A DISCRETE-TIME SINGLE SERVER QUEUE WITH BERNOULLI ARRIVALS AND CONSTANT SERVICE TIME

J.-R. LOUVION, P. BOYER, A. GRAVEY

Centre National d'Etudes des Télécommunications
LAA/SLC/EVP, Route de Trégastel
22301 Lannion, FRANCE

Abstract The recent development of new telecommunication techniques like ATM has pointed out the need for new discrete-time models since the requirements are more important than usual. The Geo/D/1 finite and infinite models are identified as useful for modelling this new technique; they are herein completely solved. This enables to compute analytically all the parameters of these queues and therefore to reach values inaccessible with other methods like simulation.

1 - INTRODUCTION

Many communication systems operate on a discrete-time basis, i.e. there exists a natural elementary unit of time in the behaviour of the system. Classical examples are synchronous communication systems or packet switching systems with time-slotting.

Their evaluation is often carried with approximations of time-continuous systems, either because they are convenient and give computable results or because it is not possible to do anything else [Ko Ko 77].

The recent development of ATD techniques brings a new set of evaluation problems, since the constraints are much more severe:

- the absence of retransmission due to error-recovery implies very low packet loss probabilities, namely $10^{-4}$ to $10^{-10}$.
- the requirements of time transparency imply very low fluctuations in the end-to-end transfer delay.

These demands make practically impossible the use of simulation which is efficient in another range of values (until $10^{-5}$). The need is then crucial for new computable models [BBLR 87]; one of them is the Geo/D/1 queue.

Hereafter is emphasized the usefulness of the Geo/D/1 queue with finite or infinite capacity in ATM modelling, the full solution of these models, some computational techniques and some characteristic results.

2 - GEO/D/1 AND GEO/D/1/K MODELS IN ATM MODELLING

In an ATM environment, a terminal equipment delivers a continuous digital information stream which is sliced into packets by a packetizer; these packets are then inserted on a time-discretized multiplex (see for example the PRELUDE experiment [GoAC 86]); the unit of time associated is called a cell in ATM context.

An ATD switch will direct the packets arriving on the incoming multiplexes onto the outgoing multiplexes. The behaviour of the reemission queue located in front of the outgoing multiplexes is then very important.
The arrival process in such a queue has the following characteristics:

- the arrivals can only occur at discrete epochs, since the switching matrix is synchronized.
- successive arrivals are independent, since there is no time-relation between two successive packets, owing to asynchronism.
- the probability of an arrival occurring at a discrete epoch is constant, since the communications carried by the input multiplexes are statistically equally directed towards the output multiplexes.

It may thus be modelled as a Bernoulli process.

Furthermore, in all ATM experiments, a consensus seems to arise for fixed-length packets denoted cells which fit exactly with the time structure of the multiplex. Since the server of a queue may represent a transmission line, a channel or a multiplex, the constant service time will suit to the type of problems met in ATM. It will be the time necessary for the switching matrix to receive a cell from an input multiplex or to emit it on an output multiplex.

This leads to the discrete-time single-server queue with Bernoulli arrivals and constant service time; in the following, the shorthand notation Geo will be used for a Bernoulli process.

Meisling gave, a long time ago, the mean queue length and the mean waiting time of an infinite Geo/G/1 queue [Meis 57]; later on, Minoli, within a narrowband context, produced a numerical method for obtaining these results in the finite Geo/D/1 queue [Mino 79]; at the same time that the work herein presented, De Somer provided a computational resolution of the Markov system associated to the infinite Geo/D/1 queue [Deso 86]. However, the model has never been fully analytically solved, with results numerically accessible through computation.

3 - MODEL AND NOTATIONS

Consider a single-server queue with a FCFS discipline of service. The time is discretized, i.e. interarrival and service are multiple of the same elementary unit of time denoted $\theta$ in the following; this time is the internal turnover time of a switching matrix; for convenient reasons, it will also be called time-slot, although this refers usually to the time-discretization unit of a multiplex. Furthermore, arrivals and services are synchronized, i.e. a service will begin or end at potential arrival epochs.

Let the stream of arrivals be a Bernoulli process. The arrivals only take place at the epochs $k \theta$, with $k$ integer. The discrete-time Bernoulli process is such that:

- at each discrete epoch $k \theta$, the number of arrivals is 1 with the probability $q$ and 0 with the probability $1 - q$.
- successive arrivals are independent.

It should be noted that the interarrivals are independent with the same geometric distribution; besides, the geometric distribution, like the exponential distribution, enjoys the memoryless property. The arrival rate is $\lambda = \frac{q}{\theta}$.

