Analysis of Steady-State and Transient Delay in Discrete-Time Single-Arrival and Batch-Arrival Systems

J. Walraevens, D. Claeys and H. Bruneel
Department of Telecommunications and Information Processing
Ghent University - UGent
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium.
E-mail: {jw,dclaeys,hb}@telin.UGent.be

Abstract

We perform an analysis of the transient delay in a discrete-time FIFO buffer with batch arrivals. As transient delay is an ambiguous concept, we first discuss different possible definitions of the term (delay of the k-th customer, delay of a customer arriving at time t, etc). In this paper, we focus on the analysis of the delay of a customer arriving in slot t, also sometimes called virtual delay in single-arrival systems. It turns out that the modeling in batch-arrival systems is more intricate. In analysis, we relate transient delay to transient unfinished work and characterize the latter. Some time-dependent as well as limiting steady-state delay measures are calculated. We also study a variation that is related to active probing measurements. A substantial part of the article finally focuses on some fundamental differences between alternative definitions of transient delay.

Key words: Queueing theory, generating functions, transient analysis, virtual delay

1 Introduction

The delay (waiting time, sojourn time) experienced by customers is one of the most important performance measures of a queueing system. There is a vast literature dealing with the analysis of steady-state delay in a broad range of queueing systems. Fewer research efforts have addressed transient delay measures, though it might be of interest to characterize the evolution in time of delay experienced by customers. The latter is the topic of this paper.

The definition of 'transient delay' is not unambiguous. Basically, one can take a *customer-based* approach or a *time-based* approach. With the first approach, transient delay is defined as the delay of the k-th arriving (or departing) customer ($k \geq 1$). This transient delay is analyzed in for instance [1–8]. With the second approach, transient delay is defined as the delay of a customer arriving at time t ($t \geq 0$). In single-arrival systems, this is sometimes called the virtual delay or virtual waiting time at time t, as it is the delay that a 'virtual' customer that would arrive at time t would experience. Examples of analyses of this transient delay can be found in [1, 9, 10].

We note that the two alternative types of transient delay are substantially different. Only for *single-arrival* systems, the *steady-state* delay distributions are identical, see further. The differences between the *steady-state* delay distributions in case of batch arrivals were discussed by Burke [11]. Burke claims that the customer-based approach (as was done by himself) is the correct one, and the time-based approach (as was used in papers and books before Burke's paper appeared) is erroneous. This is a valid point of view if the steady-state delay is concerned, since one basically wants to characterize the delay of a random customer. However, in case of transient delay, the time-based approach makes sense as well; one does not necessarily know the ordinal number of a particular customer (the customer-based approach), while one might know the time of arrival of this customer (time-based approach). Therefore, characterization of the delay of a customer arriving at time t is interesting. It is also of interest to analyze the fundamental differences between both approaches. These two issues are the subject of the current contribution.

More precisely, we analyze the delay of a customer arriving during slot t ($t \geq 0$) in a discrete-time $Geo^X/G/1$ queue with a FIFO (First-In-First-Out) scheduling discipline. This paper is therefore complementary to our paper [8], where the delay of the k-th arriving customer ($k \geq 1$) in the same queue is analyzed. We note that almost all articles characterizing transient delay assume *single-arrival* queueing systems. One of the main challenges in the current analysis, however, is how to deal with the batch nature of the arrival process. Indeed, in single-arrival systems, the delay is equal to the sum of the unfinished work at the arrival instant (i.e., the time needed to process all customers present in the system at that time) and the service time of the arriving customer. In batch-arrival systems, other customers can arrive at the same instant as the virtual customer, so we have to deal with this issue. The queueing model is described in more detail in the following section and the analysis of the delay of a customer arriving in slot t is presented in section 3. We relate the probability generating functions (pgf) of the delay of a customer arriving in slot t and of the unfinished work at the beginning of slot t and use this relation to compute the transform function of the former sequence of pgfs (for all $t \geq 0$). In section 4, we calculate the pgf for $t \to \infty$. The transient mean delay is characterized in section 5. We then treat a slightly different model in section 6, where we assume the virtual customer to be an additional 'test' customer. This is, for instance, useful in active probing measurements, as extra customers are injected into the system in this type of measurements and the delay of these customers is measured (see, for instance, [12]). Different definitions of (transient) delay lead to different results. We therefore devote a substantial part of this article (section 7) to a comparison between the different transient (and steady-state) delays, and investigate fundamental identities and differences. We also provide some numerical examples. We conclude with some final remarks in the last section.

2 Queueing model

We consider a discrete-time single-server queueing system with infinite buffer space. Customers are served on a First-Come-First-Served basis. The numbers of arrivals during the consecutive time units (denoted as slots) are independent and identically distributed (i.i.d.) stochastic variables. Denote the number of arrivals in an arbitrarily chosen slot by a. We use the notations $a(n)$ and $A(z)$ to indicate its probability mass function (pmf) and probability generating function (pgf) respectively, i.e., $a(n) \triangleq \text{Prob}\left[a = n\right], n \geq 0$ and $A(z) \triangleq \text{E}\left[z^a\right] = \sum_{n=0}^{\infty} a(n)z^n$. The service times of customers are defined as the numbers of slots it takes to serve them and are assumed to be generally distributed with pmf $s(n)$, $n \geq 1$, and pgf $S(z) = \sum_{n=1}^{\infty} s(n)z^n$. The mean number of arrivals in a slot and the mean service time are given by $\lambda = A'(1)$ and $1/\mu = S'(1)$, respectively. The load equals $\rho = \lambda/\mu$.

