Discrete-time queueing systems for data networks performance evaluation

G. Pujolle

Laboratoire MASI
Université P. et M. Curie (ParisVI)
45 avenue des Etats-Unis
78000 Versailles
France

Abstract:
Most of the data communication networks are known through measurements obtained only at discretized points in time. To study the performance of such systems, the model must use a discrete-time queueing network. A slot is the unit of time for the discrete-time model. The length of the slot depends upon the knowledge of the system. The smaller average the interarrival and service times are in comparison with the slot duration, the less we know about the system.

In our model, the input process consists of bulk arrival depending on the number of customers in the queue. The service time process is also a bulk Bernoulli process. More precisely, the time axis is segmented into a contiguous sequence of time intervals (slots) of duration \( \Delta \), the probability of \( i \) arrivals in a slot has a geometric distribution with parameters \( p_i \), and the probability of \( i \) departures in a slot is also a geometric distribution with parameters \( r_i(k) \) where \( k \) is the number of customers in the queue. We show that a queueing network consisting of such queues has a product-form solution. This discrete-time queueing network can be seen as the equivalent of Jackson's network and can be used very easily when the system is not completely known.

Applications to the performance evaluation of ISDN is provided in the last part of this paper.

I- Introduction

Consider a continuous-time queue where arrivals and departures take place at any instant of time as shown in Fig. 1.

In most of the systems, we do not know the exact instants of arrivals and departures. We define a slot i.e. a unit of time, during which the number of departures and arrivals can be counted. For example, this slot may be a quantum in a computer, the time to send a packet in a packet switched network or just a unit of time expressed in microseconds, or nanoseconds, seconds, etc.

The problem is: the smaller the average interarrival and service times are in comparison with the slot duration, the less we know about the system. For example, let us assume that the slot duration is 1 ms. If the mean interarrival time is

---

Figure 1 - The continuous-time model
10 ms, on the average the number of arrivals per slot is 0.1. This necessitates that we know very well the behavior of the system. Now, on the other hand, if the slot duration is 10 seconds, the mean number of arrivals is 1000. It turns out that most of the customers enter and leave this system in the same slot.

If the slot time decreases and goes to 0, then the system becomes continuous and we can use the familiar continuous time modelling techniques.

So, now assume that the continuous time model is replaced by a discrete time model where the slot duration is defined following the knowledge of the system. If the knowledge is perfect, the slot time goes to 0. Generally, when we assume that the interarrival times follow a Poisson process, or when the service time distribution is exponential, the knowledge is not perfect. The classical approximation is to use a slot duration going to 0. Here, we assume that we know the system through the statistics associated with the duration of a slot. The Fig. 1 is now replaced by the system described in Fig. 2.

![Figure 2 - The discrete-time model](image)

We see in this model that some customers can enter the system and also leave it in the same slot. This assumption is very important because the reversibility property in a discrete time queueing system could be obtained only with this assumption.

II - Assumption of the model

Let us assume that the time axis is segmented into a contiguous sequence of time intervals of duration $\Delta$. First, we consider a simple model with just one queue. Then, we shall examine a general queueing system.

Let us assume that the arrival and service processes are extended Bernoulli processes [9], namely random processes for a discrete-time system. Formally, an extended Bernoulli process is a discrete-time integer-valued process $\{X_k, k = 0,1,2,...\}$ where the $X_k$s are iid with a geometric distribution. This process is defined as follows:

Let $p_i$ be the probability to have $i$ arrivals in a slot time. Let $p$ and $r$ be two constants such that $0 < p < 0.5$ and $0 < r < 0.5$. Let $p^* = 1 - p$ and $r^* = 1 - r$. The quantities $p$ and $r$ are the parameters of the arrival and departure processes defined as:

1. for all $i > 0$, $p_i = \frac{p}{p^*} p_{i-1}$

and by normalization:

$p_0 = \frac{1 - 2p}{1 - p}$ and $p_i = \left(\frac{p}{p^*}\right)^i p_0$

It turns out that the mean number of arrivals per slot is $\lambda = p/(1-2p)$.

2. $\forall k > 0$ with $1 \leq i \leq k$,

$r_i(k) = \frac{r}{r_S} r_{i-1}(k)$

$r_0(k)$ is obtained by normalization. Therefore,

$r_i(k) = \left(\frac{r}{r_S}\right)^i r_0(k)$

or

$r_i(k) = \frac{r^i r_S^{k-i}}{r_S^k + r_S^{k-1} r + r_S^{k-2} r^2 + ... + r^k}$

We see that these relationships are an extension of the simple Bernoulli probability where $p_0=p^*$ and $p_1 = p$ and $p_i = (p/p^*) p_0$. In the sequel, we shall call these probabilities: "extended Bernoulli probabilities", and we shall call an "extended Bernoulli queue" a queue where the customers are served with an "extended Bernoulli probability". In other words, it is a discrete-time integer-valued process where the random variables are iid with a geometric distribution of rate $\alpha = p/p^*$.

