Delay Analysis of a Variable-Capacity Batch-Server Queue with General Class-Dependent Service Times

Jens Baetens[1,a)], Bart Steyaert[1], Dieter Claeys[1,2,3] and Herwig Bruneel[1]

[1]*Ghent University, Dept. of Telecommunications and Information Processing, SMACS Research Group, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium*
[2]*Ghent University, Dept. of Industrial Systems Engineering and Product Design, Technologiepark 903, Zwijnaarde, Belgium*
[3]*Department of Agile and Human Centered Production and Robotic Systems, Flanders Make*

[a)]Corresponding author: jens.baetens@ugent.be

Abstract. In manufacturing, a batch server groups multiple customers that require the same type of service based on a specific characteristic, such as temperature or destination. In this paper, we extend previous work with the analysis of the delay in a variable-capacity batch-service queueing system with general class-dependent service times and customer-based correlation in the arrival process. The impact of asymmetry and correlation in the arrival process on the mean delay of a random customer and the tail distribution of the delay is investigated as well.

INTRODUCTION

A machine in a production line is often capable of processing multiple products simultaneously. The number of customers (products) that are grouped together, also called the service (production) capacity, is often assumed to be a constant, e.g. Banerjee and Gupta [4]. In many applications, this service capacity varies in time and depends on the state of the server(s) and on the waiting customers. Germs and van Foreest [7] studied the $M(n)^{X(n)}/G(n)^{Y(n)}/1/K+B$ model where the arrival rate, service time distribution and service capacity depend on the number of customers waiting in the queue with size $K+B$. In our previous papers [1, 2, 3], we analysed a batch queueing system with a service capacity that not only depends on the number of waiting customers but also on their respective classes. Such a server is often found in production lines where a machine can process multiple product types with different characteristics like required temperature or color. Other applications of customer differentiation with variable-capacity queueing systems can be found in logistics, where all packets with the same destination can be transported together.

Differentiation between multiple customer types has already been studied in the context of polling systems and priority queueing systems (e.g. [8, 6]). In such types of systems, each class of customers has its own queue, and the order in which customers are processed, can be changed. While both of these techniques allow for optimizing the order in which the customers are processed, implementing these techniques is not always feasible due to a requirement for a global FCFS-service discipline or due to the increased cost of a more complex system. For this reason, we consider a common infinite-size queue for both classes, and do not allow customer reordering.

In this paper, we analyse the delay of a two-class discrete-time variable-capacity batch-service queueing model. In the arrival process, we introduce correlation between the classes of consecutive customers in order to model the common tendency for same-class customer to arrive in clusters by using the probabilities α and β as the probability that the class of the current customer is the same as the previous customer, which was respectively of class A or B. The number of customer arrivals in a random slot follows a general distribution with probability generating function (pgf) $E(z)$ with mean λ. The batch server in this model is capable of grouping and serving all consecutive same-class customers at the head of the queue, which yields a service process of variable capacity. The pgf of the class-dependent service time of a batch of class A (B) customers is assumed to be given by the function $S_A(z)$ ($S_B(z)$). In the numerical experiments in this paper, we focus on the impact of asymmetry and correlation in the arrival process on the mean delay and the tail distribution of the delay of a random customer.

ANALYSIS

In order to study the delay of a random customer for this model, which is the time between the customer's arrival and departure instants, represented by the variable d (with pgf $D(z)$), we must first calculate the stationary distribution of the number of customers in the queue upon arrival of the random customer. Using a spectral decomposition technique, we calculate the pgf of the delay of a random customer. The delay tail probability $\Pr[d > T]$ can be approximated by applying the dominant singularity approach. These results will only be valid when the system is stable. In [2], we have shown that the condition for stability for this system is given by $\lambda\left(S'_A(1) + S'_B(1)\right) < \frac{1}{1-\alpha} + \frac{1}{1-\beta}$.

Joint pgf of queue occupancy at customer arrival and remaining service time

In order to find the joint pgf of the queue occupancy at customer arrival epochs and the remaining service time, we need the probability that, during a random slot, the server is busy processing a class A or B batch (denoted by π_A or π_B) or idle and the previous service was of class A or B customers (denoted by $\pi_{I,A}$ or $\pi_{I,B}$). In [2], we found expressions for π_A, π_B and $\pi_I = \pi_{I,A} + \pi_{I,B}$. Using the partial pgfs $Q_A(z)$ and $Q_B(z)$ of the queue occupancy after service initiation of a class A or B batch, which are calculated in [2], the expressions for the state of a random slot can be written as

$$\begin{aligned}
\pi_A &= Q_A(1)S'_A(1)(1-E(0))/Q \ , \ \pi_B = Q_B(1)S'_B(1)(1-E(0))/Q \ , \\
\pi_{I,A} &= Q_A(0)S_A(E(0))/Q \ , \ \pi_{I,B} = Q_B(0)S_B(E(0))/Q \ , \\
Q &= Q_A(1)S'_A(1)(1-E(0)) + Q_B(1)S'_B(1)(1-E(0)) + Q_A(0)S_A(E(0)) + Q_B(0)S_B(E(0)) \ .
\end{aligned}$$

