \psi(\omega) \) by

\[ \psi(\omega) = \frac{1}{2\tau} \int_{-\tau}^{\tau} N(t)e^{-i\omega t} dt \quad (3) \]

where \( \psi(0) = E[M-\tau<t<\tau] \). \( E \) is the operator of the expected value. An excess period is defined to begin when the queueing system starts to be completely occupied, e.g. \( t_1 \) in figure 1, and to end when it next becomes only partially filled, e.g. \( t_2 \) in figure 1, this means \( N(t_1^-)<c \) and \( N(t_2^+)<c \).

![Figure 1. Queueing system occupation versus time.](image)

We now assume that there are \( k \) excess periods during the busy period \(-\tau<t<\tau\), \( t_0=-\tau \) and \( t_{2k+1}=\tau \). After substituting (1) into (3) and partial integration we obtain

\[ \psi(\omega) = \frac{-1}{i2\omega} \int_{-\tau}^{\tau} U(t)e^{-i\omega t} dt = \frac{-1}{i2\omega} \sum_{j=0}^{k} \int_{t_{2j}}^{t_{2j+1}} U(t)e^{-i\omega t} dt. \quad (4) \]

It is useful to remember that

\[ dN(t) = \begin{cases} 0, & \int_{-\tau}^{t} U(s) ds \geq c, \quad -\tau \leq t \leq \tau. \\ U(t) dt, & \text{otherwise,} \quad -\tau \leq t \leq \tau. \end{cases} \]

We next introduce the characteristic function \( \chi_{[a,b]} \), which is 1 for \( a \leq t \leq b \) and 0 otherwise. We define \( \mathcal{F}\{f(t)\} \) to be the Fourier Transform of a function \( f(t) \). Hence (4) can be expressed as Fourier transform, namely
\[
\psi(\omega) = \frac{-1}{i2\omega \tau} \sum_{j=0}^{k} \int_{t_j-\tau}^{t_j+\tau} \chi_{[t_j, t_{j+1}]} U(t)e^{-i\omega t} dt = \frac{-1}{i2\omega \tau} \sum_{j=0}^{k} \mathcal{F}\left\{ \chi_{[t_j, t_{j+1}]} U(t) \right\}. \tag{5}
\]

We now calculate \(\psi(\omega)\overline{\psi}(\omega)\), which is for \(\omega=0\) the square of the expected value of the queueing system occupation \(N\).

\[
\psi(\omega)\overline{\psi}(\omega) = \frac{1}{4\tau^2 \omega^2} \sum_{j,l=0}^{k} \int_{-\tau}^{\tau} dt \int_{-\tau}^{\tau} d\tilde{t} \chi_{[t_j, t_{j+1}]} \chi_{[\tilde{t}_l, \tilde{t}_{l+1}]} U(t)U(\tilde{t})e^{-i\omega t} e^{i\omega \tilde{t}} \tag{6}
\]

Changing the variables to \(v = t - \tilde{t}\) and \(t = \tilde{t}\) yields

\[
\psi(\omega)\overline{\psi}(\omega) = \frac{1}{4\tau^2 \omega^2} \sum_{j=0}^{k} \int_{-2\tau}^{2\tau} dv \int_{-\tau}^{\tau} \chi_{[t_j, t_{j+1}]} U(t)U(t-v) dt. \tag{7}
\]

In the following we assume the process \(U(t)\) to be ergodic, this means

\[
\frac{1}{2\tau} \int_{-\tau}^{\tau} \sum_{j=0}^{k} \chi_{[t_j, t_{j+1}]} U(t)U(t-v) dt = \mathbb{E}\left[U(t)U(t-v)\right]_{t \in \Omega} \tag{8}
\]

where \(\Omega = \bigcup_{j=0}^{k} \{t: t_j \leq t \leq t_{j+1}\}\). Furthermore, we assume the process \(U(t)\) to be wide sense stationary, which means that \(\mathbb{E}[U(t)]=\eta\) is independent of \(t\), and that the autocorrelation \(\mathbb{E}[U(t+v)U(t)] = R_U(v)\) only depends on the lag \(v\). Obviously, \(R_U(v)\) is symmetric, that means \(R_U(v)=R_U(-v)\). Hence, expression (8) becomes \(R_U(v)\).