The modelling of discrete-time queues let arise a difficulty: simultaneity; arrivals and departures actually take place at the end of time-slots, whatever the time where this event may occur in the time-slot; they may thus be simultaneous, which is difficult to take into account in a model, because it is always assumed that only one event can occur at a time.

Therefore, it is assumed that the arrivals take place just before the end of a time-slot and the departures just after the beginning. With this agreement, if an arrival and a departure are simultaneous, the arriving customer sees the departing one just to leave and the leaving customer leaves behind the one just arrived.

Let the service time be constant with a value multiple of the elementary time, say $p \theta$. The offered load is then $\rho = \frac{q}{\theta} p \theta = q p$.
4 - RESOLUTION OF THE GEO/D/1 AND GEO/D/1/K QUEUES

The state of the system will be described by N the number of customers in system, and W the waiting time of an entering customer.

It must be noticed that the state of the system at the arrival of any entering customer is statistically identical to the steady-state: this is the GASTA property (Geometric Arrivals See Time Averages), discrete-time analogue for the well-known PASTA property (Poisson Arrivals See Time Averages) [Half 83].

4.1 The infinite capacity queue

4.1.1 Noteworthy results

The generating function of the number of customers N can be obtained using the method of the imbedded Markov chain:

\[ N(z) = (1 - \rho) \frac{(z - 1)(qz + 1 - q)^p}{z - (qz + 1 - q)^p} \]

In particular, \( N(0) \), \( N'(1) \) and \( N''(1) \) lead respectively to the probability that the system is empty, the mean and variance of the number of customers in system:

\[ P(N = 0) = 1 - \rho \]
\[ E(N) = \frac{\rho}{1 - \rho} \left( 1 - \frac{\rho}{2} - \frac{q}{2} \right) \]
\[ \text{var}(N) = \rho(1 - \rho) + \frac{\rho(p - 1)q^2(3 - 2p)}{2(1 - \rho)} + \left( \frac{\rho(p - 1)q^2}{2(1 - \rho)} \right)^2 + \frac{\rho(p - 1)(p - 2)q^3}{3(1 - \rho)} \]

Using Little's formula, we also obtain a Pollaczek's formula giving the mean waiting time:

\[ E(W) = \frac{\rho}{1 - \rho} \frac{p - 1}{2} \theta \]

Since in a FCFS queue a customer, after its service, leaves behind all the other customers which entered the queue during its sojourn, the generating function of the distribution of the sojourn time \( T(z) \) can be related to \( N(z) \) via the relation:

\[ T(z) = \frac{z}{z - qz + 1 - qz^p} \]

Seeing that the generating functions of the waiting time \( W(z) \) and the sojourn time \( T(z) \) are linked by \( T(z) = zW(z) \), and using (1), we obtain \( W(z) \) quite easily:

\[ W(z) = (1 - \rho) \frac{(z - 1)}{z + q - 1 - qz^p} \]

The probability of null waiting time can then be derived:

\[ P(W = 0) = \frac{1 - \rho}{1 - q} = \frac{P(N = 0)}{1 - q} \]

\( P(W = 0) \) is thus greater than \( P(N = 0) \), although these two quantities are identical in the M/G/1 FCFS queue. This can be roughly explained with the agreement of simultaneity (see §3): an entering customer which finds only one customer before having nearly completed its service, will not wait although the queue is not empty.

4.1.2 Full solution

The probability distribution of W is obtained via a method used in [Sysk 86] where an analysis of the M/D/1 queue is carried out.
The first step consists in comparing the waiting times of two successive customers, say \( k \) and \((k + 1)\). Let \( I_{k+1} \) be the interarrival time between them, and \( W_k \) (resp. \( W_{k+1} \)) the waiting time of the \( k^{th} \) (resp. \((k + 1)^{th}\)) customer. The following Lindley's equation is then achieved:

\[
\{ W_{k+1} \leq i \} = \{ W_k \leq i \theta + I_{k+1} - p \theta \}
\]

The interarrivals are independent geometric random variables; the Lindley's formula then gives in steady-state:

\[
P(W \leq i \theta) = \sum_{j=\max(0, i+1-p)}^{\infty} P(W \leq j \theta) q(1-q)^{j-D-i-1}
\]

This equation then allows to obtain the full distribution of \( W \) using a progressive resolution with sets of size \( p \) (the equation is first solved in the set \( \{0, \ldots, p-1\} \); the results are then used to obtain the next \( p \) values, and so on):

\[
P(W \leq i \theta) = \frac{1-p}{(1-q)^{i+1}} \sum_{k=0}^{j} [q(1-q)^{p-1}] \binom{kp-i-1}{k} \quad \text{if} \quad jp \leq i \leq (j+1)p-1
\]

