Approximate Flow Time Distribution of a Queue with Batch Service

Sven Bartelsen, Manfred Mittler, Oliver Rose

Report No. 174 June 1997

Bayerische Julius-Maximilians-Universität Würzburg
Lehrstuhl für Informatik III
Am Hubland, D - 97074 Würzburg
Tel.: +49-931-8885518, Fax: +49-931-8884601
e-mail: {bartelsen|rose}@informatik.uni-wuerzburg.de
mittler@VNET.IBM.COM

Abstract: In this paper, we investigate a single-server queue with finite capacity that features batch service operations which can be found in semiconductor manufacturing. We assume that service periods are triggered according to the minimum batch size rule. Using the embedded Markov chain technique, the resulting steady-state probabilities in case of Poisson arrivals and arbitrary service time distributions can be used to construct an approximate flow time distribution.
1 Introduction

In semiconductor manufacturing, several wafers are accumulated together to lots for transportation and processing purposes. Batch service tools can accommodate several lots at the same time. The major problem regarding these types of tools is the question of how many lots are needed to initiate a service period. Usually, a minimum batch size is used. If this size is one, the control rule corresponds to the greedy policy. The server can also force full batches, where service is started only if the server can be filled completely. Due to sophisticated and complex equipments, modeling and analysis techniques have to be employed to determine the optimum batch size that yields the shortest flow time (time a wafer spends in the manufacturing system) or the highest equipment utilization. This is in particular the case when rework occurs.

In the past, much progress has been made in the derivation of flow (sojourn) times in batch-service queues (see Gold and Tran-Gia (1993), and Tran-Gia and Schömig (1996)).

The aim of this paper is to extend the analysis of a finite capacity single-server queue with batch service towards an approximation of the flow time distribution of customers through this system. We apply the minimum batch size rule and use the embedded Markov chain technique to derive state probabilities under stationary conditions in case of Poisson arrivals, general service time distributions, and FIFO queueing discipline.

2 Model description and notation

The model consists of a finite queue with a maximum capacity of $S$ lots of wafers and a single batch server (as depicted in Figure 1). We use the terminology of queuing analysis and refer to these lots as customers. The customers enter the system according to a Poisson process with rate $\lambda$. A lot is not admitted to the system if it finds the queue fully occupied upon its arrival. The server only starts processing if there are at least $a$ customers in the queue. At most $b$ customers are served simultaneously. Service times are drawn from a common general distribution $H$ and are mutually independent of one another. Finally, the server picks up customers from the queue according to the FIFO dispatching rule. The following Kendall-like notation of this model reads $M/G^{[a,b]}/1 - S - FIFO$.

For a random variable (r.v.) $X$, the following notation is used throughout this paper:
$f_X(t)$ : probability density function of $X$, 
$F_X(t)$ : probability distribution function of $X$, 
$E[X]$ : mean of $X$, 
$c_X$ : coefficient of variation of $X$.

3 Analysis

The analysis of this system is based on the approach of Gold and Tran-Gia (1993) who analyzed an $M/G^{[a,b]}/1 - S$ - FIFO system. They derived the stationary distribution of number of customers in the queue at arbitrary time instants. The outline of the analysis is as follows. First, we derive the state probabilities of the embedded Markov chain. From these probabilities we obtain the state probabilities at arbitrary instants which then can in turn be employed to calculate the approximation of the flow (sojourn) time and additional performance measures. For this purpose, the distribution of both the number of customers in the system and the number of waiting customers at arbitrary instants are needed.

3.1 State probabilities

Since we deal with a queuing system with Poisson arrivals and general service time distributions it is instructive to employ the embedded Markov chain technique, where the observation instants are just prior to the service completion instants. A detailed description of how to set up the state space and the state transition matrix $P$ can be found in Gold and Tran-Gia (1993). For this purpose, let $N_0$ denote the random variable of the number of customers in the queue at embedded time instants. Also, let $d_k$ be the probability of $k$ customers arriving during a service period $H$, which can be calculated using the Poisson-Transform of the service time distribution function (see Bolgiano Jr. and Piovoso (1968)). To simplify the derivation of the Poisson-Transform, Gold and
Tran-Gia (1993) relied on substitute distribution functions which are matched according to the first two moments (cf. Kühn (1979)). For these special cases, he obtained explicit expressions for the probability $d_k$.

3.2 Queue size and system size distribution at arbitrary time instants

Gold and Tran-Gia (1993) show how to compute the queue size distribution at arbitrary time time instants $N^*_q$. This will not be recalled here. In a similar way one can derive the system size distribution at arbitrary time instants $N^*$.