The spectral density \(\Phi_U(\omega)\) of the process \(U(t)\) is the Fourier transform of the autocovariance function \(C_U(v) = R_U(v)\eta^2\) [6]. With this and the convolution theorem for Fourier transforms, expression (7) becomes

\[
\psi(\omega)\overline{\psi}(\omega) = \frac{1}{2\tau \omega^2} \mathcal{F}\left\{ \chi_{[-2\tau, 2\tau]} \ast (\Phi_U(\omega) + \mathcal{F}\left\{ \eta^2 \right\}) \right\} = \frac{1}{\tau \omega^2} \int_{-\infty}^{\infty} \frac{\sin(2\alpha \tau)}{\alpha} (\Phi_U(\omega - \alpha) + 2\pi n/2 \delta(\omega - \alpha)) d\alpha. \tag{9}
\]

In the following we approximately evaluate the integral of (9) at \(\omega=0\). The term \(\sin(2\alpha \tau)/\alpha\) has its main lobe at \(\alpha=0\). The greater \(\tau\) the more important becomes this main lobe. Its width
decreases with increasing $\tau$. Or in other words, the greater $\tau$ the more important become low frequencies. We therefore assume that $\omega$ is proportional to $1/\tau$. This yields for small $\alpha$ that
\[
1/\tau \omega^2 \approx \omega / (\omega^2 + \alpha^2),
\]
where $\approx$ means approximately proportional to. 
Equation (9) can be given as an approximation, when substituting (10) into (9), taking $\omega \to 0$, and limit and integration can be interchanged to obtain
\[
\psi(0) \overline{\psi}(0) \approx \int \lim_{\omega \to 0} \left( \frac{\omega}{\omega^2 + \alpha^2} \sin(2\alpha \tau) \left( \Phi_U(\omega - \alpha) + 2\pi \eta^2 \delta(\omega - \alpha) \right) \right) d\alpha.
\]
Since $\lim_{\omega \to 0} \frac{\omega}{\omega^2 + \alpha^2} = \pi \delta(\alpha)$ and $\lim_{x \to 0} \frac{\sin(2\alpha \tau)}{x} = 2\pi \tau$, we obtain
\[
\psi(0) \overline{\psi}(0) = (E[N|t < t < \tau])^2 \approx 2\pi \tau \left( \Phi_U(0) + 2\pi \eta^2 \right).
\]
Equation (12) shows that during a busy period the squared expected value of the queueing system occupation is approximately proportional to an expression in which the following parameters are involved
- the expected value of the process $U(t)$, that means the difference between arrival rate and service rate
- the spectral density $\Phi_U$ at frequency 0 of the process $U(t)$
- the duration of the busy period, given as $\tau$.

The spectral density at frequency zero is equal to the integral over all lags of the autocovariance function of the process $U(t)$. This shows that the occupation of the queueing system does not depend on the detailed but the global behaviour of the autocovariance function of the process $U(t)$. For ATM systems typically, the service rate is constant and hence, only the spectral density at frequency zero of the arrival process influences the queueing system occupation. Furthermore, expression (12) shows that the expected value of $U(t)$, the difference of mean arrival and service arrival rate, also decisively influences the queueing system occupation. The $\Phi_U(0)$ becomes for non-correlated arrivals equal to the variance. That means $\Phi_U(0)$ consists of the variance plus an extra term which characterizes the correlation. Thus, the influence of correlation on the queueing behaviour becomes more evident. It is therefore suitable to consider also the variance of $U$, which gives an integral information about the whole spectral density, as a key parameter. The duration of the busy period strongly depends on the fluctuations of the service process of the queueing system [7]. Therefore, it is obvious to replace the relevant parameter busy period by the relevant parameter: variance of the process $U(t)$. Hence, the key parameters for the mean occupation - also for the loss as seen later on - of a queueing system are
- the expected arrival and the expected service rate, more precisely the expectation of the process $U(t)$
- the variance of the process $U(t)$
- the spectral density at frequency zero of the process $U(t)$.