Another way to define the transition rates is:

$p_i = \alpha p_{i-1}$

$p_i = \alpha^i p_0$

and
and let $0 < \beta < 1$

\[
\begin{align*}
\displaystyle r_1(k) &= \beta \cdot r_{1-1}(k) \\
\displaystyle r_i(k) &= \beta^i \cdot r_0(k)
\end{align*}
\]

and $r_0(k)$ is obtained by normalization.

Therefore:

\[
\begin{align*}
\displaystyle r_0(k) &= \frac{1 - \beta}{1 - \beta^{k+1}} \\
\displaystyle r_i(k) &= \beta^i \cdot (1 - \beta) \\
\end{align*}
\]

The behavior of the queue is as follows. A customer entering an empty queue may be served in the same slot. Departures take place after the arrivals.

Related works have been reported in the literature assuming that arrivals occur just after the beginning of the time slot, while departures take place just before the end of a slot, e.g. a customer cannot enter and leave the queue in the same slot. Main studies were performed by Bharath-Kumar [1], Hsu [3], Hsu and Burke [4], Hunter [5], Kobayashi [7]. Other studies, concerning applications of discrete time systems are given in the references.

We have to note that with this last assumption, the departure process of a tandem queueing network is not a renewal process. The proof is easy: let us assume that the system is empty with N queues in series. During N time slots we are sure that no customer is leaving the system. The departure process depends upon the state of the system.

With our assumptions and assuming extended Bernoulli processes as arrival and service processes, we have:

\[
P(0) = P(0) \cdot [p_0 + p_1 r_1(1) + p_2 r_2(2) + \ldots + p_k r_k(k) + \ldots] + P(1) \cdot [p_0 r_1(1) + p_1 r_2(2) + \ldots + p_k r_{k+1}(k+1) + \ldots] + P(2) \cdot [p_0 r_2(2) + \ldots + p_k r_{k+1}(k+1) + \ldots] + \ldots + P(i) \cdot [p_0 r_i(i) + \ldots + p_k r_{k+i}(k+i) + \ldots] + \ldots
\]

\[
\forall k > 0
\]

\[
P(k) = P(0) \cdot [p_0 r_0(k) + \ldots + p_{k+j} r_j(k+j) + \ldots] + P(1) \cdot [p_0 r_0(k-1) + \ldots + p_{k+j-1} r_{j-1}(k+j-1) + \ldots] + \ldots + P(k) \cdot [p_0 r_0(k) + \ldots + p_{k} r_{k}(k+j) + \ldots] + \ldots + P(i) \cdot [p_0 r_i(k+i) + \ldots + p_{j} r_{j-k}(i-k+j) + \ldots] + \ldots
\]

The steady state solution is:

\[
P(n) = \left( \frac{p^* r_i}{p^* r} \right)^n \left( 1 - \frac{p^* r_i}{p^* r} \right) \quad \text{when} \quad \frac{p^* r_i}{p^* r} < 1
\]

This could be very easily checked on the general equations. Another way to check the correctness of this solution is through the detailed balance equations or reversibility equations, i.e.:

\[
\forall g,h \quad p(g,h) \cdot P(g) = p(h,g) \cdot P(h).
\]

In effect, $V_k$ and $n$ such that $n-k \geq 0$ we have:

\[
[p_0 r_0(n) + p_1 r_{k+1}(n+1) + \ldots] \cdot P(n) = [p_k r_0(n) + p_{k+1} r_1(n+1) + \ldots] \cdot P(n-k)
\]

This shows [6] that the output process of the general discrete-time queue under study is also an "extended Bernoulli process".

Let us note that the previous equation or detailed balance equation can be decomposed under sub-detailed balance equations:

\[
\forall i, \quad p_i r_{k+i}(n+i) \cdot P(n) = p_k r_i(n+i) \cdot P(n-k)
\]

Summing on $i$, these sub-detailed balance equations give again the detailed balance equation. Namely, if the sub-detailed balance equations hold, the detailed balance equation also holds. We shall use this property in the next section of this paper.

Moreover, due to the reversibility property, the number in the queue at any given time $t_0$ is independent of the departure process prior to time $t_0$. This means that such queues in series have a product form solution.

