(will be inserted by the editor)

Delay analysis of two batch-service queueing models with batch arrivals: $Geo^X/Geo^c/1$

Dieter Claeys, Joris Walraevens, Koenraad Laevens and Herwig Bruneel

SMACS Research Group, Department TELIN, Ghent University
Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
e-mail: {dclaeys, jw, kl, hb}@telin.ugent.be

The date of receipt and acceptance will be inserted by the editor

Abstract In this paper, we compute the probability generating functions (PGF's) of the customer delay for two batch-service queueing models with batch arrivals. In the first model, the available server starts a new service whenever the system is not empty (without waiting to fill the capacity), while the server waits until he can serve at full capacity in the second model. Moments can then be obtained from these PGF's, through which we study and compare both systems. We pay special attention to the influence of the distribution of the arrival batch sizes. The main observation is that the difference between the two policies depends highly on this distribution. Another conclusion is that the results are considerably different as compared to Bernoulli (single) arrivals, which are frequently considered in the literature. This demonstrates the necessity of modeling the arrivals as batches.

Keywords: batch service, batch arrivals, immediate versus full-batch service policy, customer delay

MSC classification 60K25, 68M20, 90B22

1 Introduction

Servers capable of processing several customers simultaneously are generally referred to as batch servers. Examples include elevators in high buildings, transport vehicles, ship locks, ... Furthermore, in telecommunications, it is often the case that information packets are grouped in larger entities (batches) and these batches are transmitted instead of all packets individually. This is mainly done for efficiency reasons, since only one header per aggregated batch has to be constructed instead of one header per single information unit. Optical burst switched (OBS) networks, for instance, apply this method abundantly (Chen et al., 2004; Qiao and Yoo, 1999). At the edges, IP packets with the same destination and Quality of Service (QoS) requirements are aggregated into optical bursts which are injected into the network.

The model that resembles batch service the most is a multiserver system, as they share the feature that several customers can be served simultaneously (we call the maximum number c the capacity). However, batch-service systems might be less performant: if a customer arrives when the batch server is processing less than c customers, this customer cannot join the ongoing service, whereas the customer would be served immediately by one of the available servers in the multiserver system. In view of this, the batch server has to make a decision, called the service policy, when he becomes available and finds less than c customers in the queue. The server could, for instance, start serving the already present customers immediately (Chang and Choi, 2005; Chaudhry and Templeton, 1983; Janssen and van Leeuwaarden, 2005; Zhao and Campbell, 1996). Although the present customers benefit from this approach, capacity is wasted: customers that arrive later cannot join the ongoing service.

An alternative for this so-called **immediate-batch** service policy (IBSP) is the **full-batch** service policy (FBSP) (Chang and Takine, 2005; Chaudhry and Templeton, 1983). In this case, the available server only starts operating if the system contains at least as many customers as his capacity, which, in turn, has a negative effect on the delay of the customers waiting to form a full batch. Hence, the choice whether to wait (FBSP) or not (IBSP) is important in order to optimize the performance of the batch-serving system, for instance in terms of the (mean) customer delay.

In this paper, we consider two discrete-time batch-service models which only differ in their service policy. In the first, the IBSP is adopted while the FBSP is employed in the second. These models are described in section 2. We then compute, for both models separately, the probability generating function (PGF) of the customer delay: section 3 covers the IBSP while the FBSP is dealt with in section 4. Several moments can be extracted from the obtained PGF's, which serve as performance measures in section 5 through which the two systems are compared.

The analyses of the delays in sections 3 and 4 form the first major contribution of this paper. Especially, the inclusion of *batch arrivals* (i.e. several customers can arrive simultaneously) in the queueing model and delay analysis is novel (it has not been done for discrete-time or continuous-time models, as far as we know). Delay analysis for single arrivals and batch service has been covered in e.g. Chaudhry and Templeton (1983); Downton (1955); Kim and Chaudhry (2006); Medhi (1975), but the batch-arrival nature of the models in our paper, makes the delay analysis much more involved. In our conference paper Claeys et al. (2008), we were the first to study the customer delay in a preliminary batch-arrival, batch-service queueing model. In that paper, we considered service times of one slot and the server applies the FBSP. We here extend this to the more general case of geometrically distributed service times and the analysis of both the IBSP and the FBSP system. The approach, based on PGF's, is promising because it leads to semi-analytical expressions, whereas methods based on matrix-analytical techniques would require more numerical calculations. The resulting expressions give us the tools to compare the two, practically important batch server policies in a quite general setting (service times of more than one slot, general number of arrivals per slot). Herein lies a second contribution of this paper: we are able to study the influence of the distribution of the number of customer arrivals in a slot on the performance of both systems. This leads to the important observation that the results differ considerably as compared to single (Bernoulli) arrivals and that the difference between both policies is highly effected by the distribution of the number of arrivals in a batch. Hence, we conclude that the inclusion of batch arrivals in the queueing model is a necessity and this paper provides this analysis.

2 Description of the models

Before describing the models, we first make the following notational convention: we denote the mass function of a discrete random variable X by $x(n)$. Hence, $x(n) \triangleq \Pr[X = n], n \geq 0$. $X(z) \triangleq \mathrm{E}\left[z^X\right] = \sum_{n=0}^{\infty} x(n) z^n$ is the corresponding PGF.

