# A note on system contents and cell delay in FIFO ATM-buffers

Bart VINCK and Herwig BRUNEEL [*]
SMACS Research Group [†]
Laboratory for Communications Engineering, University of Ghent
Sint-Pietersnieuwstraat 41, B–9000 Gent, Belgium

**Abstract**

This paper first presents a simple proof of a relationship between the distribution of the number of customers present during an arbitrary slot and the delay experienced by an arbitrary customer, in the context of discrete-time queues with a single server with deterministic service times of 1 slot. Second, this result is applied to derive a relationship between the massfunctions of the same stochastic variables for the multiple server case. In both cases no condition is imposed on the arrival process other than that necessary for the system to be stable.

## 1 Introduction

In the context of ATM switches and multiplexers, one often encounters situations in which customers, i.e. ATM cells, line up in a queue and require one slot of service from a single server or one out of more identical servers. All customers are treated alike, and served according to a FIFO discipline.

In the context of a single-server queue and assuming various types of stochastic models for the arrival process, the delay of an arbitrary customer and the system occupancy during an arbitrary slot have been analysed on several occasions (see e.g. [1, 2, 3, 4]). In this letter we prove that regardless of the nature of the arrival process, a close relationship between the distributions of both exists, so that one can always be derived from the other, making one of both analyses superfluous.

The result has been observed to hold under various more restrictive conditions, none of which is necessary to impose when following the proof presented in this letter. In [3] and [4], the authors allow for correlated as

[*] Both authors wish to thank the Belgian National Fund for Scientific Research (NFWO) for support of this work.

[†] SMACS: Stochastic Modeling and Analysis of Communication Systems.

well as uncorrelated arrivals in the context of infinite buffers. However, they require the system to reach a slot-wise limit-distribution, thereby excluding all arrival processes that lead to a different limit-behaviour. With some additions, their proof could be extended to handle periodic behaviour and finite-capacity buffers, but any extension beyond that is not obvious. We also notice that our new proof is more intuitively appealing and mathematically less complex than theirs.

The result obtained for the single-server queue with deterministic service times of 1 slot can be used to establish similar relationships in the context of queues with a more complicated service-process, such as queue with a general independent service time ditribution or queues with more than one server. In the context of ATM switches the latter is the most important extension. In Sect. 4 we derive a relationship that allows to derive the distribution of the delay of an arbitrary customer from the distribution of the buffer occupancy during an arbitrary slot for the case of a queue with more than one server. The result is in agreement with the result obtained in [5], in which again more restrictive requirements are imposed and a much longer proof is given. Finally, in Sect. 5 we show the result remains valid for finite buffer queues, and we show that no restrictions on the offered load need to be imposed.

## 2 Model

We consider a discrete-time queueing system with (generally) $c$ servers and deterministic service times of 1 slot for each customer. Customers arrive according to a general arrival process and are queued for service according to a FIFO discipline, whereby the order in which customers arriving simultaneously are queued for service is irrelevant. Servers can only remain idle in case not enough customers are present in the queue. Due to the discrete-time operation, time appears as divided into fixed-length slots $S_k = (t_k, t_{k+1})$; all arrivals and departures of customers occur at the slot boundaries. The number of customers arrivang at $t_k$ is denoted as $A(t_k)$, the *buffer occupancy*, i.e., the number of customers present, during $S_k$ is denoted as $U(S_k)$, and the delay of customer $C_j$ (chronologically indiced according to arrival and departure instant) is denoted as $D(C_j)$.

For any sample path of the queueing process, the *load* $\rho$, i.e., the average arrival rate, is defined as

$$\rho := \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} A(t_k) \ , \tag{1}$$

assuming the limit exists. Likewise, using the notation # to indicate the number of elements in a set, the distribution of the buffer occupancy $U$

during an arbitrary slot can be characterised as:

$$\begin{aligned}\mathsf{Pr}\,[U=n] &:= \lim_{k\to\infty}\frac{\#\,\{S_i\,|\,U(S_i)=n, 1\le i\le k\}}{k},\ n=0,1,\ldots,\\ \mathsf{Pr}\,[U=n\,|\,U\ge 1] &:= \lim_{k\to\infty}\frac{\#\,\{S_i\,|\,U(S_i)=n, 1\le i\le k\}}{\#\,\{S_i\,|\,U(S_i)\ge 1, 1\le i\le k\}},\ n=1,2,\ldots \quad (2)\end{aligned}$$

and the delay distribution for an arbitrary customer is given by:

$$\mathsf{Pr}\,[D=n] := \lim_{l\to\infty}\frac{\#\,\{C_j\,|\,D(C_j)=n, 1\le j\le l\}}{l}\ ,\ n=1,2,\ldots\ , \quad (3)$$

again, under the assumption that these limits exist. A sufficient condition for that is that the limit in (1) exists and for that for infinite buffers $\rho < c$.