Using these probabilities, we can then calculate the partial joint pgf $N_A(z,x)$ of the queue occupancy and the remaining service time at customer arrival epochs if the customer at the head of the queue is of class A, leading to

$$\begin{aligned}
N_A(z,x) = &\frac{1-E(z)}{\lambda S'_A(1)S'_B(1)(1-z)(x-E(z))}\Big(S'_A(1)S'_B(1)(x-E(z))(\alpha\pi_{I,A} + (1-\beta)\pi_{I,B}) \\
&+ \frac{\pi_A}{Q_A(1)}q_A(0)\alpha S'_B(1)(S_A(x) - S_A(E(z))) + \frac{\pi_B}{Q_B(1)}(Q_B(z) - \beta q_B(0))S'_A(1)(S_B(x) - S_B(E(z)))\Big) \ .
\end{aligned}$$

A similar analysis for the case that the customer at the head of the queue is of class B results in

$$\begin{aligned}
N_B(z,x) := &\frac{1-E(z)}{\lambda S'_A(1)S'_B(1)(1-z)(x-E(z))}\Big(S'_A(1)S'_B(1)(x-E(z))(\beta\pi_{I,B} + (1-\alpha)\pi_{I,A}) \\
&+ \frac{\pi_B}{Q_B(1)}q_B(0)\beta S'_A(1)(S_B(x) - S_B(E(z))) + \frac{\pi_A}{Q_A(1)}(Q_A(z) - \alpha q_A(0))S'_B(1)(S_A(x) - S_A(E(z)))\Big) \ .
\end{aligned}$$

Analysis of the customer delay

We start the analysis of the delay by calculating $D_{A,n}(z)$ and $D_{B,n}(z)$, the partial pgfs of the customer delay given that there are n customers in the queue before the random arriving customer and the customer at the head of the queue is respectively of class A or B. Using recursion, we obtain that, starting from $D_{A,0}(z) := S_A(z)$ and $D_{B,0}(z) := S_B(z)$,

$$\begin{bmatrix} D_{A,n}(z) \\ D_{B,n}(z) \end{bmatrix} := \begin{bmatrix} \alpha & (1-\alpha)S_A(z) \\ (1-\beta)S_B(z) & \beta \end{bmatrix} \begin{bmatrix} D_{A,n-1}(z) \\ D_{B,n-1}(z) \end{bmatrix} = \boldsymbol{M}(z) \begin{bmatrix} D_{A,n-1}(z) \\ D_{B,n-1}(z) \end{bmatrix} = \boldsymbol{M}(z)^n \begin{bmatrix} S_A(z) \\ S_B(z) \end{bmatrix} \ .$$

The eigenvalues $\lambda_1(z)$ and $\lambda_2(z)$ of $\boldsymbol{M}(z)$ are given by

$$\lambda_{1,2}(z) = \frac{\alpha+\beta}{2} \pm \frac{1}{2}\sqrt{(\alpha-\beta)^2 + 4(1-\alpha)(1-\beta)S_A(z)S_B(z)} \ .$$

Similarly to the analysis in our previous paper [3], where the delay is analysed for a similar model without customer-based correlation in the arrival process, it can be proven that the branching points, where $\lambda_1(z) = \lambda_2(z)$, can be removed. Defining the matrices $\boldsymbol{R}(z)$ and $\boldsymbol{L}(z)$ as the right and left eigenvectors corresponding with the matrix $\boldsymbol{M}(z)$, we can show that

$$\boldsymbol{R}(z) = \begin{bmatrix} \frac{(1-\alpha)S_A(z)}{\lambda_1(z)-\alpha} & \frac{(1-\alpha)S_A(z)}{\lambda_2(z)-\alpha} \\ 1 & 1 \end{bmatrix} =: \begin{bmatrix} R_1(z) & R_2(z) \\ 1 & 1 \end{bmatrix} \ ,$$

and

$$\boldsymbol{L}(z)\begin{bmatrix}S_A(z)\\S_B(z)\end{bmatrix}=\boldsymbol{R}^{-1}(z)\begin{bmatrix}S_A(z)\\S_B(z)\end{bmatrix}=\begin{bmatrix}\frac{(1-\beta)S_A(z)S_B(z)+(\lambda_1(z)-\alpha)S_B(z)}{2\lambda_1(z)-\alpha-\beta}\\\frac{(1-\beta)S_A(z)S_B(z)+(\lambda_2(z)-\alpha)S_B(z)}{2\lambda_2(z)-\alpha-\beta}\end{bmatrix}=:\begin{bmatrix}L_1(z)\\L_2(z)\end{bmatrix}\ .$$