3 Analysis

We are interested in the delay of an arbitrary customer arriving in slot t, i.e., from all customers arriving in slot t, we select one customer (the tagged customer) at random and analyze his delay (the number of slots he resides in the system, not including his slot of arrival). Denote this delay by d_t, for all $t \geq 0$. Note that slot t is a 'special' slot, as we assume that at least one customer arrives in this slot.

We first introduce the concept 'unit of work'. We may assume that each customer consists of a number of units of work equal to its service time and that the server processes at the rate of one unit of work per slot. We can then relate d_t to the unfinished work w_t in the system at the beginning of slot t (defined as the number of units of work that are in the system at the beginning of slot t), the amount of work units f_t arriving in slot t and to be executed before the tagged customer, and the service time s_t of the tagged customer himself:

$$d_t = (w_t - 1)^+ + f_t + s_t, \tag{1}$$

with $(x)^+ = \max(x, 0)$.

As mentioned, the distribution of the number of arrivals in slot t is different from the distribution of the number of arrivals in a random slot, as it is assumed that at least one customer arrives in slot t. However, since the numbers of per-slot arrivals are i.i.d., the special nature of slot t has no impact on the unfinished work w_t *at the beginning of that slot*, and, therefore, w_t is equally distributed as the unfinished work at the beginning of slot t in a system with a number of per-slot arrivals with pgf $A(z)$ in *all* slots (also in slot t). For that reason, f_t is the only variable of the right-handside of (1) that is affected by the special nature of slot t.

We now translate equation (1) to probability generating functions. We have:

$$\begin{aligned} D_t(z) &\triangleq \mathrm{E}[z^{d_t}] \\ &= \mathrm{E}\left[z^{(w_t-1)^+ + f_t + s_t}\right] \\ &= \frac{S(z)F(z)}{z}(W_t(z) + (z-1)W_t(0)), \end{aligned} \tag{2}$$

where $F(z) \triangleq \mathrm{E}\left[z^{f_t}\right]$ and $W_t(z) \triangleq \mathrm{E}\left[z^{w_t}\right]$. We express $F(z)$ as a function of $A(z)$ later. Let us first take the z-transform (or x-transform in this case) of the previous equation with respect to t. This will allow us to use results obtained in [13] and to find (semi-)closed-form results. We define

$$\begin{aligned} D(x,z) &\triangleq \sum_{t=0}^{\infty} D_t(z)x^t, \\ W(x,z) &\triangleq \sum_{t=0}^{\infty} W_t(z)x^t. \end{aligned}$$

Expression (2) then leads to

$$D(x,z) = \frac{S(z)F(z)}{z}(W(x,z) + (z-1)W(x,0)). \tag{3}$$

Next, we analyze the transient unfinished work in a $Geo^X/G/1$ queue. We note that this unfinished work is equal to the buffer content (i.e., the number of customers in the buffer, including the one in service) in a queue with *single-slot* service times but with each customer from the original system accounting for as many customers as the length of its service time (in slots), i.e., a $Geo^X/D/1$ system with the pgf of the number of customer arrivals in one slot equal to $A(S(z))$. The transient buffer content in a $Geo^X/D/1$ queue was already analyzed in [13]. Using these results, we find

$$W(x,z) = \frac{zW_0(z) + xW(x,0)(z-1)A(S(z))}{z - xA(S(z))}, \tag{4}$$

$$W(x,0) = \frac{W_0(Y(x))}{1-Y(x)}, \tag{5}$$

with $W_0(z)$ the (given) pgf of the unfinished work at the beginning (slot 0) and $Y(x)$ the only solution for z inside the unit disk of equation $z - xA(S(z)) = 0$ ($|x| < 1$). Substitution of (4)-(5) in (3) yields

$$D(x,z) = S(z)F(z)\frac{W_0(z)(1-Y(x)) + (z-1)W_0(Y(x))}{(z - xA(S(z)))(1-Y(x))}. \tag{6}$$

Finally, we calculate $F(z)$. As already mentioned, this function depends on the fact that the number of arrivals in slot t is differently distributed than in other slots. In fact, other definitions of f_t (and of the delay d_t) are possible as well; we will look into an example later. Since f_t is the amount of work arriving in slot t and to be executed before the tagged customer, and since the order of service of the customers arriving in one slot is completely arbitrary, we have

$$F(z) = \frac{1}{1-A(0)}\sum_{m=1}^{\infty}\sum_{l=0}^{m-1}\frac{a(m)}{m}S(z)^l.$$

This can be understood as follows. With probability $a(m)/(1-A(0))$, m customers ($m \geq 1$) arrive in slot t. The tagged customer is one of these m customers and he is served as $(l+1)$-st of all these customers with probability $1/m$ ($l = 0, \ldots, m-1$). Working out the summation over l leads to

$$F(z) = \frac{\sum_{m=1}^{\infty} \dfrac{a(m)}{m}(1 - S(z)^m)}{(1 - A(0))(1 - S(z))}.$$

We now relate the sum in the numerator to the pgf A. This can be done as follows:

$$\begin{aligned}\sum_{m=1}^{\infty} \frac{a(m)}{m}(1 - S(z)^m) &= \sum_{m=1}^{\infty} a(m) \int_{y=S(z)}^{y=1} y^{m-1} dy \\ &= \int_{y=S(z)}^{y=1} \sum_{m=1}^{\infty} a(m) y^{m-1} dy \\ &= \int_{y=S(z)}^{y=1} \frac{A(y) - A(0)}{y} dy,\end{aligned}$$

and thus

$$F(z) = \frac{\int_{y=S(z)}^{y=1} \dfrac{A(y) - A(0)}{y} dy}{(1 - A(0))(1 - S(z))}.$$

