Polling with gated batch service

O. Boxma
EURANDOM & Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands. o.j.boxma@tue.nl

J. van der Wal
Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands. jan.v.d.wal@tue.nl

U. Yechiali
Dept. of Statistics and OR, Tel Aviv University, Israel. uriy@post.tau.ac.il

This paper considers a batch service polling system in which all customers in a queue present upon arrival of the server are served simultaneously. The service time is independent of the batch size. We study the joint steady-state queue-length distribution as well as the waiting time distribution at each of the queues.

1. Introduction

We consider polling systems where service is given in batches of unlimited size. When the server visits a queue, all customers present are served *in a single batch.* We call this *gated batch service.* The batch service time is independent of the size of the batch. Some examples of such systems are discussed in the literature below. Examples more related to manufacturing are ovens, transportation of material by one server to a number of different machines and, as an approximation, situations in which the set-up time is an order of magnitude larger than the actual production time.

Literature review

Polling systems with unlimited batch service have been studied much less extensively then those in which service is given to customers in a 'one at a time' fashion or in batches of limited size. Unlimited batch service models are considered in the context of teletext, videotex and TDMA systems, as well as for central data-base operations. Ammar and Wong [1987] studied a teletext system with N queues, fed by independent Poisson arrival streams. Service times in all queues are deterministic (slotted, unit time each), there are no switch-over times, and the service discipline is locally gated. They showed that the policy which minimizes mean response time is of a cyclic nature, with cycle length $L \geq N$ slots, in which queue i is visited k_i times, where $\sum_{i=1}^{N} k_i = L$. Yet, the problem of finding the exact length L was only partially resolved. Liu and Nain [1992] examined a TDMA model with both the locally gated and exhaustive regimes for the case of zero switching times and homogeneous arrival process to all queues. Dykeman et al. [1986] used Howard's policy-iteration algorithm to control a videotex system. They indicated that, even with equal and deterministic service requirements, and with no switching times, the structure of the optimal policy could be very complicated. Van Oyen and Teneketzis [1996] formulated a central data base system and an Automated Guided Vehicle as a polling system with an infinite-capacity batch service and zero switching times, where the controller observes only the length of the queue at which the server is located. Van der Wal and Yechiali [2003] explored dynamic server's visit-order policies in non-symmetric polling systems with switch-in and switch-out times, where service is in batches of unlimited size. They concentrated on so-called 'Hamiltonian tour' policies in which - in order to give a fair treatment to the various queues - the server attends every non-empty queue exactly once during each cycle. The server then dynamically generates a new visit schedule at the start of each

round, depending on the current state of the system and on the various non-homogeneous system parameters. Three service regimes were considered: Locally Gated, Exhaustive and Globally Gated, and 3 different performance measures were examined. For each combination of service regime and performance measure, the characteristics of the optimal Hamiltonian tour were derived. Some of the resulting optimal policies are elegant index-type rules; others are the solutions of NP-Hard problems; while special cases are reduced to Assignment problems with specific cost matrices.

Contribution of the paper

In the present paper we consider a polling system consisting of N queues which are visited in cyclic order by a single server. When the server polls a queue and it is not empty, all customers present upon arrival are served in one single batch. We study the joint steady-state queue-length distribution as well as the waiting time distribution at each of the queues.

2. Gated batch service

2.1. Preliminaries

We study the following polling model. A single server S cyclically visits N queues $Q_1, \ldots, Q_N$. Customers arrive at these queues according to independent Poisson processes, with rate λ_i at Q_i, $i = 1, \ldots, N$. If, upon the arrival of S at Q_i, there are $X_i^i > 0$ customers present at Q_i, then S serves exactly those customers, in one batch. The service time of this batch is a random variable, that we shall generically denote by B_i, with Laplace-Stieltjes Transform (LST) $\tilde{B}_i(\cdot)$. S subsequently switches to Q_{i+1}. The switch-over time of S from Q_i to the next queue is a random variable, that we shall generically denote by D_i, with LST $\tilde{D}_i(\cdot)$. The analysis that follows will also go through when we distinguish between switch-over times following visits with actual service and those following visits with null service (i.e., with length zero), but we shall make no such distinction. We shall furthermore make all the usual independence assumptions regarding the involved inter-arrival intervals, service times and switch-over times.