In this paper, we study two models, which only differ in their service policy. The other properties are identical for both models:

- The time axis is divided into fixed-length slots.
- The amount of customers that arrive during slot k is denoted by A_k and A_k can be larger than one (we call this batch arrivals). The numbers of customer arrivals during consecutive slots constitute an independent and identically distributed (IID) process, with common mass function $a(n)$ and PGF $A(z)$.
- The queue is infinitely large.
- There is one batch server of capacity c (c is a constant), meaning that the server can process up to c customers simultaneously.
- A service period is the period between the start and end of the service of one batch of customers. The service can only start and end at slot

boundaries, implying that an arriving customer has to wait for service at least until the next slot mark. This part is not included in the customer delay (also called waiting time): we denote the customer delay as the number of slots between the end of the customer's arrival slot and the beginning of the slot whereat the service of the customer is effectively started.
- The consecutive service times - a service time is the length of a service period, expressed in a number of slots - are IID and have a geometric distribution, with mass function $s(n) = (1-\alpha)\alpha^{n-1}(n \geq 1)$, whereby α $(0 \leq \alpha < 1)$ represents the probability that an ongoing service during a random slot is not finished at the end of that slot. The corresponding PGF is equal to $S(z) = (1-\alpha)z/(1-\alpha z)$, and the mean service time $\mathrm{E}[S]$ equals $1/(1-\alpha)$ slots. Note that in some papers the service time is included in the customer delay, while it is not here. However, one can include the service time by multiplying the PGF of the delay by $S(z)$.
- The queueing discipline is first-come-first-served (FCFS).

In case of IBSP, the server starts another service if the system is not empty upon the server becoming available (without waiting to fill the capacity). In case of FBSP, the server waits to start processing until the beginning of the first slot at which at least c customers have accumulated in the system. At this moment, the server starts the service of exactly c customers.
The stability condition for both models is that the load $\rho \triangleq \frac{\lambda}{c(1-\alpha)} < 1$, whereby λ represents the mean number of arrivals in a random slot.

3 IBSP

As mentioned above, this section deals with the PGF of the customer delay ($\tilde{W}(z)$) for the IBSP model. Let us concentrate on an arbitrary tagged customer T arriving in say slot J. The number of other arrivals in slot J and before customer T is denoted by B and $\tilde{U}$ is the system content (the sum of the number of customers in the queue and those in the server) at slot mark J. Note that in the case of single arrivals, $B = 0$, which would simplify the analysis. Depending on the value of $\tilde{U}$, we consider two cases:

- $\tilde{U} = 0$: the server is idle during slot J. At slot mark $J+1$, B customers are ahead of customer T. The delay of T then consists of $\left\lfloor \frac{B}{c} \right\rfloor$ service periods, with $\lfloor . \rfloor$ the floor function, i.e. $\lfloor x \rfloor = \max\{n \in \mathbb{N} \mid n \leq x\}$.
- $\tilde{U} \geq 1$: the server is working during slot J. We observe two subcases:
  - If the service period ends at the end of slot J (with probability $1-\alpha$), the delay consists of $\left\lfloor \frac{\tilde{Q}+B}{c} \right\rfloor$ service periods, with $\tilde{Q}$ the number of customers in the queue (without the server) at slot mark J.

– If the service period does not end at the end of slot J (with probability α), the delay consists of $1+\left\lfloor \frac{\tilde{Q}+B}{c} \right\rfloor$ service periods. We indeed can consider the service period that continues during slot $J+1$ as a new service period, due to the memoryless character of the geometric distribution of the service periods, but we have to keep in mind that the server not necessarily serves c customers during slot $J+1$. Note that because of the geometric distribution of the service times, we do not have to keep track of the remaining service time. This would not be the case for a general distribution, which would complicate the analysis considerably.

The combination of these cases produces

$$\begin{aligned}\tilde{W}(z) = \tilde{U}(0)\mathrm{E}\left[z^{\sum_{i=1}^{\lfloor B/c \rfloor} S_i}\right] + (1-\alpha)\mathrm{E}\left[z^{\sum_{i=1}^{\lfloor (\tilde{Q}+B)/c \rfloor} S_i}\right] \\ -(1-\alpha)\mathrm{E}\left[z^{\sum_{i=1}^{\lfloor (\tilde{Q}+B)/c \rfloor} S_i} \left\{\tilde{U}=0\right\}\right] + \alpha\mathrm{E}\left[z^{\sum_{i=1}^{1+\lfloor (\tilde{Q}+B)/c \rfloor} S_i}\right] \\ -\alpha\mathrm{E}\left[z^{\sum_{i=1}^{1+\lfloor (\tilde{Q}+B)/c \rfloor} S_i} \left\{\tilde{U}=0\right\}\right] \quad ,\end{aligned}$$

with $\mathrm{E}\left[z^x\{\text{condition}\}\right] \triangleq \mathrm{E}\left[z^x|\text{condition}\right] \Pr\left[\text{condition}\right]$ and S_i the length of the i^{th} service period after slot J. Taking into account the IID nature of the service times and that $\tilde{U}=0 \Rightarrow \tilde{Q}=0$, yields:

$$\begin{aligned}\tilde{W}(z) = \tilde{U}(0)\alpha\left[1-S(z)\right]\sum_{i=0}^{\infty}\sum_{j=0}^{c-1} b(ic+j)S(z)^i \\ +\left[1-\alpha+\alpha S(z)\right]\sum_{i=0}^{\infty}\sum_{j=0}^{c-1} h(ic+j)S(z)^i \quad ,\end{aligned}$$

with $H \triangleq \tilde{Q}+B$. The IID customer arrivals further imply that B is independent of $\tilde{Q}$, so that $H(z)=B(z)\tilde{Q}(z)$.