### III- Networks of "extended Bernoulli queues"

We consider an open network with $J$ queues whose service times are defined by extended Bernoulli processes with parameters $r_j$, $j = 1, 2, \ldots, J$. We shall denote $r_j(k)$, $V_k$, for $j = 1, \ldots, J$, and $i=1, \ldots, k$, the probability that $i$ customers are served out of $k$ in the queue $j$. We have:

\[
r_j^i(k) = \frac{r_j^i(k)}{r_j^0(k-1)}
\]

The external arrivals form an "extended Bernoulli process" $p_i$, $i=0, 1, \ldots$ with $p_i=(p/p^*)p_i$. The probability that an arriving customer enters directly queue $j$, $j=1, 2, \ldots, J$, is $x_{0j}$. The routing probabilities that a customer leaving station $i$ goes to station $j$, are $x_{ij}$, $i=1, 2, \ldots, J$, and $j=1, 2, \ldots, J$. The quantity $x_{i,j+1}$ is the probability that a customer leaving station $i$ departs from the network.

Let $e_i$ be the mean number of visits at station $i$. The values of $e_i$, $i=1, \ldots, J$, are given by the well-known equation: $E = X_0 + E X_i$, where $E=(e_1, e_2, \ldots, e_J)$, $X_0=(x_{01}, \ldots, x_{0J})$, and $X=(x_{ij}$,
Let \( k_j \) be the number of customers in queue \( j \) and \( k=(k_1,k_2,\ldots,k_J) \). Finally, define \( k(m)=(k_1,m),\ldots,k_J,m) \) the state of the system at time slot \( m \).

**Theorem**

The equilibrium joint distribution of the open queueing network described above is:

\[
P(k_1,k_2,\ldots,k_J) = \prod_{j=1}^{J} P_j(k_j)
\]

where \( P_j(k_j) = P_j (1 - P_j) \) and

\[
\rho_j = \frac{p_j e_j}{p_j r_j}
\]

if \( \forall \, j \), \( \rho_j < 1 \).

**Proof**

The process \( k(m) \) is a Markov process with transition rate \( p(k,k') \). The reversed process \( k(-m) \) is defined by transition rates \( p'(k',k) \). We are going to show that:

\[
p(k,k') P(k) = p'(k',k) P(k')
\]

which will prove the theorem (see for example [6]).

Let \( T_{k,k'} \) be the transition function from state \( k \) to state \( k' \). We have \( T_{k,k.k=k'} \). The problem is that the probability intensity \( p(k,T_{k,k'.k}) \) is very difficult to write due to the very large number of possible transitions in one slot. So, we are going to decompose the pseudo-detailed balance equation \( p(k,T_{k,k',k}) P(k) = p'(T_{k,k',k}) P(k') \) into pseudo-sub-detailed balance equations: (we use the word pseudo because they are not exactly the detailed balance equations: i.e. \( p(g,h) \) is replaced by \( p'(g,h) \))

\[
\sum_{i=1}^{J} p(k,T_{k,k',k}) P(k) = \sum_{i=1}^{J} p'(T_{k,k',k}) P(k')
\]

where \( T_{k,k'} \) is the transition function related to queue \( i \). In the sequel, we prove that the following equation holds:

\[
p(k,T_{k,k',k}) P(k) = p'(T_{k,k',k}) P(k')
\]

The following figure gives the different arrivals and departures where only feasible states are taken into account:

The quantity \( k'' \) represents the total number of arrivals; the quantity \( k''' \) is the total number of departures. The state \( k' \) is as follows:

\[
k' = (k_1 + k_{i1} - k_{i2}, k_2 + k_{i2} - k_{i3}, \ldots, k_J + k_{Ji} - k_{Ji+1})
\]

Then, the pseudo-sub-detailed balance equation for station \( i \) is:

\[
p_{k_{0i}}(x_{0i},k_{i1}) \cdot r_{k_{i1}} (k_{i1} + k_{i1} - k_{i2}, k_{i2} + k_{i2} - k_{i3}, \ldots, k_{iJ} + k_{iJ} - k_{iJ+1}) P(k) = \\
p_{k_{ji+1}}(x_{ji+1} e_j) \cdot r_{k_{i1}} (k'_{i1} + k_{i1} - k_{i2}, k_{i2} + k_{i2} - k_{i3}, \ldots, k_{iJ} + k_{iJ} - k_{iJ+1}) P(k')
\]
This equation holds since:

\[
P(k) = \left( \frac{P}{p^*} \right)^{k_0} \cdot \left( \frac{r_j e_j}{r_j^*} \right) \cdot \left( \frac{r_i e_i}{r^*} \right) \quad \ldots
\]

We have several other properties for this network. In steady-state, customers leave the network in an "extended Bernoulli process" with the probabilities \( p_i \), \( i = 0, 1, \ldots \), and the state of the system at a given time \( t_0 \) is independent of departures from the system prior to time \( t_0 \).