For this batch-service gated polling model, we determine the Probability Generating Function (PGF) of the joint steady-state queue length distribution, as well as the LST of the waiting time distribution of a class-i customer, $i = 1, \ldots, N$. Let us now introduce some further notation. In the sequel, $I_{[\cdot]}$ shall denote an indicator function. Furthermore, for $i = 1, \ldots, N$:

$$\begin{aligned} A_i(t) &= \text{number of arrivals to } Q_i \text{ during a time interval of length } t. \\ X_i^j &= \text{number of jobs in queue } Q_j \text{ when } Q_i \text{ is polled.} \\ V_i &= V_i(X_i^i) = B_i I_{[X_i^i > 0]} = \text{the visit time of } S \text{ to } Q_i. \\ G_i(z_1, \ldots, z_N) &= \mathbb{E}[\prod_{j=1}^{N} z_j^{X_i^j}]. \end{aligned}$$

It is easily seen that the following "laws of motion" hold for the X_i^j:

$$\begin{aligned} X_{i+1}^j &= X_i^j + A_j(V_i(X_i^i)) + A_j(D_i), \quad j \neq i, \\ X_{i+1}^j &= A_j(V_i(X_i^i)) + A_j(D_i), \quad j = i. \end{aligned} \tag{1}$$

While we present these laws of motion in terms of steady-state quantities, in reality we are expressing the number of jobs in Q_j at the nth visit of S to Q_{i+1} into that at Q_j at the nth visit of S to Q_i. So we look one queue ahead. By doing this N successive times, we can express the number of jobs in Q_j at the $(n+1)$th visit of S to Q_i into those at the nth visit of S to Q_i.

Introducing $\sigma(z_1, \ldots, z_N) = \sum_{j=1}^{N} \lambda_j (1 - z_j)$, it follows that, for $i = 1, \ldots, N$ (with $G_{N+1} = G_1$):

$$\begin{aligned} G_{i+1}(z_1, \ldots, z_N) &= \mathbb{E}[z_1^{X_i^1} \ldots z_{i-1}^{X_i^{i-1}} z_{i+1}^{X_i^{i+1}} \ldots z_N^{X_i^N} I_{[X_i^i > 0]}] \tilde{B}_i(\sigma(z_1, \ldots, z_N)) \tilde{D}_i(\sigma(z_1, \ldots, z_N)) \\ &\quad + \mathbb{E}[z_1^{X_i^1} \ldots z_{i-1}^{X_i^{i-1}} z_{i+1}^{X_i^{i+1}} \ldots z_N^{X_i^N} I_{[X_i^i = 0]}] \tilde{D}_i(\sigma(z_1, \ldots, z_N)) \\ &= G_i(z_1, \ldots, z_{i-1}, 1, z_{i+1}, \ldots, z_N) \tilde{B}_i(\sigma(z_1, \ldots, z_N)) \tilde{D}_i(\sigma(z_1, \ldots, z_N)) \\ &\quad + G_i(z_1, \ldots, z_{i-1}, 0, z_{i+1}, \ldots, z_N)[1 - \tilde{B}_i(\sigma(z_1, \ldots, z_N))] \tilde{D}_i(\sigma(z_1, \ldots, z_N)). \end{aligned} \tag{2}$$

To develop insight into the structure of the solution of this recursion, we first consider the special case of $N = 2$ queues in Subsection 2.2; the general case will subsequently be solved in Subsection 2.3.

2.2. The two-queue case

For $N = 2$, Formula (2) becomes:

$$\begin{aligned} G_1(z_1,z_2) &= G_2(z_1,1)\tilde{B}_2(\sigma(z_1,z_2))\tilde{D}_2(\sigma(z_1,z_2)) \\ &\quad + G_2(z_1,0)[1-\tilde{B}_2(\sigma(z_1,z_2))]\tilde{D}_2(\sigma(z_1,z_2)), \end{aligned} \tag{3}$$

$$\begin{aligned} G_2(z_1,z_2) &= G_1(1,z_2)\tilde{B}_1(\sigma(z_1,z_2))\tilde{D}_1(\sigma(z_1,z_2)) \\ &\quad + G_1(0,z_2)[1-\tilde{B}_1(\sigma(z_1,z_2))]\tilde{D}_1(\sigma(z_1,z_2)). \end{aligned} \tag{4}$$