A result which is similar to that of the mean occupation can now be derived easily for the mean cell loss during the busy period. Let $\Phi(\alpha)$ denote the Fourier transform of the mean cell loss. The busy period of the loss starts, see figure 1, at the time $t_I$ and ends up with $t_{2k}$. In the
following, these instants will be named as $-\tau$ and $\tau$ respectively. The Fourier transform of the mean loss is then given by

$$\varphi(\omega) = \frac{-1}{i2\omega\tau} \sum_{j=1}^{l_2} \int_{j}^{l_2+1} U(t)e^{-i\omega\tau} dt.$$  

(13)

This equation is form-invariant to (4). Finally, we obtain for the squared mean loss a similar expression as for the mean occupation.

$$\varphi(\omega)\overline{\varphi}(\omega) = \left( E[M | -\tau < r < \tau] \right)^2 \approx 2\pi\tau (\Phi_U(0) + 2\pi\eta^2)$$  

(14)

It is important to point out that the factor $2\tau$ in (14) corresponds to the elapsed time between the first and the last loss during a busy period. Furthermore, it is very important to note that it is an approximate proportionality. In other words, both equations (12) and (14) do not serve for the computation of the mean occupation and the mean cell loss in a queueing system, but they clearly show the importance of parameters. These key parameters are, as already pointed out above, the spectral density at frequency zero of the process $U(t)$, the variance, the mean arrival rate and the mean service rate. This is valid for the mean occupation as well as the mean loss. The analytical study of the mean occupation of the $\Sigma 2SM/D/1$ queue [15] clearly proves that the mean occupation only depends on the key parameters mentioned above. ($2SM$ means two state Markov process, which is a generalization of the Bernoulli process.)

III. The Fitting Procedure

The probably best known process characterizing non-renewal arrivals is the Markov Modulated Poisson Process (MMPP) with two phases. Several fitting procedures for the MMPP(2) are given in the literature, e.g. [1,4,10,16,19].

The procedure in [16] focuses on the parameter estimation of an MMPP(2) based on observational interarrival times. The method bases on iteration and is motivated by the maximum likelihood estimation. The method in [4] allows observational data of a process to be completely fitted to the distribution function of an SSMP(2). It is furthermore possible to fit the first lag of the autocovariance function. This method assumes that the arrivals in state i occur according to a general process (not Poisson). The methods of [10,19] are similar. Both fit the mean arrival rate and the long term variance-to-mean ratio of the number of arrivals in $(0,t)$ with $t \to \infty$. Furthermore, [10] proposes to fit the variance to mean ratio and the third moment, both of the number of arrivals during a finite interval. [19] proposes to fit the covariance of the number of arrivals of two consecutive infinitely long time intervals and the squared coefficient of variation of the arrival times. The common goal of the four last procedures is that they propose to accurately fit the arrival process of a queueing system.

The main goal in the fitting procedure in [1] is to fit the parameters of the MMPP(2), such that important effects of the queueing behaviour can be seen, e.g. cell loss versus buffer capacity, cell and burst scale behaviour. The procedure consists of splitting the superimposed traffic stream into an underload and an overload phase. During underload, the instantaneous arrival rate is smaller than the mean arrival rate. The definition of overload is analogous. It has been shown, [1] figure 1, that such a fitting procedure is more accurate than the former ones.
In the following, we present a technique for approximating superimposed traffic streams to an SSMP(2). We assume that when the SSMP is in state $Y_i, i \in \{1, 2\}$, arrivals occur according to a Poisson process of rate $\lambda_i$. The modulator's transition probabilities are defined as $p = pr(Y_{i+1} = 2|Y_i = 1)$ and $q = pr(Y_{i+1} = 1|Y_i = 2)$. The aim of our fitting procedure is to fit the most relevant parameters, that means those which most decisively influence the performance of the queueing system. It can be shown [4] that such an SSMP(2) is equivalent to the MMPP(2).