In the next step, we relate $\tilde{W}(z)$ with $B(z)$, $\tilde{Q}(z)$ and $\tilde{U}(0)$. We therefore make use of the following relation between the Kronecker delta function $\delta(.)$ ($\delta(.)$ is 1 when its argument is zero and 0 otherwise) and the c complex c^{th} roots of unity ($\varepsilon_m \triangleq e^{\imath 2\pi m/c}, 0 \le m \le c-1$, with $\imath$ the imaginary unit):

$$\delta(j-r) = \frac{1}{c}\sum_{m=0}^{c-1} \varepsilon_m^{ic+j-r} \quad , \quad \forall i \in \mathbb{N} \; .$$

We also introduce the function $u(z)$, defined as $u(z) \triangleq |S(z)|^{1/c} e^{\imath \mathrm{Arg}\left(S(z)\right)/c}$, with $|z|$ the absolute value of z and $\mathrm{Arg}(z)$ the principal value of the argument of z (i.e. it is a mapping in the interval $]-\pi,\pi]$). We then obtain subsequently:

$$
\begin{aligned}
&\tilde{W}(z)\\
&= \tilde{U}(0)\alpha\,[1-S(z)]\sum_{i=0}^{\infty}\sum_{j=0}^{c-1}\sum_{r=0}^{c-1} b(ic+j)u(z)^{ic+j-r}\delta(j-r)\\
&\quad + [1-\alpha+\alpha S(z)]\sum_{i=0}^{\infty}\sum_{j=0}^{c-1}\sum_{r=0}^{c-1} h(ic+j)u(z)^{ic+j-r}\delta(j-r)\\
&= \tilde{U}(0)\alpha\,[1-S(z)]\sum_{i=0}^{\infty}\sum_{j=0}^{c-1}\sum_{r=0}^{c-1} b(ic+j)u(z)^{ic+j-r}\sum_{m=0}^{c-1}\frac{1}{c}\epsilon_m^{ic+j-r}\\
&\quad + [1-\alpha+\alpha S(z)]\sum_{i=0}^{\infty}\sum_{j=0}^{c-1}\sum_{r=0}^{c-1} h(ic+j)u(z)^{ic+j-r}\sum_{m=0}^{c-1}\frac{1}{c}\epsilon_m^{ic+j-r}\\
&= \frac{\tilde{U}(0)\alpha}{c}\,[1-S(z)]\sum_{m=0}^{c-1} B(u(z)\varepsilon_m)\frac{[u(z)\varepsilon_m]^{-c}-1}{[u(z)\varepsilon_m]^{-1}-1}\\
&\quad + \frac{[1-\alpha+\alpha S(z)]}{c}\sum_{m=0}^{c-1} B(u(z)\varepsilon_m)\tilde{Q}(u(z)\varepsilon_m)\frac{[u(z)\varepsilon_m]^{-c}-1}{[u(z)\varepsilon_m]^{-1}-1}\\
&= \frac{1-u(z)^c}{cu(z)^c}\sum_{m=0}^{c-1} B\left(u(z)\epsilon_m\right)\frac{u(z)\epsilon_m}{1-u(z)\epsilon_m}\\
&\quad \left\{\tilde{U}(0)\alpha\,[1-S(z)] + [1-\alpha+\alpha S(z)]\,\tilde{Q}(u(z)\epsilon_m)\right\}\ . \qquad (1)
\end{aligned}
$$

Equation (1) thus gives us the required relation between $\tilde{W}(z)$, $B(z)$, $\tilde{Q}(z)$ and $\tilde{U}(0)$. $B(z)$ can be found by taking into account that the probability of having k arrivals during slot J is proportional to both $a(k)$ and the number k itself, because T can be any of the k arrivals in a slot during which k arrivals occur (see e.g. Bruneel (1983)), and that the position of T in a slot of k arrivals is uniformly distributed between 1 and k, leading to

$$
b(n) = \sum_{k=n+1}^{\infty}\frac{1}{k}\frac{a(k)k}{\lambda},\ \ n \geq 1; \qquad B(z) = \frac{A(z)-1}{\lambda(z-1)}\ .
$$

Now it only remains to determine $\tilde{U}(0)$ and $\tilde{Q}(z)$ in (1). First, due to the IID nature of the customer arrivals, $\tilde{Q}$ and $\tilde{U}$ are equally distributed as the queue, respectively system, content at a random slot mark. Secondly, in Claeys et al. (2007), we have studied the system content $\tilde{U}$ at a random slot mark in a more general model with general service times and a lower limit l on the system content before starting a service ($l = 1$ corresponds with IBSP). For the particular model of this paper it leads to the following expression:

$$\tilde{U}(z) = \frac{S(A(z))}{z^c - S(A(z))}\left[u_0(z^c-1) + \frac{S(A(z))-1}{\mathrm{E}\left[S\right](A(z)-1)}\sum_{n=1}^{c-1} u_n(z^c - z^n)\right] \ , \quad (2)$$

whereby the remaining unknowns u_n, $0 \leq n \leq c-1$, are calculated by solving the following set of linear equations:

$$\begin{cases} u_0(z_i^c - 1) + \frac{S(A(z_i))-1}{\mathrm{E}[S](A(z_i))-1)}\sum_{n=1}^{c-1} u_n(z_i^c - z_i^n) = 0 \ , \ \ 1 \leq i \leq c-1, \\ u_0 c + \sum_{n=1}^{c-1} u_n(c-n) = c - \mathrm{E}\left[S\right]\lambda. \end{cases}$$

$\tilde{U}(0)$ is then found by substituting z by 0 in (2): $\tilde{U}(0) = u_0$. Furthermore, this queueing model satisfies the conditions in Kim et al. (2002), so that $\tilde{Q}(z) = \tilde{U}(z)/S(A(z))$. Mark that we comment on an alternative calculation of $\tilde{U}(z)$ in appendix A.