It follows from (4) that $G_2(z_1,1)$ is expressed in $G_1(1,1)$ and $G_1(0,1)$; similarly, $G_2(z_1,0)$ is expressed in $G_1(1,0)$ and $G_1(0,0)$. By substituting (4) with $z_2 = 1$ (respectively, $z_2 = 0$) into (3), we are able to express $G_1(z_1,z_2)$ into known terms plus the four unknown constants $G_1(1,1)$ (which actually equals 1), $G_1(0,1)$, $G_1(1,0)$ and $G_1(0,0)$:

$$\begin{aligned} G_1(z_1,z_2) &= \{G_1(1,1)\tilde{B}_1(\sigma(z_1,1))\tilde{D}_1(\sigma(z_1,1)) \\ &\quad + G_1(0,1)[1-\tilde{B}_1(\sigma(z_1,1))]\tilde{D}_1(\sigma(z_1,1))\} \\ &\quad \times \tilde{B}_2(\sigma(z_1,z_2))\tilde{D}_2(\sigma(z_1,z_2)) \\ &\quad + \{G_1(1,0)\tilde{B}_1(\sigma(z_1,0))\tilde{D}_1(\sigma(z_1,0)) \\ &\quad + G_1(0,0)[1-\tilde{B}_1(\sigma(z_1,0))]\tilde{D}_1(\sigma(z_1,0))\} \\ &\quad \times [1-\tilde{B}_2(\sigma(z_1,z_2))]\tilde{D}_2(\sigma(z_1,z_2)). \end{aligned} \tag{5}$$

It remains to determine $G_1(0,1)$, $G_1(1,0)$ and $G_1(0,0)$. Those three unknown constants may be found by the substitutions $\{z_1 = 0, z_2 = 1\}$, $\{z_1 = 1, z_2 = 0\}$ and $\{z_1 = 0, z_2 = 0\}$ into (5), resulting in three linear equations with three unknowns.

The above yields the following insight. To determine the PGF $G_i(z_1,z_2)$, what really matters is whether a queue is empty or not when server S visits it. If it is non-empty, the actual queue size does not have an effect on the visit time. Hence the joint queue length distribution at a visit epoch of S at, say, Q_1 is determined by the four possible events *both Q_1 and Q_2 non-empty at the last previous visit of S to Q_1*, ..., *both Q_1 and Q_2 empty at the last previous visit of S to Q_1*. Q_1 being non-empty at the previous visit has probability $\mathbb{P}(X_1^1 > 0) = G_1(1,1) - G_1(0,1) = 1 - G_1(0,1)$, etc.

It should be noticed that the process $\{(U_1^{(n)}, U_2^{(n)}),\ n = 1,2,\ldots\}$, with $U_i^{(n)} = 1$ (0) denoting that Q_i is non-empty (resp., empty) at the nth visit of S to Q_1 is a two-dimensional Markov chain. This Markov chain is irreducible, aperiodic and positive-recurrent, and hence has a unique non-negative steady-state solution. With an obvious notation, we have: $\mathbb{P}(U_1 = 1, U_2 = 1) = 1 - G_1(1,0) - G_1(0,1) + G_1(0,0)$, ..., $\mathbb{P}(U_1 = 0, U_2 = 0) = G_1(0,0)$.

2.3. The N-queue case

The insight obtained in the previous subsection for the case of 2 queues readily allows us to obtain the structure of the solution of the case of an arbitrary number of queues. N successive substitutions of (2) result in an expression of $G_1(z_1,\ldots,z_N)$ into the 2^N unknown constants $G_1(1,1,\ldots,1)$, ..., $G_1(0,0,\ldots,0)$. These 2^N constants (of which the first actually equals 1) can be obtained by determining the unique steady-state solution of an N-dimensional irreducible, aperiodic and positive-recurrent Markov chain $\{(U_1^{(n)},\ldots,U_N^{(n)}),\ n = 1,2,\ldots\}$, with $U_i^{(n)} = 1$ (0) denoting that Q_i is non-empty (resp. empty) at the nth polling instant of S to Q_1.