Let $X$ denote the random variable characterizing the number of arrivals during one time slot. A time slot is a fixed length interval. Its length is equal to the time it takes to transmit one cell. For the considered SSMP(2) the following set of equations can easily be derived

$$E[X] = \frac{1}{1 + \alpha} (\lambda_1 + \alpha \lambda_2), \quad Var[X] = E[X] + \frac{\alpha}{(1 + \alpha)^2} (\lambda_1 - \lambda_2)^2$$

(15, 16)

$$E[X^3] = -2E[X] + 3E[X^2] + \frac{1}{1 + \alpha} (\lambda_1^3 + \alpha \lambda_2^3), \quad \Phi_X(0) = E[X] + \left(\frac{2}{p + q} - 1\right) (Var[X] - E[X])$$

(17, 18)

with $\alpha = p/q$.

Now we consider the statistical problem of fitting an SSMP(2) with Poisson arrivals, in the following called MMPP(2), to observational data. Let $\mu_X^{(n)}$ denote the nth moment of the observational data, $\sigma_X^2$ the sample variance and $\Phi_X(0)$ the spectral density at frequency zero. Techniques for measuring on-line such data are presented in [8]. The knowledge of the key parameters is of particular interest for testing performance and network behaviour of ATM. From equations (16) and (18) a priori follows that observational data can only be matched accurately to an MMPP(2) if $\sigma_X^2 \geq \mu_X^{(1)}$ and $\Phi_X(0) \geq \mu_X^{(1)}$ respectively. The first restriction obviously comes from the assumption that arrivals occur for each state according to a Poisson process. The second limitation is valid for all SSMPs with a two state modulator. Equation (18) is independent of the Poisson assumption, since the spectral density of any SSMP only depends on the modulator and the mean rate of each state. To summarize, relevant statistical parameters of observational data cannot always be matched to an MMPP(2) or even the more general SSMP(2). To demonstrate the relevance of the key parameters, found in the last section, we stick to the MMPP(2). The relationship between the transition probabilities $p$ and $q$ in (19) is useful when matching observational data to an MMPP(2). Namely, all the pairs of transition probabilities $(p, q) \in A$ match the observed mean and the observed variance to an MMPP(2)

$$A = \left\{(p, q) : 1 \geq p \geq q > 0, \delta = \frac{\sigma_X^2 - \mu_X^{(1)}}{(\mu_X^{(1)})^2}, \text{with} \begin{cases} 0 \leq q \leq 1 \text{ iff } 0 \leq \delta \leq 1 \\ 0 \leq q < 1/\delta \text{ iff } \delta > 1 \end{cases} \right\}$$

(19)

Its proof is given in [9]. A further restriction for the choice of the pairs of transition probabilities is given by equation (18). Given that the observed mean and variance match
with the MMPP(2), the observed spectral density at frequency zero can be matched as long as \((p,q) \in \mathcal{B}\) where

\[
\mathcal{B} = \left\{ (p,q) : p + q = \frac{2(\sigma_X^2 - \mu_X^{(1)})}{\Phi_X(0) + \sigma_X^2 - 2\mu_X^{(1)}}, \ 0 \leq p, q \leq 1 \right\}.
\] (20)

Figure 2. Comparison of sets \(\mathcal{A}\) and \(\mathcal{B}\). For case a) an empty intersection of is possible. In case b) the intersection never becomes empty.