Let us now consider the same network but closed, e.g., external arrivals and departures out of the network are prohibited. Let \( K \) be the fixed total number of customers. We have the following result:

**Theorem**

The equilibrium joint distribution of the closed network is:

\[
P(k_1, k_2, \ldots, k_j) = G \prod_{j=1}^{k_j} P_j
\]

where \( P_j(k_j) = p_j \) and

\[
p_j = \frac{r_j e_j}{r_j^*}
\]

\( G \) is the normalizing constant.

**Proof**

The relations established in the case of the open network, for transitions arising from the movement of a customer from one queue to another apply here also. We only need to remove the transmission from outside and to outside. We have: \( p(k, k') \cdot P(k) = p'(k', k) \cdot P(k') \) where, as before, \( p'(k', k) \) is the transition rate of the reversed process.

Several classes of customers may be taken into account. If we assume there is no constraints on the service times and on the routing probabilities, each class of customers works in parallel. Even if customers can change class, provided that the independence between the classes is maintained, we have the following general theorem on mixed networks of extended Bernoulli queues.

Let the quantity \( e_{ir} \) be the relative frequency of the number of visits to station \( j \) by a customer of class \( r \). These values are obtained solving the system \( E = X_0 + EX \), where \( E = \{ e_{jr}, \text{ with } j = 1, \ldots, J \text{ and } r = 1, \ldots, R \} \) where \( R \) is the total number of classes. The vector \( X_0 \) gives the routing probabilities arriving from outside (0 if the class is closed) and the matrix \( X = \{ x_{ir}, e_{jr} \} \) the routing probabilities.

Let \( k_r \) be the number of customers of class \( r \) in the station \( j \) and \( k_j = (k_{j1}, k_{j2}, \ldots, k_{JR}) \). Finally, let \( r_j \) be the parameter of the extended Bernoulli probability concerning the service process of customers of class \( r \) at station \( j \).

**Theorem**

The equilibrium joint distribution of the mixed queueing network is:

\[
P(k_1, k_2, \ldots, k_j) = G \prod_{j=1}^{k_j} P_j(k_j)
\]

\[
P_j(k_j) = (k_{j1}, k_{j2}, \ldots, k_{JR}) = k_{j1} \prod_{r=1}^{R} e_{jr}^r
\]

where \( G \) is the normalizing constant.

**IV- Remarks and applications**

The "extended Bernoulli process" is very easy to understand. The number of customers in a batch arrival is geometrically distributed: let \( \alpha = p/p^* \), then we have \( p_j = \alpha^{k_j} (1-\alpha) \) and \( p_0 = 1-\alpha \).

Hence, the extended Bernoulli process is a discrete-time integer-valued process \( \{ X_t, k=0, 1, 2, \ldots \} \) where the \( X_t \)'s are iid with a geometric distribution of rate \( \alpha \).

The product form of the general queueing system can be explained as follows. If a sequence of arriving messages is modeled as an "extended Bernoulli process", then it is equivalent to assume that during a slot the arrival process is a binomial distribution. Knowing that the binomial process is reproductive, the probability distribution of the sum of several binomial distribution is a binomial distribution. We can say that the "extended Bernoulli process" plays the role of the Poisson process for the discrete-time.

A general queueing network with simple Bernoulli probabilities do not have a product form because when two Bernoulli streams are merged together, the resulting sequence is not Bernoulli, i.e. the Bernoulli process does not have the reproductive property.

Broadband networking is going to be a very important phenomenon in the information society of the future. Many of the switching solutions for broadband networking fall under the general category of fast packet switching. In
fact, two techniques may be used: frame switching and cell switching. The distinction between frame switching and cell switching is really quite simple. Frame switching operates with a variable length frame. Cell switching is working with fixed-length packets that contain a 48 octets payload and a five-octets header.

The mathematical model we developed in the previous part may be used to obtain the performance of future high-speed data networks. The frames or the cells (the packets) are supposed to arrive in the network following an "extended Bernoulli process". The messages are decomposed into fragments. This may be modelled using a batch arrival process. Fast packet switches are boxes with N inputs and N outputs which routes the packets arriving on its inputs to their requested outputs. Several architectures designs have emerged in the recent years. They may be classified into three categories: the shared-memory type, the shared-medium type and the space-division type [10]. In all three cases, however, the behavior is similar. The number of packets served by unit of time may be approximated by a truncated geometric distribution. Therefore, the service process of a fast packet switch is well approximated by an "extended Bernoulli distribution" as soon as the number N of input/output is large enough and the unit of time larger than the switch synchronization time.

So, fast packet switching networks may be modelled through a general queueing network using the discrete-time queueing system we solved in the previous part.

References