We now diagonalize the matrix $\boldsymbol{M}(z)$ and sum over all possible values of the queue length and remaining service times at customer arrival epochs, which leads to

$$D(z)=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\begin{bmatrix}n_A(n,m) & n_B(n,m)\end{bmatrix}\boldsymbol{R}(z)\begin{bmatrix}\lambda_1(z)^n & 0\\0 & \lambda_2(z)^n\end{bmatrix}\boldsymbol{L}(z)\begin{bmatrix}S_A(z)\\S_B(z)\end{bmatrix}z^m\ ,$$

where $n_A(n,m)$ and $n_B(n,m)$ are the probability mass functions corresponding to the partial joint pgfs $N_A(z,x)$ and $N_B(z,x)$ respectively. Working out both sums eventually leads to the following result

$$D(z)=\Big(R_1(z)N_A(\lambda_1(z),z)+N_B(\lambda_1(z),z)\Big)L_1(z)+\Big(R_2(z)N_A(\lambda_2(z),z)+N_B(\lambda_2(z),z)\Big)L_2(z)\ .$$

Tail probability of the delay

The tail distribution of the delay d of a random customer can be studied using an approximation technique described in Bruneel et al. [5]. This technique uses the inverse Z-transform to express the probability $\Pr[d>T]$ as a weighted sum of the negative n-th powers of the singularities of $D(z)$. Since $D(z)$ is an analytical function inside the complex unit disk $|z|<1$, all of the singularities will have a modulus larger than 1, which means that the probability $\Pr[d>T]$ is dominated by the singularity on the real axis with the smallest modulus. This dominant singularity is equal to the smallest zero on the real axis of the common denominator of $Q_A(\lambda_1(z))$ and $Q_B(\lambda_1(z))$, and is the solution of $z_1=E(\lambda_1(z_1))$. With this dominant singularity, we can approximate the pgf of $D(z)$ by

$$D(z)\approx K_1/(z_1-z)\ ,\quad K_1:=\frac{\Big(r_1(z_1)NN_A(\lambda_1(z_1),z_1)+NN_B(\lambda_1(z_1),z_1)\Big)L_1(z_1)}{DN'(\lambda_1(z_1))}\ ,$$

where $NN_A(z,x)$ and $NN_B(z,x)$ respectively represent the numerators of $N_A(z,x)$, $N_B(z,x)$, and $DN(z)$ and $DL_1(z)$ correspond with the denominators of $N_A(z)$ (and $N_B(z)$) and $L_1(z)$. For T sufficiently large, the probability $\Pr[d>T]$ can then be approximated by $\Pr[d>T]\approx K_1/(1-z_1)\cdot z_1^{-T-1}$.

NUMERICAL RESULTS

In the numerical examples, we use a geometrically distributed arrival process with mean λ, and the service times of a batch of class A and B customers are geometrically distributed as well with a mean of 3 slots. In Figure 1, we show the impact on the mean delay $E[d]$ of asymmetry, that is the difference between the probabilities that a random customer is of class A or B. In this figure, we vary the probability α while keeping $\alpha+\beta=1$, which implies an uncorrelated arrival process. We see that asymmetry has a significant impact for higher arrival rates. We also see that, for α very close to 0 or 1, the arrival rate only has a small impact on $E[d]$ which becomes equal to 3 slots, the mean service time of a batch. In Figure 2 for a system with $\lambda=0.4$, we see that a higher degree of asymmetry in the arrival process drastically decreases the slope of $\Pr[d>T]$.

In Figure 3, we show the impact of correlation on $E[d]$ in a symmetric arrival process by using $\alpha=\beta$. These results show that the impact of correlation is more significant when the system is operating near a point of instability. If there is a high degree of correlation, the system processes on average large sequences of customers, which means that a newly arrived customer will have a high probability to be processed in the next service period. The effect that correlation has on the tail of the delay is shown in Figure 4. We clearly see that increased correlation significantly decreases the slope of the tail. In contrast to Figure 2, we also observe that the impact of an increasing degree of correlation is significant for all values of α, while increasing the amount of asymmetry in the arrival process is relatively small when α is close to 0.5.