Obviously, the set \(\mathcal{A} \cap \mathcal{B}\) may be empty, which means that either mean and variance or the spectral density at frequency zero do not matched with the MMPP(2). In the case of \(\mathcal{A} \cap \mathcal{B} = \emptyset\), the optimal \((p,q)\) can be found by e.g. minimizing

\[
\left| \left( E[X^3] \right) - \mu_X^{(3)} \right|^2.
\]

To summarize, the algorithm for matching the key parameters of observational data to an MMPP(2) is given in the list below.

**Algorithm**

Remark: The rates \(\lambda_1\) and \(\lambda_2\) used for the steps 3 and 5 are given by:

\[
\lambda_1 = \lambda_2 + M \quad \text{and} \quad \lambda_2 = E[X] - \frac{q}{p+q} M \geq 0 \quad \text{with} \quad M \equiv (p+q) \sqrt{\frac{\text{Var}[X] - E[X]}{pq}}
\]

1. Determine the sets \(\mathcal{A}\) given by (19) and \(\mathcal{B}\) given by (20)
2. if \(\mathcal{A} \cap \mathcal{B} \neq \emptyset\) goto 3, otherwise goto 4
3. Compute the optimal pair \((p,q)\) based on (17), goto 5;
4. Compute the optimal pair \((p,q)\) such that the matching error between observed and theoretical spectral density becomes minimal. Remark: For an approximately minimal error choose the pair \((p,q) \in \mathcal{A}\) which has the smallest distance to the set \(\mathcal{B}\)
5. Compute \(\lambda_2\) and then \(\lambda_1\).
IV. Numerical Example

The following numerical example demonstrates the relevance of the key parameters found in section II. More examples are presented in [9]. We assume that the observational data will be produced by an SSMP(3) with Poisson arrivals. We fit this arrival law with an MMPP(2) while matching the key parameters with the presented fitting algorithm. Afterwards, we compute the expected loss of the SSMP(3)/D/1/c queue and the MMPP(2)/D/1/c queue. An overview on how to compute efficiently the SSMP(n)/G/1/c queueing system is presented in [9]. The comparison of the expected loss for the observational data (SSMP(3)) with the matched MMPP(2) shows that the proposed fitting guarantees good results.

Arrival law: \( A = \begin{pmatrix} 0.05 & 0.95 & 0 \\ 1E-4 & 0.99989 & 1E-5 \\ 0 & 0.95 & 0.05 \end{pmatrix} \), \( \lambda_1 = 50, \lambda_2 = 0.50, \lambda_3 = 5.0 \)

V. Conclusions

Firstly, we have considered the expected occupation and the expected loss of a finite queue during a busy period. This consideration has shown that the queueing behaviour strongly depends on three parameters, the so-called key parameters, which are the expected arrival rate, the expected service rate and the spectral density at the frequency zero of the difference between arrivals and service. This difference is supposed to be ergodic and wide sense stationary. A typical example of such a process is the SSMP. The motivation for using an SSMP for describing traffic has been given by multimedia traffic. Furthermore, the aggregation of ATM traffic can be described easily by such a process. The disadvantage is that the state space of the underlying Markov chain dramatically grows. To exploit the results about the key parameters of section II, we presented in section III a method to fit the key
parameters of observational data to an MMPP(2). The importance of the key parameters has been illustrated by a numerical example, others are given in [9], which make evident that only the key parameters have a relevant influence on the queuing behaviour. The use of the fitting algorithm is manifold, e.g. SSMPs with a tremendous state space can be reduced significantly, observational data can be fitted accurately. The computation time of the queuing problem with SSMP(n) batch arrivals is very short. The findings can serve as a framework for an engineering tool which allows approximate calculation of loss and delay jitter in an ATM network. Furthermore, the knowledge of the key parameters brings some basic and important insight for the parameter choice of the performance test and the end-to-end test of ATM traffic.

References