# 3 Single server queue

In this section we consider the queue with a single server and an infinite buffer. In Sect. 5 we will show the result remains valid for the finite buffer case too. We will derive our relationship from an argument which applies to an individual sample path. We assume that the system is empty during $S_0$ and that

$$\rho < 1\ . \quad (4)$$

As as result, the system empties infinitely often and departures group in finite-length busy periods. Each customer receives service in the busy period he initiates or he arrives in.

Let $S_{k(m)}$ be the last slot of the $m$-th busy period, and let us concentrate on the intervals $[t_0, t_{k(m)+1}]$ that contain a number of entire busy periods. For each $m$ we denote the number of customers arriving (and leaving) during that period as $N(m) = \sum_{i=1}^{k(m)} A(t_i)$. In each such interval, all customers that have arrived are served, and we observe that the number of customers served equals the number of slots in the interval during which at least one customer was present, since (only) during each such slot the server is busy and exactly one customer is served. We thus have that

$$\#\,\{S_i\,|\,U(S_i)\ge 1, 1\le i\le k(m)\} = N(m)\ . \quad (5)$$

A second observation is less obvious. It says that during any busy period the number of slots with system occupancy $n$, for any value of $n = 1, 2, \ldots$, equals the number of customers served that experience a delay of $n$ slots. In order to see this, let us introduce the notion of *the position* a customer is in during a slot, whereby, under the condition of a FIFO service order, the customer being served is in position 1, the customer to be served next is in position 2, etc. More specifically we look at a customer's *position on arrival.* Then, again under FIFO discipline, that position is equal to the delay (expressed in slot times) the customer experiences (see Fig. 1). If the queueing discipline were LIFO, however, and customers are assigned the same positions on arrival so that customers arriving simultaneously are served in reversed order as under FIFO, that position on arrival would be equal to the system occupancy during the customer's service slot (i.e. the slot during which he receives service) under LIFO discipline. However, since the evolution of the buffer occupancy does not depend on the nature of the service discipline, as long as the server remains working whenever at least one customer is present, that buffer occupancy is also the buffer occupancy during the same slot when the queue operates under FIFO discipline.

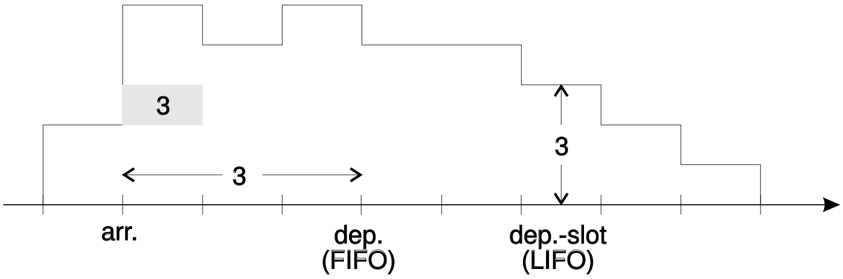

Figure 1: The Single Server Queue

The relationship from an customer to the slot during which he would receive service in case the service discipline were LIFO, is clearly a one-to-one correspondence; therefore the following associated sets have equal cardinality:

$$\#\left\{S_i \,|\, U(S_i) = n, 1 \le i \le k(m)\right\} = \#\left\{C_j \,|\, D(C_j) = n, 1 \le j \le N(m)\right\} \ , \tag{6}$$

for all $m$.