The rationale behind this solution structure is that, for determining the steady-state joint queue length distribution at a visit of S to Q_1, what really matters is whether $Q_1,\ldots,Q_N$ were empty or

not at the last previous visit of S to Q_1; not what their actual queue lengths were. The probabilities of those events are obtained by solving an N-dimensional Markov chain with 2^N states.

Remark
It easily follows from (1) that the mean number of customers in Q_j when S polls Q_i, $f_i^j := \mathbb{E}X_i^j$, satisfies (with $\mathbb{E}V_i$ the mean visit period of S at Q_i):

$$\begin{aligned} f_{i+1}^j &= f_i^j + \lambda_j \mathbb{E}V_i + \lambda_j \mathbb{E}D_i, \quad j \neq i, \\ f_{i+1}^j &= \lambda_j \mathbb{E}V_i + \lambda_j \mathbb{E}D_i, \quad j = i. \end{aligned} \tag{6}$$

Summing (6) over all i yields:

$$f_j^j = \lambda_j \sum_{i=1}^{N} (\mathbb{E}V_i + \mathbb{E}D_i), \tag{7}$$

where $\mathbb{E}V_i = \mathbb{P}(X_i^i > 0)\mathbb{E}B_i = [G_i(1,\ldots,1,\ldots,1) - G_i(1,\ldots,0,\ldots,1)]\mathbb{E}B_i$, the 1 (resp. 0) appearing at the ith position. Notice that those $G_i(\ldots)$ have to be determined via the solution of a Markov chain, as discussed above. Also notice that f_j^j equals the mean number of arrivals at Q_j during one cycle time and that, via (6), f_i^j is readily expressed in f_j^j and the mean visit periods at $Q_j, \ldots, Q_{i-1}$. In particular, focussing on the number of customers in Q_1, f_1^1 is given by (7) while

$$f_i^1 = \lambda_1 \sum_{k=1}^{i-1} (\mathbb{E}V_k + \mathbb{E}D_k), \quad i = 2, \ldots, N. \tag{8}$$

2.4. Waiting times

In this subsection we study the waiting time W_i of an arbitrary customer at Q_i in steady state. First we make the following observation about queue lengths. Once the PGF $G_i(z_1, \ldots, z_N)$ of the joint queue length distribution when S polls Q_i has been determined for $i = 1, \ldots, N$, it is straightforward to derive the PGF of the joint queue length distribution at the instant at which S begins a switch-over time from Q_i to the next queue, $i = 1, \ldots, N$. Subsequently, it is not hard to determine the PGF $G_i^{visit}(\cdot)$ of the joint queue length distribution during a visit to Q_i (respectively, the PGF $G_i^{switch}(\cdot)$ of the joint queue length distribution during a switch from Q_i to Q_{i+1}). Taking an appropriate weighted average, one finally obtains the PGF of the joint steady-state queue length distribution, and hence also the mean steady-state queue length at any queue Q_i. That brings us back to waiting times: Application of Little's formula yields the mean time a type-i customer spends in the system (waiting plus in service).

It is somewhat more complicated to derive the (LST of the) waiting time *distribution*. Consider a tagged type-i customer. Conditioning on the type of interval during which that tagged customer arrived: a visit of S to Q_j, or a switch-over from Q_j to Q_{j+1}, one can determine the conditional waiting time LST. We refrain from working out the details. Instead, we sketch the approach by determining the conditional LST of the waiting time of a tagged type-3 customer who has arrived during a non-empty visit of S to Q_1 (respectively, during a switch of S from Q_1 to Q_2). Let $p_{1,2}^{(0)} = G_1^{visit}(1, 0, 1, \ldots, 1)$ denote the probability that Q_2 is empty at an arbitrary visit epoch of S to Q_1. Similarly, define $q_{1,2}^{(0)} = G_1^{switch}(1, 0, 1, \ldots, 1)$ to be the probability that Q_2 is empty at an arbitrary switch-over epoch of S from Q_1 to Q_2. Let $B_1^{(res)}$ denote the residual part (overshoot) of the ongoing batch service at Q_1, with density $\mathbb{P}(B_1 > x)/\mathbb{E}B_1$. We define $D_1^{(res)}$ similarly. Then