4 FBSP

In this section, we compute the PGF of the customer delay ($W(z)$) for FBSP. We again consider an arbitrary tagged customer T arriving during slot J. As opposed to multiserver queues and IBSP, the delay that customer T experiences in this case consists of two components; the first (we denote its length by W_1) is the time required to serve batches of 'older' customers. The second part (with length W_2) is the time needed, starting from the end of the first waiting time, to completely fill the batch containing the tagged customer. Several quantities influence W_1 and W_2:

- the number of customers M in the system in front of T at the beginning of slot $J+1$,
- the number of customers F in the system behind T at slot mark $J+1$,
- the number of customers G in front of the tagged customer at the end of the first waiting time,
- the number of customers H behind the tagged customer at the end of the first waiting time, and
- the system content $P \triangleq G+1+H$ at the end of the first waiting time.

Let us consider the example depicted in Fig. 1 to clarify W_1, W_2, M, F, G, H, P and the relations between them. Assume that $c = 10$, $M = 34$ and $F = 2$. Hence, the first waiting time equals 3 service times (due to the memoryless character of the geometric distribution, we can consider the potential service period that continues at the end of slot J as a service period starting at slot $J+1$). After this first waiting time, G equals 4 and H is the sum of F and the number of arrivals during the first waiting time.

Hence, the system content at this moment, P, equals $G+1+H$. If $P \geq 10$, the server processes the batch with the tagged customer immediately and W_2 is equal to zero. Otherwise (which is the case in this example), W_2 equals the time until the beginning of the first slot whereby the system content is at least 10.

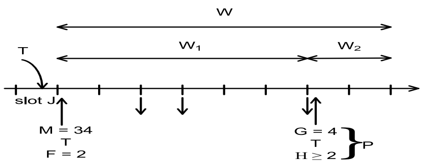


Fig. 1. Illustration of W_1, W_2, M, F, G, H, P and the relations between them

As a first step in the analysis, we compute the joint PGF of M and F. Define B (X) as the number of customer arrivals during slot J and before (after) the tagged customer (note again that for single arrivals $B=0$ and $X=0$). Then it is obvious that $F=X$ and that $M=Y+B$, with Y the number of customers ahead of T at the beginning of slot $J+1$ that were already in the system at the beginning of slot J. The following relation between Y and the system content at the beginning of slot J (U) holds:

$$Y = \begin{cases} U-c & \text{if } U \geq c \text{ and the ongoing service} \\ & \text{period ends at the end of slot } J \\ & \\ U & \text{else} \end{cases} .$$

Due to the IID arrivals, U is equally distributed as the system content at the beginning of a random slot in the steady state. Furthermore, the foregoing relation equals the slot-by-slot evolution of the system content (see appendix A) except without A_J, the number of arrivals in slot J. Consequently, $Y(z)$ equals $U(z)/A(z)$. Since Y is independent of B and X, we obtain:

$$D(z,x) \triangleq \mathrm{E}\left[z^M x^F\right] = \frac{U(z)}{A(z)} \mathrm{E}\left[z^B x^X\right] \quad .$$

As in section 3, $\mathrm{E}\left[z^B x^X\right]$ can be obtained by taking into account that an arbitrary customer is more likely to arrive in a slot with more customer arrivals (Bruneel, 1983): $\Pr\left[B=n, X=m\right] = a(n+m+1)/\lambda$, leading to

$$D(z,x) = \frac{U(z)}{A(z)} \frac{A(z)-A(x)}{\lambda(z-x)} \quad . \tag{3}$$

We denote the corresponding mass function by $d(n,m) \triangleq \Pr[M=n, F=m]$. As mentioned before, the G customers in front of T at the end of the first waiting time are served in the same batch as the tagged customer. We have

$$W_1 = \left\lfloor \frac{M}{c} \right\rfloor \text{service times} \;\;, \quad G = M \bmod c \;\;, \quad H = F + \sum_{j=1}^{W_1} A_{J+j} \;\;,$$

with 'mod' the modulo operator. As a next step, we compute the joint PGF of W_1, G and H:

$$W(z,x,y) \triangleq \mathrm{E}\left[z^{W_1} x^G y^H\right] = \sum_{n=0}^{\infty} \sum_{l=0}^{c-1} \sum_{k=0}^{\infty} d(nc+l,k) S(zA(y))^n x^l y^k \;\; .$$

In order to relate $W(z,x,y)$ with $D(z,x)$, we define the function $u(z,y)$ as

$$u(z,y) \triangleq |S(zA(y))|^{1/c} e^{\imath \mathrm{Arg}(S(zA(y)))/c} \;\; .$$

We relate $W(z,x,y)$ with $D(z,x)$ along the same lines as we found (1), leading to:

$$W(z,x,y) = \frac{1}{c} \sum_{m=0}^{c-1} D(u(z,y)\varepsilon_m, y) \frac{u(z,y)^c - x^c}{u(z,y)\varepsilon_m - x} \frac{u(z,y)\varepsilon_m}{u(z,y)^c} \;\; . \qquad (4)$$