Substitution in (6) finally yields

$$D(x,z) = S(z) \frac{\int_{y=S(z)}^{y=1} \dfrac{A(y) - A(0)}{y} dy}{(1 - A(0))(1 - S(z))} \frac{W_0(z)(1 - Y(x)) + (z-1)W_0(Y(x))}{(z - xA(S(z)))(1 - Y(x))}. \tag{7}$$

By taking derivatives at $z = 1$, transform functions of the sequences of moments of the delay d_t ($t = 0, \ldots, \infty$) can be calculated. We will, for instance, calculate the transform function of the sequence of mean delays $\{\bar{d}_t\}_{t=0}^{\infty}$ in section 5. We note that it is usually not straightforward to invert these generating functions, mainly because of the implicitly defined function $Y(x)$, but that a couple of approaches have been suggested in literature, such as numerical inversion techniques [14, 15], iterative techniques [13], as well as analytical asymptotical techniques [16]. We also refer to section 5 in our article [8] for some discussion on this.

4 Limit for t $\rightarrow \infty$

The limit of $D_t(z)$ for $t \rightarrow \infty$ equals the pgf of the delay of an arbitrary customer arriving in a randomly chosen slot with at least one arriving customer in steady state (if such a steady state exists). Through the final value theorem [17, 18], we find that

$$\begin{aligned}D_\infty(z) &\triangleq \lim_{t \rightarrow \infty} D_t(z) \\ &= \lim_{x \rightarrow 1} (1 - x) D(x, z).\end{aligned}$$

By substituting (7) in this equation and taking the limit for x going to 1 (since $Y(1) = 1$, see [13], the use of l'Hôpital's rule is necessary in the last step), we find

$$D_\infty(z) = S(z) \frac{\int_{y=S(z)}^{y=1} \dfrac{A(y) - A(0)}{y} dy}{(1 - A(0))(1 - S(z))} \frac{z - 1}{Y'(1)(z - A(S(z)))}.$$

Finally, we calculate $Y'(1)$. By taking the first derivative of the equation $Y(x) = xA(S(Y(x)))$ and using the property that $Y(1) = 1$, we find

$$Y'(1) = \frac{1}{1-\rho},$$

and

$$D_\infty(z) = (1-\rho)\frac{\int_{y=S(z)}^{y=1} \dfrac{A(y)-A(0)}{y}dy}{(1-A(0))(1-S(z))}\frac{S(z)(z-1)}{z-A(S(z))}. \tag{8}$$

5 Mean delay

Here, we calculate the transform function of the sequence of mean delays $\{\bar{d}_t\}_{t=0}^{\infty}$ (with $\bar{d}_t \triangleq \mathrm{E}[d_t]$), defined as

$$\bar{D}(x) \triangleq \sum_{t=0}^{\infty} \bar{d}_t x^t.$$

It can be calculated from $D(x,z)$ as

$$\bar{D}(x) = \left.\frac{\partial D(x,z)}{\partial z}\right|_{z=1}.$$

By means of (7), we find

$$\bar{D}(x) = \frac{1}{2\mu(1-x)} + \frac{\rho}{2(1-A(0))(1-x)} + \frac{W_0'(1)}{1-x} + \frac{W_0(Y(x))}{(1-Y(x))(1-x)} - \frac{1-\rho x}{(1-x)^2}.$$

The sum of the first two terms gives the transform of the mean amount of work that arrives in slot t and that is part of the delay. The other terms account for that part of the delay that is caused by earlier arrived work.

The mean delay of a customer arriving in a given slot in steady state can be calculated from $\bar{D}(x)$ as well as from $D_\infty(z)$:

$$\begin{aligned}\bar{d}_\infty &= \lim_{x\to 1}(1-x)\bar{D}(x)\\ &= D_\infty'(1)\\ &= \frac{1}{2\mu} + \frac{\rho}{2(1-A(0))} + \frac{A''(1)}{2\mu^2(1-\rho)} + \frac{\lambda S''(1)}{2(1-\rho)}.\end{aligned} \tag{9}$$

6 A variation

In this section, we treat a small variation on the model. Here, we want to study the delay $\hat{d}_t$ of an *extra* customer arriving in slot t, and its pgf $D_t(z)$. The rationale is that one approach to measure the delay in a queue is to send in a test customer and measure his delay, the so-called active-probing approach (see for instance [12]). However, one should be careful, since this is not identical to the delay of a random customer, or even to the delay of a customer arriving in a random slot (or in slot t in the transient case). Therefore, it is interesting to study the differences between these different delays.

The model and analysis in this case are not much different from those in section 3. The only difference between $\hat{d}_t$ and d_t is the work that arrives in the same slot as the customer we are interested in and that is part of his delay, i.e., we need a different pgf $\hat{F}(z)$ that replaces function $F(z)$ from section 3.

The pgf of the total number of arrivals in slot t equals $zA(z)$, since an extra customer arrives in this slot, on top of the usual arrivals. Therefore, we find in the same way as in the original model that

$$\hat{F}(z) = \frac{\int_{y=S(z)}^{y=1} A(y)dy}{1-S(z)}.$$

We note that the difference between $\hat{F}(z)$ and $F(z)$, and thus also between $\hat{D}_t(z)$ and $D_t(z)$ and even between the two defined delays, does not depend on the slot t.