4 Approximate flow time distribution

The above results can now be used to construct an approximation of the flow time distribution.

4.1 Continuous time

Let $A$ denote the random variable of the interarrival time which is exponentially distributed with rate $\lambda$:

$$A(t) = 1 - e^{-\lambda t}$$ and $$a(t) = \lambda e^{-\lambda t}.$$ (4.1)

Since we are interested in the flow time distribution of an arbitrary customer, only customers that are admitted to system need to be considered. Thus, upon arrival an admitted customer experiences a queue size distribution $N^*_q(i)$ with $0 \leq i < S$, which is given by the following equation:

$$N^*_q(i) = \frac{N^*_q(i)}{\sum_{j=0}^{S-1} N^*_q(j)}. \tag{4.2}$$

Further, the probability of the server being idle upon arrival of an admitted customer and seeing a system size of $j$ ($0 \leq j < a$), $N^*_a(j)$, reads as:

$$N^*_a(j) = \frac{N^*_a(j)}{\sum_{i=0}^{S-1} N^*_a(i)}.$$ (4.3)

In deriving equations describing the flow time distribution, two cases have to be examined: The arrival of a customer during an idle or a busy period of the server.
4.1.1 Arrival during an idle period

During an idle phase less than \( a \) customers are already in the system. Thus, a newly arriving customer finding \( j \) customers ahead of himself in the queue must wait another \( a - j - 1 \) interarrival times to be served. Let \( F_{j\text{idle}} \) denote the random variable of his flow time, which follows immediately and is given by:

\[
F_{j\text{idle}} = A + \cdots + A + H, \tag{4.4}
\]

where the sum of \( l \) interarrival times is Erlang-distributed, i.e.:

\[
e_l(t) = \frac{\lambda(\lambda t)^{l-1}}{(l-1)!} e^{-\lambda t}. \tag{4.5}
\]

4.1.2 Arrival during a service period

During a service period there are at least \( a \) customers in the system. A newly arriving customer finds \( i \) customers in the queue. The first part of his waiting time is determined by the ongoing service phase. This quantity is represented by the remaining service time conditioned on experiencing a queue size of \( i \). Its respective random variable is denoted by \( H_i^{rem} \). A customer finding \( i \) customers ahead of him in the queue upon arrival cannot be viewed as an arbitrary observer of the system any more because the probability of seeing a queue size of \( i \) is not independent of the history of the server activity. However, a quantity describing the remaining service time is needed and to keep the following analysis tractable, it is assumed that the remaining service time is independent of the condition stated above, which leads to the recurrence time \( H^r \). Thus, \( H_i^{rem} \) is approximated by the recurrence time \( H^r \) (see Eqns. (4.6) and (4.7)).

The second part of a customer’s waiting time is determined by \( k = \text{floor}(i/b) \) batches that will be served prior to him.

In order to be served immediately after the last service phase of customers having already occupied the system prior to his arrival, at least \( a \) customers need to be present in the queue. For this purpose, let \( l = (i + 1) \mod b \) denote the number of customers left in the queue at the time the considered customer could be served given that no further customers have arrived after this customer. Clearly, if \( l \geq a \) holds our customer does not have to wait any longer and is served immediately. However, if \( l \) is less than \( a \), it took the remaining \( a - j \) customers longer to fill the batch than the server to deal with all other batches ahead of them in the queue. Therefore, let \( F_i \) denote the random variable of the flow time conditioned on seeing \( i \) customers in the queue upon arrival. The following
equations summarize the approximated flow time distribution of an arbitrary customer:

\[ l = 0, l \geq a: \]
\[ F_i = H^r + H + \cdots + H + H, \]
\[ \text{k-fold} \]  
(4.6)

\[ 0 < l < a: \]
\[ F_i = \max \left( A + \cdots + A, H^r + H + \cdots + H \right) + H. \]
\[ \text{k-fold} \]  
(4.7)

4.1.3 Flow time distribution

From the above, we are able to determine the distribution of the flow time of an arbitrary customer \( F(t) \) by applying the law of total probability:

\[
F(t) = \sum_{j=0}^{a-1} F_{j \text{die}}(t) \cdot N^*_a(j) + \sum_{i=0}^{a-1} F_i(t) \cdot (N^*_{q_a}(i) - N^*_a(i)) + \\
\sum_{i=a}^{S-1} F_i(t) \cdot N^*_q(i).
\]

(4.8)

In continuous time, Eqns. (4.6) and (4.7) have some serious drawbacks because one needs to compute an \( l \)-fold convolution of the service time, which results in solving integrals directly by numerical means or using Laplace transforms. But, the most serious problem is to solve for the maximum of two random variables in continuous time since to the authors' knowledge no helpful tools such as Laplace transforms are available. Therefore, we will leave the continuous time domain and consider the problem in the discretized time domain, where time only proceeds unit-wise.