As mentioned in the above example, the system content at the end of the first waiting time, P, determines whether W_2 is zero or not. Since $P = G + 1 + H$, the joint PGF of W_1 and P equals

$$\mathrm{E}\left[z^{W_1} x^P\right] = x W(z,x,x) \;\; . \qquad (5)$$

Now, we focus on W_2, starting from the following property:

$$\Pr[W_2 > m | P = p] = \Pr[p + \tilde{A}_1 + \cdots + \tilde{A}_m < c] \quad m \geq 0 \;\; , \qquad (6)$$

with $\tilde{A}_j$ the number of arrivals during the j^{th} slot after the first waiting time. Multiplying both sides of (6) by t^m and summing over all m yields

$$\begin{aligned} \frac{\mathrm{E}\left[t^{W_2} | P = p\right] - 1}{t-1} &= \sum_{m=0}^{\infty} t^m \sum_{n=0}^{c-1} \Pr\left[p + \tilde{A}_1 + \cdots + \tilde{A}_m = n\right] \\ &= \sum_{m=0}^{\infty} t^m \sum_{n=0}^{c-1} \frac{1}{n!} \frac{\partial^n}{\partial x^n} x^p A(x)^m \bigg|_{x=0} \\ &= \sum_{n=0}^{c-1} \frac{1}{n!} \frac{\partial^n}{\partial x^n} \frac{x^p}{1 - tA(x)} \bigg|_{x=0} \;\; . \qquad (7) \end{aligned}$$

Mark that the second step follows from the probability generating property of PGF's and the fact that $x^p A(x)^m$ is the PGF of $p + \tilde{A}_1 + \cdots + \tilde{A}_m$

(due to the IID arrival process). Furthermore, the last step requires that $|tA(x)| < 1$ about $x = 0$. Note that if $A(z) = 1 - \lambda + \lambda z$ - i.e. the single arrival case -, the previous analysis for $\mathrm{E}\left[t^{W_2}|P = p\right]$ is not necessary, as W_2 given $P = p$ is then the sum of $(c-p)^+$ geometrically distributed interarrival times. Expression (7), together with the notion that W_1 only influences W_2 through P, enables us to compute the joint PGF of W_1 and W_2:

$$\begin{aligned}
\mathrm{E}\left[z^{W_1}t^{W_2}\right] &= \sum_{p=1}^{\infty} \Pr\left[P = p\right] \mathrm{E}\left[z^{W_1}t^{W_2}|P = p\right] \\
&= \sum_{p=1}^{\infty} \Pr\left[P = p\right] \mathrm{E}\left[z^{W_1}|P = p\right] \left\{ 1 + (t-1) \sum_{n=0}^{c-1} \frac{1}{n!} \frac{\partial^n}{\partial x^n} \frac{x^p}{1 - tA(x)}\bigg|_{x=0} \right\} \\
&= \mathrm{E}\left[z^{W_1}\right] + (t-1) \sum_{n=0}^{c-1} \frac{1}{n!} \frac{\partial^n}{\partial x^n} \frac{\mathrm{E}\left[z^{W_1}x^P\right]}{1 - tA(x)}\Bigg|_{x=0} \quad . \qquad (8)
\end{aligned}$$

The PGF of the total time that customers remain in the queue before they are served can finally be obtained by putting $t = z$ in (8):

$$W(z) = \mathrm{E}\left[z^{W_1}\right] + (z-1) \sum_{n=0}^{c-1} \frac{1}{n!} \frac{\partial^n}{\partial x^n} \frac{\mathrm{E}\left[z^{W_1}x^P\right]}{1 - zA(x)}\Bigg|_{x=0} \quad . \qquad (9)$$

$\mathrm{E}\left[z^{W_1}\right]$ can be deduced by substituting x by 1 in (5):

$$\mathrm{E}\left[z^{W_1}\right] = \frac{1}{c} \sum_{m=0}^{c-1} D(u(z,1)\varepsilon_m, 1) \frac{u(z,1)^c - 1}{u(z,1)\varepsilon_m - 1} \frac{u(z,1)\varepsilon_m}{u(z,1)^c} \quad . \qquad (10)$$

Substitution of (10), (3), (4) and (5) in (9) yields an expression for the PGF of the delay as a function of $U(z)$, which can be found through the earlier mentioned, more general result in Claeys et al. (2007) (FBSP corresponds with $l = c$):

$$U(z) = \frac{(z^c - 1)S(A(z)) \sum_{n=0}^{c-1} u(n)z^n}{z^c - S(A(z))} \quad .$$

Alternatively, $U(z)$ can be obtained for this particular model as explained in appendix A. Further, the set of equations to determine the $u(n)$'s are given in equation (11) in appendix A.

5 Behaviors of the queueing systems

The PGF's from the previous sections enable us to compute moments of the customer delay by taking derivatives at $z = 1$. In this section, we study and compare the means and variances for both systems. As the inclusion of batch arrivals is the main contribution in the delay analysis, we consider, next to the Bernoulli (single) arrivals, several batch arrival processes and investigate the influence of these arrival processes on the behaviors of both systems.