Dividing both the left and the right hand side of (5) by $k(m)$ and taking the limit for $m \to \infty$ we get, in view of (1):

$$\mathsf{Pr}\left[U \ge 1\right] = \lim_{m\to\infty} \frac{\#\left\{S_i \,|\, U(S_i) \ge 1, 1 \le i \le k(m)\right\}}{k(m)} = \lim_{m\to\infty} \frac{N(m)}{k(m)} = \rho \ . \tag{7}$$

Similarly, dividing both sides of (6) by the corresponding sides of (5) and taking the limit for $m \to \infty$ we get, in view of (2) and (3):

$$\mathsf{Pr}\left[U = n \,|\, U \ge 1\right] = \mathsf{Pr}\left[D = n\right] \ , \ n = 1, 2, \ldots \ . \tag{8}$$

Finally, combining (7) and (8), we find the relationship we were looking for:

$$\mathsf{Pr}\left[U = n\right] = \begin{cases} 1 - \rho & , \quad n = 0 \\ \rho\,\mathsf{Pr}\left[D = n\right] & , \quad n \ge 1 \end{cases} \tag{9}$$

## 4 Multiple server queue

In this section we consider a discrete-time queue with $c$ servers, again equipped with an infinite buffer and each customer requiring 1 slot of service from any of the servers. We assume that customers are queued according to a FIFO discipline. The condition we must impose in order to guarantee the stability of the queue now reads

$$\rho = \lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} A(t_i) < c \ . \tag{10}$$

The presence of more than one server, voids both observed interpretations of the *position upon arrival*, so that the reasoning from the previous section cannot be transferred mechanically. We can however make use of the obtained result when we divide each slot in $c$ *minislots* of equal length and consider an *equivalent single server queue* that servers 1 customer per minislot and welcomes the same arrivals. Also in the equivalent queue customers will be served according to a FIFO discipline. The slot-boundaries of the minislots are denoted as $\hat{t}_k$, whereby $\hat{t}_{kc} = t_k$, for any $k \in \mathbb{N}$, see Fig. 2.

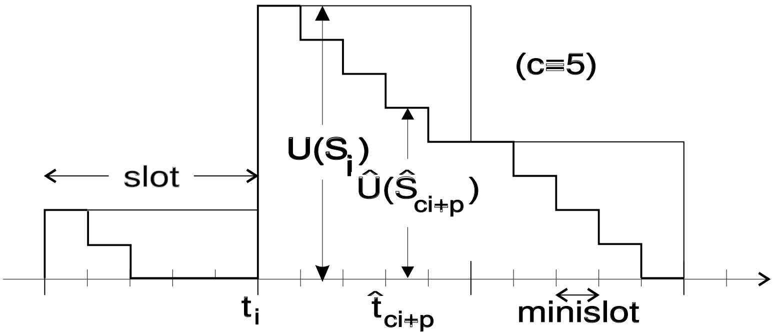

Figure 2: The Multiple Server Queue

In the equivalent single server queue the number of "slots" is raised by a factor $c$, so that the load is $\rho/c < 1$. Therefore (9) holds, i.e.,

$$\mathsf{Pr}\left[\hat{U} = n\right] = \begin{cases} 1 - \rho/c & , \quad n = 0 \\ \rho/c \, \mathsf{Pr}\left[\hat{D} = n\right] & , \quad n \geq 1 \end{cases} \tag{11}$$

where $\hat{U}$ is the system contents during an arbitrary minislot and $\hat{D}$ is the delay — expressed in a number of minislots — of an arbitrary message in the equivalent single server queue. We now translate this in a relationship between the massfunctions of the stochastic variables related to the actual multiple server queue.

First, we observe that during each slot the multiserver queue and the equivalent single server queue serve the same customers. A customer that leaves in the multiserver queue at $t_{k+1}$ leaves the equivalent single server queue on $\hat{t}_{kc+1}$, $\hat{t}_{kc+2}$, $\ldots$or $\hat{t}_{(k+1)c}$, while the arrival instants are the same in both queues. Therefore,

$$D(C_j) = n \Leftrightarrow \hat{D}(C_j) \in \{cn, cn-1, \ldots, cn-c+1\} \ , \tag{12}$$

so that for any $l \in \mathbb{N}$:

$$\#\left\{C_j \,|\, D(C_j) = n, 1 \leq j \leq l\right\} = \sum_{p=0}^{c-1} \#\left\{C_j \,|\, \hat{D}(C_j) = cn - p, 1 \leq j \leq l\right\} \ . \tag{13}$$

Dividing by $l$ and taking the limit for $l \to \infty$ yields:

$$\mathsf{Pr}\left[D = n\right] = \sum_{p=0}^{c-1} \mathsf{Pr}\left[\hat{D} = cn - p\right] \ , \ n = 1, 2, \ldots \ . \tag{14}$$