4.2 Discrete time

In discrete time, the time axis is divided into equidistant time units of length \( \Delta t \). The system under investigation is observed only at these time instants. In this domain, the exponential distribution changes to a geometric distribution shifted to the right by one unit and is denoted by \( \text{Geom}(1) \). It has the Markov-property immediately after the observation instants, which is exploited by embedding a Markov chain immediately after service completion. It can be shown that the analysis of our system in continuous time can easily be modified to match the analysis in discrete time. A detailed treatment of
discrete time analysis can be found in Ackroyd (1980) and Tran-Gia and Ahmadi (1988). In the following, we give a brief introduction into the subject of discrete time analysis.

4.2.1 Random variables and operators

Let $X$ denote a random variable in discrete time. Then, with $k \in \mathbb{N}$, its distribution is defined as:

$$x(k) = P\{X = k \cdot \Delta t\}, \quad (4.9)$$

and its distribution function:

$$X(k) = P\{X \leq k \cdot \Delta t\}. \quad (4.10)$$

Further its mean concerning time units is given by:

$$E[X] = \sum_{k=0}^{\infty} k \cdot x(k). \quad (4.11)$$

As stated before, the shifted geometric distribution $A = Geom(1)$ represents the time-discrete equivalent to the time-continuous negative-exponential distribution. Its respective distribution is given by the next equation:

$$A(k) = (1 - q_A) \cdot q_A^{k-1}, \quad (4.12)$$

with

$$q_A = \frac{E[A] - 1}{E[A]}. \quad (4.13)$$

The forward recurrence time $X^r$ of $X$ with observation instants prior to discrete time instants is given by:

$$x^r(k) = \frac{1}{E[X]} \left(1 - \sum_{i=0}^{k} x(i)\right). \quad (4.14)$$

The main reason for abandoning the continuous time domain and switching to discrete time, has its reason in the effective computation of the convolution of two independent random variables $X_1$ and $X_2$. The sum $X = X_1 + X_2$ is distributed like:

$$x(k) = \sum_{l=-\infty}^{\infty} x_1(l) \cdot x_2(k - l) = x_1(k) \circledast x_2(k), \quad (4.15)$$

where $\circledast$ denotes the discrete convolution operator. It can be computed efficiently using techniques usually applied in signal processing like Fast Fourier Transforms (FFT).

Also, the maximum of two independent random variables $Y = \max(X_1, X_2)$ can be evaluated and is distributed like:

$$y(k) = x_1(k) \cdot x_2(k - 1) + X_1(k) \cdot x_2(k) \quad (4.16)$$
The distribution of the number of arrivals during a service phase $d_k$ in case of $Geom(1)$ arrivals can be obtained as follows (see Tran-Gia and Ahmadi (1988)):

$$d_k = \sum_{i=k}^{\infty} h(i) \binom{i}{k} (1 - q_A)^k \cdot q_A^{i-k},$$

(4.17)

where $h(i)$ is the distribution of the service time $H$. Analogously, we derive the distribution of number of arrivals during a recurrence time of the service time:

$$d^r_k = \sum_{i=k}^{\infty} h^r(i) \binom{i}{k} (1 - q_A)^k \cdot q_A^{i-k}.$$  

(4.18)

### 4.2.2 Analysis

The $Geom(1)/G^{[a,b]}/1 - S$ queue is analyzed using the embedded Markov chain technique as described earlier. The system becomes memoryless immediately after service completion.

In contrast to continuous time domains, discrete time domains entail that two events can occur at the same time instant. In this case, we model a service completion instant prior to a customer arrival instant, such that this customer can be served immediately if less than $b$ customers ahead of him have been awaiting service.

The equations mentioned to calculate the queue and system size distribution at arbitrary time instants remain unchanged and can be applied.

**Approximate flow time distribution**

Despite of having switched to discrete time, Eqns. (4.2) - (4.8) given in Section 4.1 retain their validity.

Our approximation of the flow time distribution is exact if the remaining service time conditioned on $i$ customers already in the queue, $H_i^{rem}$, is equal to the recurrence time $H^r$. This is the case for $H = Geometric(0)$, an unshifted geometric distribution. Suppose a newly started service will only take one time unit. During that time period one customer can enter the system, but his arrival will be taken into account after service completion in this particular case. Thus, he experiences no waiting time due to the ongoing service period. Thus, we approximate $H_i^{rem}$ by $H^r$. If the service time distribution is memoryless, $H_i^{rem}$ will not depend on how many customers are currently in the queue since its history is not remembered any more. A distribution that satisfies this property is the unshifted geometric one.
5 Numerical results

We have used the iterated matrix-vector multiplication to compute the state probabilities at embedded time instants, which led to simulation-conforming results.