The function $\hat{D}(x,z)$ then becomes

$$\begin{aligned}\hat{D}(x,z) &\triangleq \sum_{t=0}^{\infty} \hat{D}_t(z)x^t \\ &= S(z)\frac{\int_{y=S(z)}^{y=1} A(y)dy}{1-S(z)} \frac{W_0(z)(1-Y(x)) + (z-1)W_0(Y(x))}{(z-xA(S(z)))(1-Y(x))}.\end{aligned}$$

The derived functions are given as follows in this variation.

$$\begin{aligned}\hat{D}_\infty(z) &= (1-\rho)\frac{\int_{y=S(z)}^{y=1} A(y)dy}{1-S(z)} \frac{S(z)(z-1)}{z-A(S(z))}, \\ \bar{\hat{D}}(x) &= \frac{2+\lambda}{2\mu(1-x)} + \frac{W_0'(1)}{1-x} + \frac{W_0(Y(x))}{(1-Y(x))(1-x)} - \frac{1-\rho x}{(1-x)^2}, \\ \bar{\hat{d}}_\infty &= \frac{1}{\mu} + \frac{\rho}{2} + \frac{A''(1)}{2\mu^2(1-\rho)} + \frac{\lambda S''(1)}{2(1-\rho)}. \end{aligned} \tag{10}$$

7 Some discussion on the different delays

In this section, we discuss the differences and similarities of the three transient delays, namely the delay of an arbitrary customer arriving in slot t (main topic of this article), the delay of an extra test-customer arriving in slot t (delay analyzed in section 6) and the delay of the k-th arriving customer (delay analyzed in [8]).

7.1 Delay of an arbitrarily chosen customer versus delay of a test customer, both arriving in slot t

We start with comparing the delays d_t and $\hat{d}_t$, analyzed in sections 3 and 6 respectively. We first note that the difference between these two delays is independent of slot t, as the only difference is the work that arrives in slot t and that is part of the studied delay, i.e., through the variables with pgfs $F(z)$ and $\hat{F}(z)$. Therefore, it is also equal to the difference of the steady-state delays. The difference between the transient mean delays at time t then equals

$$\bar{\hat{d}}_t - \bar{d}_t = \frac{1-(1+\lambda)A(0)}{2\mu(1-A(0))}, \tag{11}$$

which can be found from (9) and (10). This difference can be positive as well as negative. It is natural that this difference is positive, as the test customer is an *extra* customer. However, if the probability $A(0)$ of no arrivals in a slot is large, the test customer most likely arrives in a slot with no other arrivals. In the original model, we know that slot t is a slot with arrivals. If the probability that more than two customers arrive given that at least one customer arrives is high (which should be the case if the probability $A(0)$ of no arrivals in a slot is large, while the mean arrival rate λ is also high), this mean delay can indeed turn out to be higher than in the model with a test customer. The difference is then negative.

Deterministic batch sizes

Let us illustrate this in case of deterministic batch sizes. The number of arrivals in a slot is 0 or m (m fixed), i.e., $A(x) = 1 - \lambda/m + \lambda x^m/m$. The difference in (11) becomes

$$\bar{\bar{d}}_t - \bar{d}_t = \frac{\lambda - (m-1)}{2\mu}.$$

While for $m = 1$ (Bernoulli arrivals) the difference is positive, the difference is negative for $m > 1$. I.e., active probing leads to an underestimation of the mean delay of a customer arriving in a specified slot when customers arrive in batches. This is also depicted in Figure 1, where we show the relative difference $(\bar{\bar{d}}_t - \bar{d}_t)/\bar{d}_t$ in case of single-slot service times ($S(z) = z, \mu = 1$). It can be seen that active probing underestimates the mean delay considerably (upto 60% in the figure), especially for high arriving batch sizes and low arrival rates. This is logical as in those cases the test customer has a high probability of arriving in a slot with no other arrivals, while a random customer arrives in a batch with $m - 1$ other arrivals. It can also be seen from the figure that the relative difference goes to a limiting value for $m \to \infty$. From formulas (9) and (11), we find that this limit is equal to $-(1 - \lambda)$. A general conclusion that can be drawn is that one should be careful with measuring the delay by active probing methods when the customers arrive in batches.

[Figure 1 about here.]

7.2 Delay of a customer arriving in slot t versus delay of the k-th arriving customer

In this subsection, we study the difference between the delay of a customer arriving in slot t and the delay of the customer that is the k-th to arrive in the system. The latter was studied in [8] and we first review the expressions of interest from this paper. We denote the delay of the k-th arriving customer as $\tilde{d}_k$, its pgf by $\tilde{D}_k(z)$ and its expected value by $\bar{\bar{\tilde{d}}}_k$ ($k \geq 1$). In order to have a fair comparison between both delays, we assume zero unfinished work at time zero in both cases, i.e. $W_0(z) = 1$. Therefore, the first arriving customer does not have to wait and his delay equals his service time, leading to $\tilde{D}_1(z) = S(z)$. Then, in [8], the following pgfs and transform functions are derived:

$$\begin{aligned}
\tilde{D}(x,z) &\triangleq \sum_{k=1}^{\infty} \tilde{D}_k(z) x^k \\
&= \frac{xS(z)(1 - A(xS(z)))(z - V(x))}{(1 - xS(z))(1 - V(x))(z - A(xS(z)))}, \\
\tilde{D}_\infty(z) &= \frac{1-\rho}{\lambda} \frac{1 - A(S(z))}{1 - S(z)} \frac{S(z)(z-1)}{z - A(S(z))}, && (12) \\
\bar{\bar{\tilde{D}}}(x) &\triangleq \sum_{k=1}^{\infty} \bar{\bar{\tilde{d}}}_k x^k \\
&= \frac{x}{\mu(1-x)^2} + \frac{x(V(x) - A(x))}{(1-x)(1-A(x))(1-V(x))}, \\
\bar{\bar{\tilde{d}}}_\infty &= \frac{1}{\mu} + \frac{A''(1)}{2\lambda\mu(1-\rho)} + \frac{\lambda S''(1)}{2(1-\rho)}, && (13)
\end{aligned}$$

with $V(x)$ the sole solution for z in the unit disk of $z - A(xS(z)) = 0$ ($|x| < 1$).

In general, the two transient delays can be very different. In the remainder, we first discuss differences in *steady-state* delays. Then, we turn our attention to differences between *transient* delays.

7.2.1 Steady-state delays

Let us first look at the differences between the steady-state mean delays in (9) and (13):

$$\bar{\bar{d}}_\infty - \bar{d}_\infty = \frac{\mathrm{Var}[a](1 - A(0)) - \lambda^2 A(0)}{2\lambda\mu(1 - A(0))}. \tag{14}$$

Here, we used that $A''(1) = \mathrm{Var}[a] + \lambda^2 - \lambda$, with $\mathrm{Var}[a]$ the variance of the number of arrivals in a slot. This difference is non-negative. This can be observed as follows: the smallest $\mathrm{Var}[a]$ is the variance following from a Bernoulli distributed number of arrivals in a slot. This variance equals $\lambda(1 - \lambda)$ and $A(0) = 1 - \lambda$. Substituting this in (14) leads to a difference of zero. In most other cases, the difference (14) is positive. The reason is that for d_∞ each arrival slot has the same probability of being picked, while for $\tilde{d}_\infty$ each slot has a probability of being picked that is proportional to the number of arrivals in that slot. Slots with many arrivals thus have a higher probability of being picked, and customers arriving in these slots are typically subjected to higher delays. A notable exception is the case of deterministic batch sizes, as discussed next.

Deterministic batch sizes

Assume deterministic batch sizes, i.e., $A(x) = 1 - \lambda/m + \lambda x^m/m$. In this case, $D_\infty(z)$ and $\tilde{D}_\infty(z)$ are equal; we find from (8) and (12) that

$$D_\infty(z) = \tilde{D}_\infty(z) = \frac{1-\rho}{m}\frac{1 - S(z)^m}{1 - S(z)}\frac{S(z)(z-1)}{z - A(S(z))}.$$

This can be explained as follows: when choosing a random customer, each slot with arrivals has m chances to be the arrival slot of the randomly chosen customer. Since this m is fixed and this is true for all slots with arrivals, choosing a random customer or choosing a random slot with arrivals, boils down to the same. This is especially true for single arrivals (Bernoulli arrival process) as discussed before.

Mix of single and batch arrivals

In the previous paragraph, we have seen that the difference between the mean steady-state delays is zero for deterministic batch sizes. Here, we analyze whether the difference can be substantial for other batch size distributions. Therefore, we assume an arrival process which is a mix of single arrivals and batch arrivals of size m. In Figure 2, we depict the relative difference $(\bar{\bar{d}}_\infty - \bar{d}_\infty)/\bar{d}_\infty$ as a function of the size m of the batch arrivals, for three combination of the probability $A(0)$ of no arrivals and the mean arrival rate λ. It is clear from this figure that the difference can indeed be substantial (upto 40% in the figure) and depends on $A(0)$, λ and m.

[Figure 2 about here.]

7.2.2 Transient delays

We now turn our attention to the differences in the mean *transient* delays. Therefore, we calculate asymptotics from the generating functions $\bar{D}(x)$ and $\bar{\bar{D}}(x)$. These can be calculated by investigation of the transform functions (or related functions) in the neighbourhood of their dominant singularities (singularities with lowest norm), see [16] for a detailed procedure. We find

$$\bar{d}_t \sim \begin{cases} \bar{d}_\infty - \dfrac{\sqrt{x_Y}K_Y(x_Y)}{2\sqrt{\pi}(x_Y - 1)(Y(x_Y) - 1)^2} t^{-3/2} x_Y^{-t} & \text{if } \rho < 1 \\ \dfrac{2}{K_Y(1)\sqrt{\pi}}\sqrt{t} & \text{if } \rho = 1 \\ (\rho - 1)t & \text{if } \rho > 1 \end{cases}, \tag{15}$$

for $t \to \infty$ and

$$\bar{\bar{d}}_k \sim \begin{cases} \bar{\bar{d}}_\infty - \dfrac{x_V K_V(x_V)}{2\sqrt{\pi}(x_V-1)(V(x_V)-1)^2} k^{-3/2} x_V^{-k} & \text{if } \rho < 1 \\ \dfrac{2}{K_V(1)\sqrt{\pi}}\sqrt{k} & \text{if } \rho = 1 \\ \left(\dfrac{1}{\mu} - \dfrac{1}{\lambda}\right) k & \text{if } \rho > 1 \end{cases}, \tag{16}$$