$$\begin{aligned} &\mathbb{E}[\mathrm{e}^{-\omega W_3} | \text{the tagged customer arrived while } S \text{ was serving } Q_1] \\ &= (1 - p_{1,2}^{(0)}) \mathbb{E}[\mathrm{e}^{-\omega(B_1^{(res)} + D_1 + B_2 + D_2)}] \\ &+ p_{1,2}^{(0)} \int_0^\infty \mathrm{e}^{-\omega t} \mathrm{d}\mathbb{P}(B_1^{(res)} + D_1 < t) \mathrm{e}^{-\lambda_2 t} \mathbb{E}[\mathrm{e}^{-\omega D_2}] \end{aligned}$$

$$
\begin{aligned}
&+ p_{1,2}^{(0)} \int_0^\infty \mathrm{e}^{-\omega t} \mathrm{d}\mathbb{P}(B_1^{(res)} + D_1 < t)(1 - \mathrm{e}^{-\lambda_2 t})\mathbb{E}[\mathrm{e}^{-\omega(B_2+D_2)}] \\
&= \mathbb{E}[\mathrm{e}^{-\omega(B_1^{(res)}+D_1+B_2+D_2)}] \\
&+ p_{1,2}^{(0)}\mathbb{E}[\mathrm{e}^{-(\omega+\lambda_2)(B_1^{(res)}+D_1)}]\mathbb{E}[\mathrm{e}^{-\omega D_2}][1 - \mathbb{E}[\mathrm{e}^{-\omega B_2}]]. \qquad (9)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\mathbb{E}[\mathrm{e}^{-\omega W_3} | \text{the tagged customer arrived while } S \text{ was switching from } Q_1 \text{ to } Q_2] \\
&= (1 - q_{1,2}^{(0)})\mathbb{E}[\mathrm{e}^{-\omega(D_1^{(res)}+B_2+D_2)}] \\
&+ q_{1,2}^{(0)} \int_0^\infty \mathrm{e}^{-\omega t} \mathrm{d}\mathbb{P}(D_1^{(res)} < t)\mathrm{e}^{-\lambda_2 t}\mathbb{E}[\mathrm{e}^{-\omega D_2}] \\
&+ q_{1,2}^{(0)} \int_0^\infty \mathrm{e}^{-\omega t} \mathrm{d}\mathbb{P}(D_1^{(res)} < t)(1 - \mathrm{e}^{-\lambda_2 t})\mathbb{E}[\mathrm{e}^{-\omega(B_2+D_2)}] \\
&= \mathbb{E}[\mathrm{e}^{-\omega(D_1^{(res)}+B_2+D_2)}] + q_{1,2}^{(0)}\mathbb{E}[\mathrm{e}^{-(\omega+\lambda_2)D_1^{(res)}}]\mathbb{E}[\mathrm{e}^{-\omega D_2}][1 - \mathbb{E}[\mathrm{e}^{-\omega B_2}]]. \qquad (10)
\end{aligned}
$$

References

[1]M.H. Ammar and J.W. Wong, "On the Optimality of Cyclic Transmission in Teletext Systems", IEEE Transactions on Communications, Vol. COM-35, No. 1, pp. 68–73 (1987).

[2]H.D. Dykeman, M.H. Ammar and J.W. Wong, "Scheduling Algorithms for Videotex Systems under Broadcast Delivery", in Proceedings of the International Conference on Communications (ICC'86), pp. 1847–1851 (1986).

[3]Z. Liu and P. Nain, "Optimal Scheduling in Some Multiqueue Single-Server Systems", IEEE Transactions on Automatic Control, Vol. 37, No. 2, pp. 247–252 (1992).

[4]J. van der Wal and U. Yechiali, "Dynamic Visit-Order Rules for Batch-Service Polling", Probability in the Engineering and Informational Sciences, Vol. 17, pp. 351-367 (2003).

[5]M.P. Van Oyen and D. Teneketzis, "Optimal Batch Service of a Polling System under Partial Information", Methods and Models in OR, Vol. 44, No. 3, pp. 401–419 (1996).