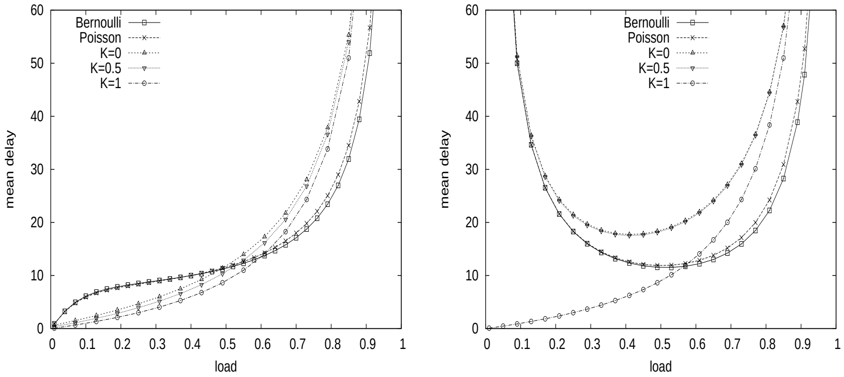
The mean and variance of the customer delay are depicted versus the load in respectively Fig. 2 and Fig. 3, for some distributions of the number of customer arrivals in a slot: Bernoulli ($a(0) = 1 - \lambda$, $a(1) = \lambda$), Poisson ($a(n) = e^{-\lambda}\lambda^n/(n!)$) and several distributions whereby the non-zero mass is centered symmetrically around c: $a(0) = (c - \lambda)/c$, $a(c-1) = a(c+1) = \lambda(1-K)/(2c)$, $a(c) = K\lambda/c$; $K(0 \leq K \leq 1)$ is a measure for the dispersion around c. We consider $K = 0$, $K = 0.5$ and $K = 1$. Note that when $K = 1$ either 0 or c customers arrive, while for $K = 0$, zero, $c - 1$ or $c + 1$ arrivals can occur in a slot. Both figures consist of two panes whereby the left corresponds to IBSP and the right to FBSP. The server capacity and the mean service time are both equal to 10. Hence, an increasing load is caused by an increasing λ. The figures exhibit that, regardless of the distribution of the number of per-slot arrivals, the batch-service queueing systems with IBSP behave to some extent similar as traditional systems: the mean and the variance of the customer delay are increasing functions of the load and reach a vertical asymptote for $\rho \to 1$. In case of FBSP, the mean and variance go to infinity when the load tends to zero (except for $K = 1$). This is the effect of the second waiting time (W_2), which goes to infinity for $\rho \to 0$. When the load increases, W_2 decreases, leading to a shorter customer delay. When the load increases further, the first waiting time W_1 becomes long, which implies that the customer delay increases again. These effects produce the U-shapes of the curves. Note that the case where the (non-zero) arriving batch size equals c is the sole exception, since in this case the service batch size is perfectly adapted to the size of arriving batches. More generally, the second waiting time vanishes for any arrival distribution whereby the number of customer arrivals can only be multiples of c.

While we discussed the influence of the mean arrival rate λ in the previous paragraph, we now discuss the influence of the entire distribution of the number of customer arrivals in a slot. Fig. 2 and Fig. 3 show that the systems behave nearly equally for Poisson arrivals as for Bernoulli arrivals, whereas they behave rather different for the other distributions.

Let us now take a look at the relative difference between the delay in case of IBSP and in case of FBSP. It is defined as

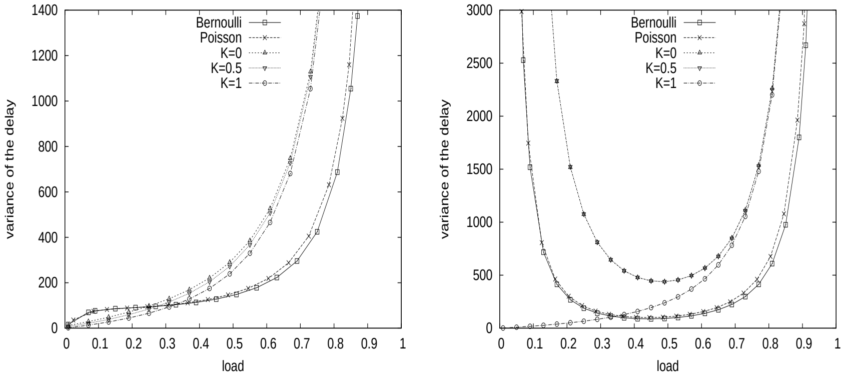
$$\left(\mathrm{E}\left[\tilde{W}\right] - \mathrm{E}\left[W\right]\right) / \left(\left[\mathrm{E}\left[\tilde{W}\right] + \mathrm{E}\left[W\right]\right] /2\right) \ ,$$

for the mean, and analogously for the variance. This implies that when the relative difference is negative (positive), IBSP (FBSP) performs better. Fig. 4 shows that IBSP is the best choice for a low load, regardless of the arrival pattern. Indeed, the long second waiting times are avoided in this



(a) IBSP (b) FBSP

Fig. 2. Mean customer delay versus the load for several distributions of the number of customer arrivals in an arbitrary slot; $c = 10$ and $\mathrm{E}[S] = 10$