Second, consider a minislot during which the buffer occupancy in the equivalent queue is $n$, $n \neq 0$. The buffer occupancy in the actual multiserver queue during the slot the minislot belongs to, is an element of $\{n, n+1, \ldots, n+c-1\}$. On the other hand, each slot with a buffer occupancy that belongs to this set, corresponds with exactly one minislot during which the buffer occupancy in the equivalent queue is $n$. We thus have that

$$\#\left\{\hat{S}_i \,|\, \hat{U}(\hat{S}_i) = n, 1 \leq i \leq kc\right\} = \sum_{p=0}^{c-1} \#\left\{S_i \,|\, U(S_i) = n + p, 1 \leq i \leq k\right\} \tag{15}$$

for all $n = 1, 2, \ldots$. Dividing both sides by $kc$ and taking the limit for $k \to \infty$ we get

$$\mathsf{Pr}\left[\hat{U} = n\right] = \frac{1}{c} \sum_{p=0}^{c-1} \mathsf{Pr}\left[U = n + p\right] \ . \tag{16}$$

Using (11), (14) and (16) we come to

$$\mathsf{Pr}\left[D = n\right] = \frac{1}{\rho} \sum_{p=-c+1}^{c-1} (c - |p|) \mathsf{Pr}\left[U = cn + p\right] \ , \ n = 1, 2, \ldots \tag{17}$$

which allows to derive the massfunction from the delay of an arbitrary customer from the massfunction of the buffer occuopancy during an arbitrary slot. It is exactly the relationship found in [5] after a much more difficult reasoning and under more restrictive conditions.

# 5 Finite buffer queues

The relationships between the probability distributions of the buffer occupancy and the customer delay derived in the previous sections are applicable to any work conserving discrete-time FIFO-queue, whereby the service times of the customers are equal to 1 slot each, regardless of the arrival process. Essentially, the only requirements for the queue are that its buffer occupancy does not grow boundless, such that the limits in (2) and (3) exist and the concept that figure in the relationships can be defined. In case we have a finite buffer this is assured regardless the load offered to the system. For finite buffers the results remain valid as long as $\rho$ is calculated from the sequence of *effective* arrivals $(A(t_k))$ in the queue, i.e. as long as $\rho$ indicates the average *carried* load (rather than the *offered* load).

It is even possible to relax the condition on the load a little further. The result in (9) and (17) applies regardless of the offered load, because it also remains valid for arrival processes which yield a carried load $\rho = 1$ (a carried load greater than 1 is physically impossible). In such case, we may have sample paths that do never empty — in the multiple server case, with an equivalent single-server queue that never empties — anymore after a last empty slot and customers that have no service slot under LIFO discipline, although they do have a position on arrival. The number of such customers, however, is always finite (and bounded by the buffer size), so that their contributions to the probabilities defined in (3) vanish, when taking the limit for $m \to \infty$. The probabilities are then determined uniquely by the customers who do have a LIFO service-slot, and for whom the argument remains unaltered.

# 6 Applicability

The relationship in (9) and (17) is especially useful in the analysis of all kinds of queueing systems occurring in ATM networks, such as ATM multiplexers or output queues in ATM switches, due to the constant length of the ATM cells (i.e. deterministic service times) and the usual FIFO serving rule in ATM. In recent years many researchers have analysed the buffer-contents distribution (and associated quantitites such as cell loss ratio) in ATM queues with various types of bursty arrival models (such as on/off models, Markov modulated models, periodic arrival models, train arrival models, etc.) using various techniques of analysis (i.e. analytical or semi-analytical approaches, numerical solutions, computer simulations). All their results can be transformed into corresponding results for the delay characteristics of these queues by means of the simple relationship established here.

## References

[1] M.J. Karol, M.G. Hluchyj, and S.P. Morgan. Input versus output queueing on a space-division packet switch. *IEEE Transactions on Communications*, COM-35(12):1347–1356, 1987.

[2] K. Sohraby. Delay analysis of a single server queue with Poisson cluster arrival process arising in ATM networks. *Proc. IEEE Globecom '89 Conf.*, Texas, 1989, pp. 611–616.

[3] H. Bruneel and B.G. Kim. *Discrete-Time Models for Communication Systems Including ATM.* Kluwer Academic Publishers, Boston, 1993.

[4] Y. Xiong and H. Bruneel. Buffer contents and delay for statistical multiplexers with fixed-length packet-train arrivals. *Performance Evaluation*, 17:31–42, 1993.

[5] H. Bruneel, B. Steyaert, E. Desmet and G.H. Petit. An analytical technique for the derivation of the delay performance of ATM switches with multiserver output queue. *International Journal of Digital and Analog Communication System*, 5:193–201, 1992.