The probability distribution of \( N \) is then obtained from the one of \( W \), because a simple equation links the number of customers in system \( N \) and the waiting time \( W \):

\[
\{N \leq j\} = \{W \leq (jp - 1) \theta\} \quad \text{if} \quad j > 0
\]

Using (7) and this relation, we obtain:

\[
\begin{align*}
P(N = 0) &= 1 - \rho \\
P(N = 1) &= (1-\rho) \left[ (1-q)^p - 1 \right] \\
P(N = j) &= (1-\rho) \left[ (1-q)^{-jp} + \sum_{k=1}^{j-1} (-1)^k (1-q)^{-k} \binom{(j-k)p+k-1}{k} \left( \frac{q}{1-q} \right)^{k} + \binom{(j-k)p+k-2}{k-1} \left( \frac{q}{1-q} \right) \right] \quad \text{if} \quad j \geq 2
\end{align*}
\]

4.2 The finite capacity queue

In a finite capacity queue, an arriving customer can be rejected if it encounters a full system; the state of the system at an arrival time and at an entry time must then be distinguished. However, thanks to the GASTA property, the results for the state at an arrival time will also be valid for the state at any time - i.e. at the end of an arbitrary time-slot.

Here will only be given these last results; the full solution can be found in [GrLB 87] or [GrLB 88].

In particular, it may be shown that the distributions of the number of customers in the finite and infinite queues are related with a proportional relation. Denoting the quantities of the finite (resp. infinite) queue with the upper index \( K \) (resp. \( \infty \)), the following relation holds:

\[
\frac{P(N^K = j)}{P(N^\infty = j)} = \frac{P(N^K = 0)}{P(N^\infty = 0)} \quad \text{for} \quad j = 0, \ldots, K-1
\]

Hence, the distribution of the number of customers in system is deduced from equation (8) through the preceding normalization.
The loss probability $P_{\text{lost}}$, i.e. the probability that an arriving customer will be rejected is given by the preceding expression of $P(N = K)$.

The probability distribution of the waiting time $W$ is obtained, like in the infinite case, from a Lindley's equation. But, this equation has a more complicated form, since the interarrivals are no more geometrically distributed. However, it may be shown that $W$ takes its values in the set \{0, \ldots, (K - 1)p - 1\}; furthermore, there exists, as for the number of customers, a proportional relation between the distributions of $W$ in the finite and infinite cases, namely:

This leads to the following distribution:

$$P(W = 0) = \left[ (1 - q) \sum_{k=0}^{K-2} (-1)^k (1 - q)^{(K-1-k)p} \left( \frac{q}{1 - q} \right)^k \left( (K - 1 - k)p + k - 1 \right) \right]^{-1}$$

$$P(W = j) = \frac{P(W = 0)}{(1 - q)^j} \sum_{k=0}^{j} (1 - q)^{kp - 1} \left( k (1 - q)^{k} (i + (1 - p)^k) \right.$$ if $j \leq K - 2$

5 - RESULTS

In [Deso 86], De Somer obtains the solution, only for the infinite queue, by a straightforward resolution of the embedded Markov chain; the results presented hereafter are in complete agreement with his.

For other more complicated models, it would then be relevant to use the direct resolution of the imbedded Markov chain.

5.1 Numerical computation

In the formulas (7), (8), (9) and (10), appear summations which are not straightforward computable, since the terms are alternate and of different orders of magnitude; but we have a closed-form expression for the infinite extension of this finite sum; the alternate finite sums are replaced by infinite sums with the notable property of having positive terms; this last sum, although slowly converging will be used because it is easier computable.

The closed-form expression we use is a Lagrange serie [Rior 79]:

$$\frac{1 - q}{1 - \rho} = (1 - q)^i \sum_{k=0}^{\infty} \left[ -q(1 - q)^{p-1} \right]^k \left( i \left( \begin{array}{c} k \\ p \end{array} \right) \right.$$ with $i$ integer

which enables to transform the preceding formulas in terms of infinite series with positive terms.

Denoting $S(q, p, x)$ the following sum:
\[ S(q, p, x) = (1 - q)^{-px} \sum_{k=k_0}^{\infty} \left[ q(1 - q)^{p-1} \right]^k \left( \frac{(k-x)p}{p-1} \right) \quad \text{with } k_0 = \left[ \frac{px - 1}{p-1} \right] + 1 \]

the formulas (7) and (8) of the infinite capacity queue for example, become:

\[ P(W > i \theta) = (1 - \rho) S(q, p, \frac{(i + 1)p}{p}) \quad (7\text{bis}) \]
\[ P(N > j) = (1 - \rho) S(q, p, j) \quad (8\text{bis}) \]