for $k \to \infty$. Here, x_Y and x_V are the dominant singularities on the real positive axis of $Y(x)$ and $V(x)$, and $K_Y(x_Y)$ and $K_V(x_V)$ are constants that can be calculated as function of the input functions $A(x)$ and $S(x)$ (but their precise form is not important for the following discussion). Since both $Y(x)$ and $V(x)$ are implicitly defined, the singularities are square-root branchpoints (see [19, 20]) and we have

$$\begin{aligned} Y(x) &\sim Y(x_Y) - K_Y(x_Y)\sqrt{x_Y - x}, \\ V(x) &\sim V(x_V) - K_V(x_V)\sqrt{x_V - x}, \end{aligned}$$

for $x \to x_Y$ and $x \to x_V$, respectively. Expressions (15)-(16) are found as follows: for $\rho < 1$, $x = 1$ is a pole with multiplicity 1, leading to the steady-state values. The second dominant singularities are the branchpoints of $Y(x)$ and $V(x)$ respectively, leading to the asymptotic terms in $t^{-3/2}x_Y^{-t}$ and $k^{-3/2}x_V^{-k}$. For $\rho = 1$, $x_Y = x_V = 1$ and $x = 1$ is a branchpoint of the form $1/(1-x)^{3/2}$ of $\bar{D}(x)$ and $\bar{\bar{D}}(x)$, leading to a mean delay $\bar{d}_t$ and $\bar{\bar{d}}_k$ going to infinity according to $\sqrt{t}$ or $\sqrt{k}$. Finally, in case of $\rho > 1$, $x = 1$ is a pole of multiplicity 2 of $\bar{D}(x)$ and $\bar{\bar{D}}(x)$, leading to the $\sim t$ and $\sim k$ patterns. In this overload case, the buffer is non-empty with probability 1 in the limit for t or k going to infinity and the mean delays are respectively increased with $\rho - 1$ slots per slot (on average ρ new arriving work units minus one finished work unit per slot) and $1/\mu - 1/\lambda$ slots per arriving customer (extra mean waiting time of $1/\mu$ slots mean interdeparture time minus $1/\lambda$ slots mean interarrival time).

The kernel solutions $Y(x)$ and $V(x)$

We notice from the asymptotics in the general case that the two transient delays (time-based and customer-based) are driven by two different "kernel solutions" $Y(x)$ and $V(x)$, respectively. Before looking into some special cases, we take a closer look at these two functions.

We have

$$Y(x) = xA(S(Y(x))), \tag{17}$$

$$V(x) = A(xS(V(x))). \tag{18}$$

These functions are in fact (defective) generating functions, as it is easy to see that the following relations between random variables lead to (17)-(18):

$$y = 1 + \sum_{i=1}^{a}\sum_{j=1}^{s_i} y_{i,j}, \tag{19}$$

$$v = a + \sum_{i=1}^{a}\sum_{j=1}^{s_i} v_{i,j}, \tag{20}$$

with y and the $y_{i,j}$ random variables with the same distribution, v and the $v_{i,j}$ also random variables with the same distribution, a the number of arrivals in a random slot and s_i the service time of the i-th customer counted in a. The $y_{i,j}$s are mutually independent, as are the $v_{i,j}$s. From (19), it can be deduced that y is the number of slots that it takes to reduce the unfinished work at the beginning of a slot by one unit. Indeed, this equals one slot if there are no new arrivals in that slot. If there are new arrivals, each new unit of work (there are $\sum_{i=1}^{a} s_i$ new units) takes a number of slots that is equally distributed as y, leading to (19). Equation (20) on the other hand demonstrates that v equals the total number of arrivals that is "generated" during a slot in the sense that not just the number of arrivals in the first slot are counted but also the arrivals in the slots when these customers are served, and the arrivals in the slots when these latter customers are served, etc.

It is clear that y and v are essentially different, but also, that they are connected in some way. The variable y can be regarded as some sort of sub-busy period, namely the time needed to reduce unfinished work by 1 unit. The variable v is in fact the number of arrivals during this sub-busy period, i.e., we can write

$$v = \sum_{i=1}^{y} a_i, \tag{21}$$

with a_i the number of arrivals during the i-th slot of the sub-busy period. Note that this does not mean that there is a clear-cut relation between the pgfs of y and v. The main reason is that the a_i and the variable y in (21) are non-independent. Indeed, the higher the a_i the higher we expect y to be. Therefore, the relation $V(x) = Y(A(x))$ that one could naively expect from (21) is incorrect. Note that, due to Wald's equation [21], there exists an easy relation between the expected values of y and v:

$$\begin{aligned} \mathrm{E}[v] &= \mathrm{E}[a]\mathrm{E}[y] \\ &= \lambda \mathrm{E}[y], \end{aligned}$$

which also follows from the derivatives at $z = 1$ of the pgfs $Y(x)$ and $V(x)$.