![Graphs showing state probabilities](image)

(a) $a = 1$

(b) $a = 4$

(c) $a = 8$

**Figure 2** $H = Geometric(10, 0)$

In the following, we have considered a queue that has a server capacity $b$ equal to 8 and a queue length of $S=64$. We have chosen service time distributions with a common mean of 10 time units, but different amounts of variation. The arrival stream is determined by a $Geom(1)$ distribution with an arrival rate of $\lambda = E[A]^{-1}$. In the following figures, the number of time units is denoted by $k$ and the flow time distribution is represented by $f(k)$. 
In Figure 2, the service time distribution is chosen to be $H = Geometric(10, 0)$, an unshifted geometric distribution with mean 10, which has the Markov property as explained above. Thus, the approximated flow time distributions are exact. They have been confirmed by simulation, but only the analytical results are plotted. Curves are presented for values of the starting threshold $a$ equal to 1, 4, and 8 by Figures 2(a), (b), and (c), respectively. In any case, flow time distributions corresponding to three different traffic intensities are shown: low ($\lambda = 0.1 \cdot b$), medium ($\lambda = 0.5 \cdot b$), and high ($\lambda = 0.9 \cdot b$).

In case of the lowest possible starting threshold ($a = 1$), the tails of the flow time distributions grow with increasing traffic intensity resulting in a monotonically growing mean.

If the starting threshold is set to higher values, the distributions corresponding to the low traffic intensity become more and more heavily tailed because customers have to wait a long time until another service phase is triggered. They also have to wait a long time, if the traffic intensity is high since many customers usually occupy the queue and are served first. Flow time distributions with small tails result from medium traffic intensities. Then, the customers’ waiting time is minimized because they arrive as frequently as to reach the starting threshold but not as frequently as to considerably fill the queue. Consequently, this leads to a low mean flow time.

In the paragraphs below, we intend to examine the quality of our approximation. For that purpose, we used two different service time distributions with strongly deviating coefficients of variation. These are on the one hand the deterministic distribution with mean 10 and coefficient of variation $c_H = 0$ ($H = Deterministic(10)$) and on the other hand a negative binomial distribution with mean 10 and $c_H = 2$ ($H = NegBin(10, 2)$). Further, to get a better insight of the accuracy of our approximation of the flow time distribution, we divided our investigation into low, medium, and high offered traffic intensity cases as already explained above.

In Figure 3(a), a starting threshold of $a = 4$ was applied, whereas in Figure 3(b) $a$ was set to a value of 8. Since the offered traffic intensity is low the distributions show a long tail, which is on account of customers waiting to reach the starting threshold. Thus, the server is idle most of the time leading to only a small fraction of time, where the remaining service time $H_t^{rem}$ can be experienced by newly arriving customers. Since this quantity is of marginal importance in a low traffic scenario, approximating $H^{rem}$ by the recurrence time $H^r$ does not influence the overall shape of the distribution grossly, regardless of what service time distribution has been used.
Figure 3  Light traffic intensity: $\lambda = 0.1 \cdot b$

Figure 4  Medium traffic intensity: $\lambda = 0.5 \cdot b, a = 4$

Figure 4 depicts the flow time distributions with a medium traffic intensity and a starting threshold of $a = 4$. Clearly, the approximation is better in the case of $H = \text{NegBin}(10, 2)$, a service time distribution of a fairly great amount of variation. This may be explained by considering the history of service times. The greater the variability the less history is carried along with time since strong fluctuations in service times lead also to strong fluctuations in queue size. Thus, the impact of the condition of being in state $i$ is diminished which leads to remaining service times $(H^\text{rem}_i)$ that are not very different from one another. In that case, an approximation by $H^r$ is useful and
accounts for acceptable results as shown in Figure 4(b).

Considering Figure 4(a) where a deterministic service time has been used, its history is thoroughly transferred by time because its end can be determined if one knows at what time a service phase was initiated. As a consequence, $H_i^{rem}$ is strongly dependent on the condition of seeing $i$ customers in the queue upon arrival. Thus, an approximation by $H^*$ will give rather coarse results.

At last, we consider a high load scenario with $a = 4$, which is delineated in Figure 5. Again, in case of a deterministic service time the approximation of the flow time distribution is only coarse due to the explanation given in the paragraph above.