For the finite capacity queue, it is convenient to calculate first the sum \( S_{K-1} = S(q, p, K - 1) \). The equations (9) and (10) can thus be expressed in a similar way:

\[ P(N = 0) = \frac{(1 - \rho)}{1 - \rho(1 - \rho)S_{K-1}} \quad (9\text{bis}) \]
\[ P(N = j) = \frac{(1 - \rho)}{1 - \rho(1 - \rho)S_{K-1}} \left[ S(q, p, j - 1) - S(q, p, j) \right] \quad \text{if } j \leq K - 1 \]
\[ P(N = K) = P_{\text{loss}} = \frac{(1 - \rho)^2 S_{K-1}}{1 - \rho(1 - \rho)S_{K-1}} \]
\[ P(W = 0) = \frac{(1 - \rho)}{(1 - q)} \left[ 1 - (1 - \rho)S_{K-1} \right]^{-1} \quad (10\text{bis}) \]
\[ P(W > i \theta) = \frac{(1 - \rho) S(q, p, \frac{(i + 1)p}{p}) - S_{K-1}}{1 - (1 - \rho) S_{K-1}} \quad \text{if } i \leq (K - 1)p \]

### 5.2 Loss in the finite queue

One of the most important results of the Geo/D/1/K model is the loss probability; it enables us to determine the size of the outgoing control queue necessary to perform a given quality in terms of probability of a packet being lost. Figure 1 page 7 shows this probability function of the offered load \( \rho \) for different values of the capacity.

It can be deduced, for the example of a 8x8 switch, that to perform a loss probability lower than \( 10^{-10} \), the control queue must have at least 32 rooms for a load of .70, 40 rooms for .75, 48 rooms for .80 and 64 rooms for .85. For a given load, the higher the matrix size, the higher the loss.

### 5.3 Relation with the M/D/1 queue

The time-slot \( \theta \) remaining unchanged, let the duration of the service \( p \theta \) increase tending to infinity while the probability of an arrival \( q \) decreases tending to 0, so that the load \( \rho = qp \) remains constant; the Bernoulli process converges then towards a Poisson process and the Geo/D/1 system towards a M/D/1 system.

A rough explanation can be given. If the service time increases, \( \theta \) being constant, it becomes large compared to \( \theta \); it can also be said that \( \theta \) becomes small compared to the service time; in other words, the discretized character of the arrival process thus tends to disappear and the Bernoulli process becomes a Poisson process.

Opposite to an M/D/1/K behaviour, the loss probability in a Geo/D/1/K queue does not depend only on the actual load; this can be observed in Figure 2 page 7 where the loss probability is plotted against the load, for different values of \( p \); the results of the M/D/1 queue can then be seen as the results of a Geo/D/1 queue with \( p \) infinite [ReKo 74].

In the range of loads .70 to .85, a rule of thumb can be given for the behaviour of the loss probability \( P_{\text{loss}} \) against the load in the particular case of capacity equal to 16, since the curves are nearly straight with the scales chosen on the axis.

2.4B.2.6
\[
\log P_{loss} = 15.7 \rho - 16.5 \quad \text{for Geo/D/1/16 with } p = 8
\]
\[
\log P_{loss} = 14.7 \rho - 15.5 \quad \text{for Geo/D/1/16 with } p = 16
\]
\[
\log P_{loss} = 14.3 \rho - 15.1 \quad \text{for Geo/D/1/16 with } p = 32
\]
\[
\log P_{loss} = 14.0 \rho - 14.8 \quad \text{for M/D/1/16}
\]

6 - CONCLUSION

The recent development of ATM has pointed out the need for new discrete-time models, which are a challenge for queueing theory. One of these new models, the single-server queue with Bernoulli arrivals and constant service time with finite or infinite capacity represent quite well the behaviour of a control queue inside an ATD switching matrix.

This model has been described, completely solved and numerically computed. It enables to obtain with a good accuracy the probability of a packet being lost in the queue and therefore to dimension it. This result has been compared to the one obtained with the approximation of the continuous-time queue M/D/1/K, showing a ratio of 3 between the results for a pronounced discrete-time process.

The basic features of ATD modelling being periodic highload and discrete arrivals, limit theorems as Palm-Kintchine for the superposition of renewal processes do not hold anymore; multiplexing schemes have thus to be taken in account yielding to \( D_1 + \cdots + D_p/D/1 \), \( Geo + D[D]/1 \), \( M + D/D/1 \), \( Geo_1 + \cdots + Geo_n/D/1 \) queues. The challenge is then always to be taken up.

7 - REFERENCES


Figure 1. Influence of the capacity in a Geo/D/1/K queue: loss probability vs load for $p = 8$.

Figure 2. Trend of the Geo/D/1/K queue towards the M/D/1/K queue: loss probability vs load for $K = 16$. 

2.4B.2.8