A second connection can also be exposed. Since the length of a sub-busy period equals 1 slot augmented with all the service times of customers arriving during the sub-busy period, we can write

$$y = 1 + \sum_{i=1}^{v} s_i, \tag{22}$$

with s_i the service time of the i-th customer arriving during the sub-busy period. Because of dependence between v and the s_i, this does again not lead to an easy relation between $Y(x)$ and $V(x)$, although Wald's equation leads to a second relation between the averages $\mathrm{E}[y] = 1{+}\mathrm{E}[v]/\mu$. However, this second relation between stochastic variables is interesting as it shows that, in some special cases, it does lead to relations between pgfs. For instance, when service times are all equal to 1 slot, (22) results in

$$y = 1 + v,$$

and after translation to pgfs

$$Y(x) = xV(x).$$

This can also be acquired from (17)-(18) by putting $S(x) = x$. So, in this case, there is a clear-cut relation between $Y(x)$ and $V(x)$, driving the transient delays. More generally, for deterministic service times of l slots, we find from (22)

$$y = 1 + l \cdot v,$$

and after translation to pgfs

$$Y(x) = xV(x^l).$$

Again this can be discovered from (17)-(18) as well, by first substituting x by x^l in (18) and multiplying both sides with x, and further observing that this leads to the same implicit function for $xV(x^l)$ as (17) does for $Y(x)$.

Single arrivals

Let us finally look at the rate of convergence of the mean transient delays $\bar{d}_t$ and $\bar{\bar{d}}_k$ to their steady-state values (assuming there is a steady state, i.e., for $\rho < 1$) for single arrivals. Then, $A(x) = 1 - \lambda + \lambda x$. Since the steady-state delays are equal in case of Bernoulli arrivals (cf. paragraph 7.2.1), a natural question to pose is which of both delays converges the fastest to this steady-state value. The rates of convergence are determined by x_Y and x_V respectively. We prove that $x_Y < x_V$ and thus that the mean delay $\bar{\bar{d}}_k$ converges faster to the steady state than $\bar{d}_t$. For Bernoulli arrivals, the functions $Y(x)$ and $V(x)$ are determined by

$$\begin{aligned} Y(x) &= x(1 - \lambda + \lambda S(Y(x))), \\ V(x) &= 1 - \lambda + \lambda x S(V(x)). \end{aligned}$$

From these formulas, it is easy to see that $Y(x) = V(x)$ for $x = 1$ only. Since $\lambda < 1$, $Y'(1) > V'(1)$, and $Y(x) > V(x)$ for all $x > 1$ where both functions exist. The derivatives of both functions are given by

$$\begin{aligned} Y'(x) &= \frac{1 - \lambda + \lambda S(Y(x))}{1 - \lambda x S'(Y(x))}, \\ V'(x) &= \frac{\lambda S(V(x))}{1 - \lambda x S'(V(x))}. \end{aligned}$$

The branchpoints x_Y and x_V are reached when the denominators of $Y'(x)$ and $V'(x)$ are zero. Since $xS'(Y(x)) > xS'(V(x))$ the former function will reach 1 for a lower x, and thus $x_Y < x_V$. Thus, the mean delay $\bar{\bar{d}}_k$ converges quicker to its steady-state value than $\bar{d}_t$ does. This can be intuitively understood as follows: since at most one customer arrives in a slot, customer k arrives in slot $t = I(k)$ with $I(k)$ a strictly increasing discrete function; for each increment of k with 1 unit, $t = I(k)$ increases with *at least* 1 unit (but possibly more), and thus $\bar{\bar{d}}_k$ has to converge quicker than $\bar{d}_t$ to their mutual steady-state value. This is also demonstrated in Figure 3, where we show x_Y and x_V as functions of the arrival rate λ for single arrivals, geometric service times ($S(x) = \mu x/(1 - (1 - \mu)x)$) and for different values of the load $\rho(= \lambda/\mu)$. Since $\mu < 1$, the arrival rate is restricted to the interval $[0, \rho]$. We see that x_Y and x_V differ the most for low arrival rates and low loads and they become equal for $\lambda \to \rho$, i.e. for $\mu \to 1$.

[Figure 3 about here.]

8 Concluding remarks

We elaborately studied the transient delay in a batch-arrival queue. We first argued that defining the transient delay is not a straightforward matter, especially in case of batch arrivals. We therefore suggested several possible definitions and analyzed the transient delay in these cases. Some discussion on the fundamental differences and identities between these delays reveals that we should be careful, especially in case of batch arrivals. Even their steady-state limits can differ. Therefore, we suggest caution when using, for instance, active probing for measuring the (transient) delay. Our results could furthermore be used to correct these measurements.

Acknowledgment. The authors would like to thank the two referees for their constructive criticism of the first version of this article. The authors further acknowledge support of the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

References

[1] J. Cohen, The single server queue, North-Holland, Amsterdam, 1982.

[2] C. Wang, On the transient delays of M/G/1 queues, Journal of Applied Probability 36 (3) (1999) 882–893.

[3] C. Wang, An identity of the GI/G/1 transient delay and its applications, Probability in the Engineering and Informational Sciences 16 (1) (2002) 47–66.

[4] T. Hofkens, K. Spaey, C. Blondia, Transient analysis of the D-BMAP/G/1 queue with an application to the dimensioning of a playout buffer for VBR video, Lecture Notes in Computer Science 3042 (2004) 1338–1343.

[5] A. Janssen, J. van Leeuwaarden, Relaxation time for the discrete D/G/1 queue, Queueing Systems 50 (1) (2005) 53–80.

[6] M. Vlasiou, B. Zwart, Time-dependent behaviour of an alternating service queue, Stochastic Models 23 (2) (2007) 235–263.

[7] O. Baron, On the law of the i^{th} waiting time in a busy period of G/M/c queues, Probability in the Engineering and Informational Sciences 22 (1) (2008) 75–80.

[8] J. Walraevens, D. Fiems, H. Bruneel, Combined analysis of transient delay characteristics and delay autocorrelation function in the $\mathrm{Geo}^{\mathrm{X}}$/G/1 queue, Stochastic Models 28 (2) (2012) 333–357.