A minor approximation is also found in Figure 5(b) because the blocking probability is about 13%. Since the remaining service time $H_i^{rem}$ is dependent on the queue size, a high blocking probability can be interpreted as a sign of great dependence on the queue size. This is the reason for the deterioration of the approximation with respect to Figure 4(b).

A tabular presentation of the first moment $E[F]$, the coefficient of variation $c_F$, and the 95%-quantile (95%-QT) are given in Table 1 for the distributions depicted in Figures 3 – 5. There, the letters 'L', 'M', and 'H' are abbreviations for the traffic intensities used. From this table, we conclude that the absolute errors of the presented quantities are small.

Figure 5 High traffic intensity: $\lambda = 0.9 \cdot b$, $a = 4$
<table>
<thead>
<tr>
<th></th>
<th>threshold a</th>
<th>Ana.- Sim.</th>
<th>$E[F]$</th>
<th>$c_F$</th>
<th>95%-QT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deterministic(10)</td>
<td>4L</td>
<td>Ana.</td>
<td>28.769</td>
<td>0.70408</td>
<td>69</td>
</tr>
<tr>
<td>Deterministic(10)</td>
<td>4L</td>
<td>Sim.</td>
<td>28.758</td>
<td>0.70451</td>
<td>69</td>
</tr>
<tr>
<td>NegBin(10, 2)</td>
<td>4L</td>
<td>Ana.</td>
<td>31.019</td>
<td>0.98741</td>
<td>89</td>
</tr>
<tr>
<td>NegBin(10, 2)</td>
<td>4L</td>
<td>Sim.</td>
<td>31.147</td>
<td>0.98070</td>
<td>89</td>
</tr>
<tr>
<td>Deterministic(10)</td>
<td>8L</td>
<td>Ana.</td>
<td>53.750</td>
<td>0.67682</td>
<td>121</td>
</tr>
<tr>
<td>Deterministic(10)</td>
<td>8L</td>
<td>Sim.</td>
<td>53.733</td>
<td>0.67674</td>
<td>121</td>
</tr>
<tr>
<td>NegBin(10, 2)</td>
<td>8L</td>
<td>Ana.</td>
<td>54.236</td>
<td>0.77075</td>
<td>131</td>
</tr>
<tr>
<td>NegBin(10, 2)</td>
<td>8L</td>
<td>Sim.</td>
<td>54.355</td>
<td>0.76893</td>
<td>131</td>
</tr>
<tr>
<td>Deterministic(10)</td>
<td>4M</td>
<td>Ana.</td>
<td>15.332</td>
<td>0.22344</td>
<td>20</td>
</tr>
<tr>
<td>Deterministic(10)</td>
<td>4M</td>
<td>Sim.</td>
<td>14.745</td>
<td>0.24342</td>
<td>20</td>
</tr>
<tr>
<td>NegBin(10, 2)</td>
<td>4M</td>
<td>Ana.</td>
<td>38.085</td>
<td>1.1322</td>
<td>126</td>
</tr>
<tr>
<td>NegBin(10, 2)</td>
<td>4M</td>
<td>Sim.</td>
<td>37.880</td>
<td>1.1492</td>
<td>127</td>
</tr>
<tr>
<td>Deterministic(10)</td>
<td>4H</td>
<td>Ana.</td>
<td>15.437</td>
<td>0.24192</td>
<td>23</td>
</tr>
<tr>
<td>Deterministic(10)</td>
<td>4H</td>
<td>Sim.</td>
<td>15.241</td>
<td>0.20630</td>
<td>20</td>
</tr>
<tr>
<td>NegBin(10, 2)</td>
<td>4H</td>
<td>Ana.</td>
<td>58.839</td>
<td>0.92776</td>
<td>167</td>
</tr>
<tr>
<td>NegBin(10, 2)</td>
<td>4H</td>
<td>Sim.</td>
<td>58.290</td>
<td>0.93480</td>
<td>167</td>
</tr>
</tbody>
</table>

6 Conclusion

In this paper, we investigated a finite capacity queue with batch service and derived an approximate flow time distribution. From the numerical results, we conclude that our approximation of the flow time distribution of the $M/G^{[a,b]}/1 - S - FIFO$ queue is practically useful for low, medium, and also high offered traffic intensities ($\lambda \leq 0.95 \cdot b$) since the differences between the first moment derived from the approximate distribution and the exact one computed by Little's Law are small. In addition, the shape of the distributions computed analytically are fairly similar to those calculated by simulation. Thus, higher quantiles of distributions can be estimated with a high degree of accuracy.

Bibliography

Ackroyd, M. H. (1980). Computing the waiting time distribution for the G/G/1 queue by signal processing methods. IEEE Transactions on Communications COM-