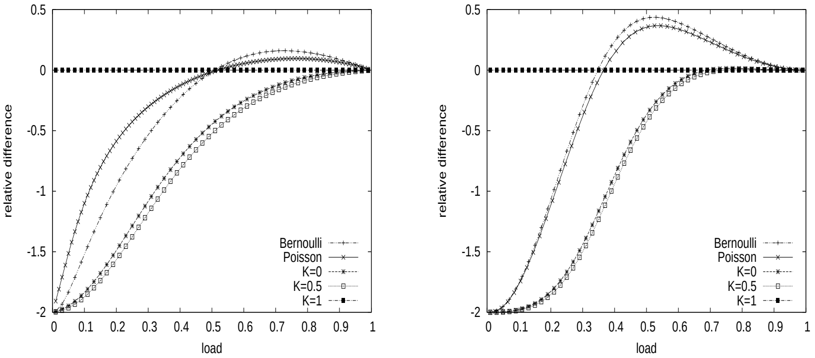


(a) IBSP (b) FBSP

Fig. 3. Variance of the customer delay versus the load for several distributions of the number of customer arrivals in an arbitrary slot; $c = 10$ and $\mathrm{E}[S] = 10$

case. When the load becomes high, the FBSP becomes considerably better in case of the Bernoulli and the Poisson distribution. This is what we expect intuitively since, when customers arrive frequently, the second waiting time is small compared to the gain of serving at full capacity. However, this does not hold for every distribution: for $K = 1$, IBSP and FBSP behave equally; for $K = 0$ and $K = 0.5$, the relative difference in the mean delay remains negative and the relative difference in the variance becomes only marginally positive. This is caused by the bursty nature of the latter distributions: no customers arrive during long times followed by a large mass of arrivals during one slot, so that the performance loss of not serving at full capacity is small compared to the second waiting times. Hence, we advise to adopt the

IBSP when the arrival process is bursty.

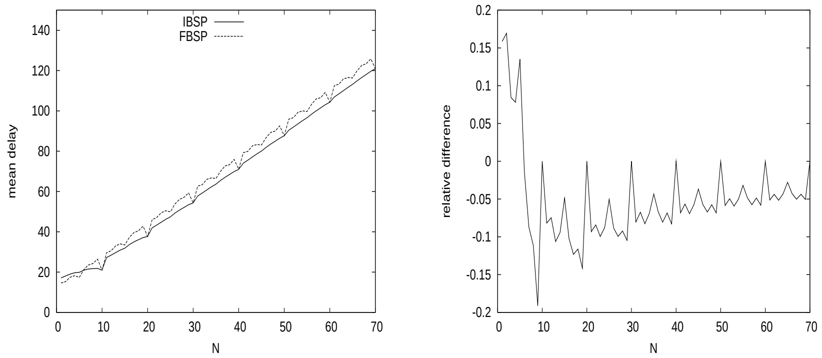


(a) $\mathrm{E}\left[W\right]$ (b) $\mathrm{Var}\left[W\right]$

Fig. 4. Relative difference versus the load

So far, we have considered the following arrival distributions: Bernoulli, Poisson and a class of distributions whereby the non-zero mass is centered symmetrically around c. We now consider another arrival process to further investigate the influence of the batch size, namely the batch Bernoulli arrival process, i.e. $a(0) = 1 - \lambda/N$, $a(N) = \lambda/N$, and investigate the effect of the batch size N on the mean customer delay. Therefore, we plot, in Fig. 5, the mean customer delay for both policies (pane a) and the relative difference in their mean values (part b) as a function of N. We assume that $\lambda = 0.7$, $\mathrm{E}\left[S\right] = 10$ and $c = 10$. First, we observe that the mean delay increases as a function of N for both policies 'on the larger' N-scale. Indeed, if N increases, the arrival process becomes more bursty and it is generally known (Bruneel, 1985) that burstiness deteriorates the performance. However, the curves exhibit some clefts when $N = kc, k \in \mathbb{N}$. This is because the server capacity is perfectly adapted to the arrival batch sizes in these cases. There are also smaller clefts when $N = kc + c/2$ or when N is even: the server capacity is reasonably adapted to the batch arrival sizes but not as good as when $N = kc$. There thus exists a pattern which repeats itself when $N = kc$, $k = \{2, 3, 4, \ldots\}$.

Let us now take a closer look at the relative difference between both policies. When $N = 1$, FBSP is better, which is in accordance with the results for the Bernoulli arrivals for $\rho = 0.7$. When N increases, this difference becomes smaller 'on the large scale' and from $N = 6$ on, IBSP is better. Indeed, we previously concluded that bursty arrivals favour IBSP. Further, we observe that the relative difference equals 0 when $N = kc$, since no second waiting time exists here in case of FBSP and the server always processes at full capacity for IBSP. The figure also shows that IBSP and FBSP differ less when c is better adapted to N. There is thus also a repetitive pattern with period c. Finally, the curve of the relative difference smooths down as N increases. Indeed, for N large, the fraction of customers in an arrival batch that experience a second waiting time becomes very small, and the relative difference of both policies converges to 0 for $N \rightarrow \infty$.



(a) For both policies (b) Relative difference

Fig. 5. Mean customer delay versus N; $\rho = 0.7$, $c = 10$ and $\mathrm{E}[S] = 10$

6 Conclusions

In this paper, we have studied two batch-service queueing models which differ in their service policy. In the immediate-batch service policy (IBSP), the available server starts processing whenever the system is not empty, while in the case of full-batch service policy (FBSP) the server waits until he can serve at full capacity. We have computed, for both models separately, the probability generating function (PGF) of the customer delay. These PGF's enable us to obtain several moments which, in turn, provide a tool to compare both policies in practical situations.

Both models share the common feature that the service times are geometrically distributed. Throughout the analysis, we have made use of the memoryless character of this distribution, which simplifies the analysis. Further research is necessary to deal with generally distributed service times.

Our main contribution is the combination of batch arrivals and delay analysis. To the best of the authors' knowledge, all previous papers concerning the customer delay consider single arrivals, which simplifies the analysis considerably. Through some numerical examples, we clearly show that the results can differ significantly from those in the Bernoulli case. We have finally compared both models comprehensivily and we conclude that the distribution of the number of customer arrivals in an arbitrary slot plays a significant role in how both policies compare to each other. Hence, the inclusion of batch arrivals in the model is a necessity, which the analysis in this paper provides.