[9] J. Abate, W. Whitt, Transient behavior of the M/G/1 workload process, Operations Research 42 (4) (1994) 750–764.

[10] D. Bertsimas, G. Mourtizinou, Transient laws of non-stationary queueing systems and their applications, Queueing Systems 25 (1-4) (1997) 115–155.

[11] P. Burke, Delays in single-server queues with batch input, Operations Research 23 (4) (1975) 830–833.

[12] A. Novak, R. Watson, Determining an adequate probe separation for estimating the arrival rate in an M/D/1 queue using single-packet probing, Queueing Systems 61 (4) (2009) 255–272.

[13] H. Bruneel, Exact derivation of transient behavior for buffers with random output interruptions, Computers Networks and ISDN Systems 22 (1991) 277–285.

[14] J. Abate, W. Whitt, Solving probability transform functional equations for numerical inversion, Operations Research Letters 12 (5) (1992) 275–281.

[15] J. Abate, W. Whitt, Numerical inversion of probability generating functions, Operations Research Letters 12 (4) (1992) 245–251.

[16] J. Walraevens, D. Fiems, M. Moeneclaey, Using singularity analysis to approximate transient characteristics in queueing systems, Probability in the Engineering and Informational Sciences 23 (2) (2009) 333–355.

[17] E. Gluskin, J. Walraevens, On two generalizations of the final value theorem: scientific relevance, first applications, and physical foundations, International Journal of Systems Science 42 (12) (2011) 2045–2055.

[18] W. Hayt, J. Kemmerly, Engineering circuit analysis (Fifth edition), McGraw-Hill, New York, 1993.

[19] M. Drmota, Systems of functional equations, Random Structures & Algorithms 10 (1-2) (1997) 103–124.

[20] P. Flajolet, R. Sedgewick, Analytic Combinatorics, Cambridge University Press, 2008.

[21] W. Feller, An Introduction to Probability, Theory and Its Applications, Vol. II, John Wiley & Sons, New York, 1971.

List of Figures

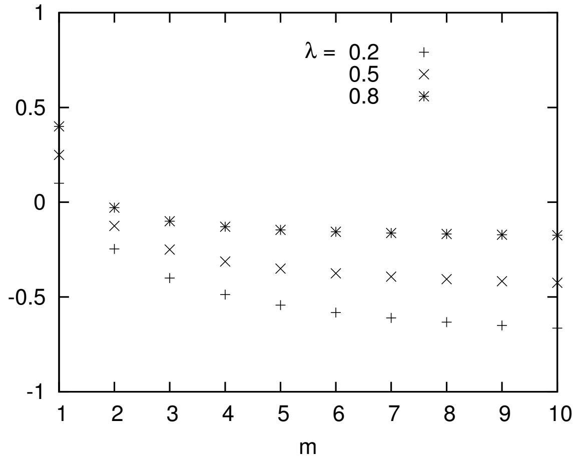

Figure 1: The relative difference between the mean delays of an arbitrarily chosen customer and of a test customer, both arriving in slot t, in case of deterministic batch sizes as a function of the batch size m and for different values of the arrival rate λ

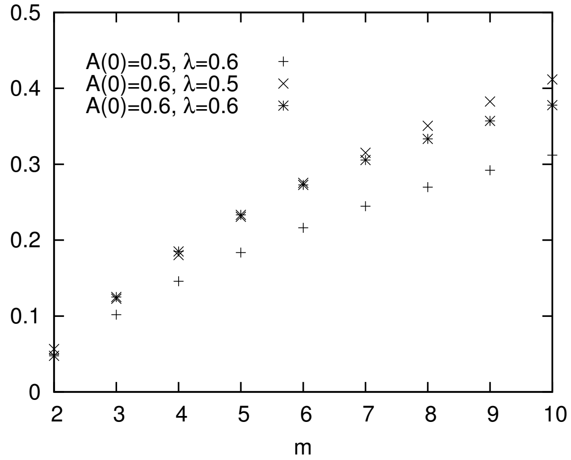

Figure 2: The relative difference between the mean steady-state delays of an arbitrary chosen customer and of a customer arriving in a randomly chosen slot in case of a mix of single arrivals and batch arrivals with size m as a function of this batch size m and for different values of the arrival rate λ and the probability $A(0)$ of no arrivals

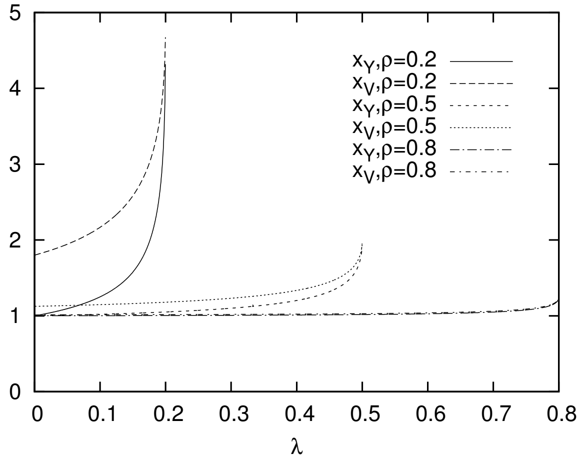

Figure 3: The convergence rates x_Y and x_V of the transient delays of a customer arriving in slot t and the k-th arriving customer, for single arrivals, geometric service times and different values of the load ρ