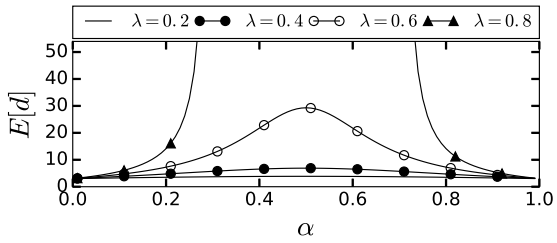

Fig. 1: Impact of asymmetry on the mean delay

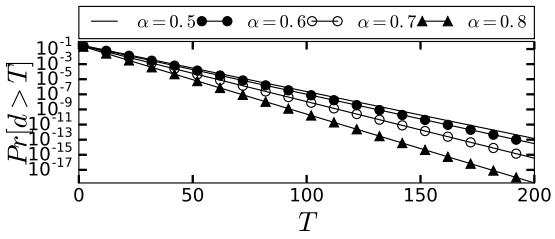

Fig. 2: Impact of asymmetry on the tail of the delay

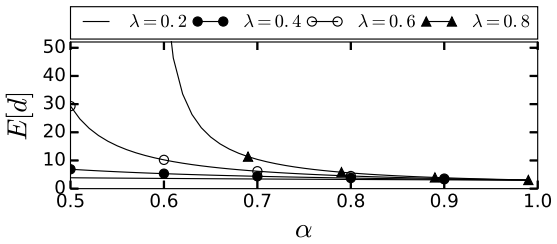

Fig. 3: Impact of correlation on the mean delay

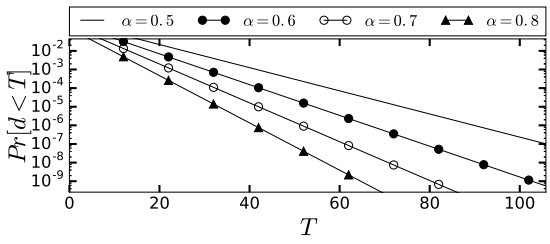

Fig. 4: Impact of correlation on the tail probability

CONCLUSIONS

In this paper, the pgf of the delay of a random customer has been calculated for a variable-capacity batch-service queueing system with class-dependent general service times and customer-based correlation. With the aid of some numerical examples, we have studied the impact of asymmetry and correlation in the arrival process on the mean delay and the tail distribution of the delay of a random customer.

A number of possible extensions on this model are possible. A first possibility is to include more than 2 customer classes. Another important extension is to introduce a maximum service capacity. Although this is an important feature to model realistic applications, we expect the model in this paper to be a good approximation assuming the system is operating under low to moderate loads, and the amount of correlation in the arrival process is not too high.

Dieter Claeys is a Postdoctoral Fellow with the Research Foundation Flanders, Belgium. Part of the research has been funded by the Interuniversity Attraction Poles Programma initiated by the Belgian Science Policy Office.

REFERENCES

[1] J. Baetens, B. Steyaert, D. Claeys, and H. Bruneel. System occupancy of a two-class batch-service queue with class-dependent variable server capacity. In *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, pages 32–44. Springer, 2016.

[2] J. Baetens, B. Steyaert, D. Claeys, and H. Bruneel. Analysis of a batch-service queue with variable service capacity, correlated customer types and generally distributed class-dependent service times. *Submitted*, 2017.

[3] J. Baetens, B. Steyaert, D. Claeys, and H. Bruneel. Delay analysis of a two-class batch-service queue with class-dependent variable server capacity. *Submitted to Mathematical Methods of Operations Research*, 2017.

[4] A. Banerjee and U.C. Gupta. Reducing congestion in bulk-service finite-buffer queueing system using batch-size-dependent service. *Performance Evaluation*, 69(1):53–70, 2012.

[5] Herwig Bruneel, Bart Steyaert, Emmanuel Desmet, and Guido H Petit. Analytic derivation of tail probabilities for queue lengths and waiting times in atm multiserver queues. *European Journal of Operational Research*, 76(3):563–572, 1994.

[6] J.L. Dorsman, R.D. Van der Mei, and E.M.M. Winands. Polling with batch service. *OR Spectrum*, 34:743–761, 2012.

[7] R. Germs and N.D. Van Foreest. Analysis of finite-buffer state-dependent bulk queues. *OR Spectrum*, 35(3):563–583, 2013.

[8] G.V.K. Reddy, R. Nadarajan, and P.R. Kandasamy. A nonpreemptive priority multiserver queueing system with general bulk service and heterogeneous arrivals. *Computers & operations research*, 20(4):447–453, 1993.