Acknowledgment

The second author is a Postdoctoral Fellow with the Research Foundation F.W.O.- Flanders, Belgium.

A Alternative calculation of the PGF's of the system content

A.1 FBSP: $U(z)$

As a first step, we express the system content at slot boundary $k+1$ (U_{k+1}) in terms of the system content at slot mark k (U_k):

$$U_{k+1} = \begin{cases} U_k + A_k & \text{if } U_k < c \\ U_k + A_k & \text{if } U_k \geq c \text{ and the ongoing service period} \\ & \text{continues at the end of slot } k \\ U_k + A_k - c & \text{if } U_k \geq c \text{ and the ongoing service period} \\ & \text{ends at the end of slot } k \end{cases} .$$

Next, this relation is translated into PGF's, leading to

$$U_{k+1}(z) = A(z) \left[\sum_{n=0}^{c-1} u_k(n) z^n + \alpha \sum_{n=c}^{\infty} u_k(n) z^n + (1-\alpha) \sum_{n=c}^{\infty} u_k(n) z^{n-c} \right] .$$

Assuming steady state, the limit $k \to \infty$ is taken, with $\lim_{k\to\infty} U_k(z) = \lim_{k\to\infty} U_{k+1}(z) = U(z)$ and $\lim_{k\to\infty} u_k(n) = u(n)$. This, together with some standard z-transform techniques, produces:

$$U(z) = \frac{(z^c - 1) S(A(z)) \sum_{n=0}^{c-1} u(n) z^n}{z^c - S(A(z))} .$$

Finally, the c unknowns $u(n)$ are determined. As the denominator $z^c - S(A(z))$ has c zeroes $z_0 = 1, z_1, \ldots, z_{c-1}$ in the closed complex unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$ (this can be proved by means of Rouché's theorem, see e.g. Klimenok (2001)), and since PGF's are normalized and bounded in this area, the unknowns are the solution of the following set of c linear equations:

$$\begin{cases} (z_i^c - 1) S(A(z_i)) \sum_{n=0}^{c-1} u(n) z_i^n = 0 \quad , \quad 1 \leq i \leq c-1, \\ c \sum_{n=0}^{c-1} u(n) = c - \mathrm{E}\left[S\right] \lambda. \end{cases} \tag{11}$$

Note that the zero $z_0 = 1$ cannot be used to determine the unknowns as its equation produces $0 = 0$.

A.2 IBSP: $\tilde{U}(z)$

$\tilde{U}(z)$ can be obtained analogously, except that we have to study the slot-by-slot evolution of the pair $(\tilde{Q}_k,N_k)$ of respectively the queue and the server content, because not necessarily c customers leave the system at the end of a service. This leads to a joint PGF $V(z,x)$ of the queue and the server content and we find $\tilde{U}(z)$ by letting $x \rightarrow z$.

References

Bruneel H (1983) Buffers with stochastic output interruptions, *Electronics Letters* 19: 461–463

Bruneel H (1985) Some remarks on discrete-time buffers with correlated arrivals, *Computers and Operations Research* 12(5):445–458

Chang SH, Choi DW (2005) Performance analysis of a finite-buffer discrete-time queue with bulk arrival, bulk service and vacations, *Computers and Operations Research* 32:2213–2234

Chang SH, Takine T (2005) Factorization and Stochastic Decomposition Properties in Bulk Queues with Generalized Vacations, *Queueing Systems* 50:165–183

Chaudhry ML, Templeton JGC (1983) *A first course in bulk queues*, John Wiley & Sons

Chen Y, Qiao C, Yu X (2004) Optical Burst Switching (OBS): A New Area in Optical Networking Research, *IEEE Network* 18(3):16–23

Claeys D, Walraevens J, Laevens K, Bruneel H (2007) A Discrete-time Queueing Model with a Batch Server Operating under the Minimum Batch Size Rule, *NEW2AN2007, Lecture Notes in Computer Science* 4712:248–259

Claeys D, Laevens K, Walraevens J, Bruneel H (2008) Delay in a discrete-time queueing model with batch arrivals and batch services, *Proceedings of the Fifth International Conference on Information Technology: New Generations* 1040–1045

Downton F (1955) Waiting Time in Bulk Service Queues, *Journal of the Royal Statistical Society, Series B (Methodological)* 17(2):256–261

Janssen AJEM, van Leeuwaarden JSH (2005) Analytic Computation Schemes for the Discrete-Time Bulk Service Queue, *Queueing Systems* 50:141-163

Kim NK, Chae KC, Chaudhry ML (2002) An invariance relation and a unified method to derive stationary queue lengths, *Operations Research* 52(5):756–764

Kim NK, Chaudhry ML (2006) Equivalences of Batch-Service Queues and Multi-Server Queues and Their Complete Simple Solutions in Terms of Roots, *Stochastic Analysis and Applications* 24:753–766

Klimenok V (2001) On the Modification of Rouche's Theorem for the Queueing Theory, *Queueing Systems* 38:431-434

Medhi J (1975) Waiting Time Distributions in a Poisson Queue with a General Bulk Service Rule, *Management Science* 21(2):777–782

Qiao CM, Yoo MS (1999) Optical burst switching (OBS) - a new paradigm for an optical Internet, *Journal of high speed networks* 8(1):69–84

Zhao YQ, Campbell LL (1996) Equilibrium probability calculations for a discrete-time bulk queue model, *Queueing Systems* 22:189–198