Asymptotics for Polling Models with Limited Service Policies

Woojin Chang
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205
USA

Douglas G. Down
Department of Computing and Software
McMaster University
1280 Main Street West
Hamilton, Ontario L8S 4L7
Canada

Abstract

In this paper we find exact asymptotic expressions for the event that the total queue length is large for a k_i-limited exponential polling model with equal service rates and two classes of customers. It is found that this behaviour divides into two very different regimes, depending on the arrival rates to the system.

1 Introduction

Polling models have seen wide application in modelling systems in the diverse areas of telecommunications, transportation, computer performance, inventory, etc. (for a couple of general references, see Boxma and Takagi [1] and Levy and Sidi [5]). A (single server) polling model is simple to describe: there are several classes of customers arriving to a single server which operates under a particular service policy. Some typical examples are the so-called *exhaustive* policy, in which the server empties the system of a class of customers before moving on to the next class and *limited* policies, in which the server has a limit on the number of customers of a class that can be served before moving on to the next class. The latter is the focus of this paper.

Over the last several years, there has been a flurry of activity in the analysis of rare events (or large deviations) in stochastic systems. It has been argued that in many practical systems, a useful performance measure is the frequency of rare events, such as data loss in a telecommunications system or large delays in filling an order at a warehouse. Thus it is somewhat surprising that there is a dearth of results for rare events in polling models. The authors are aware of only two papers addressing this problem (Choudhury and Whitt [2] and Duffield [3]) and these deal specifically with polling models under exhaustive-type service policies. The first of these papers uses numerical inversion techniques on the moment generating function for the invariant distribution of the system while the latter uses more "traditional" large deviations techniques. We are not

aware of any articles that analyze rare events for polling models with limited service policies. In fact, [2] states that limited service policies do not fit certain structural assumptions that allows their analysis to proceed. As the same structural assumptions are employed in [3], it appears that these techniques will have difficulty seeing application to limited service policies.

Recently, McDonald [6] has developed a methodology for analyzing rare events for Markov chains evolving on an orthant. We have found that the analysis of polling models with limited service policies is tailor-made for this methodology. The crux of this methodology is that engineering insight must be employed to identify possible large deviations paths a priori and then verify some technical conditions. We have found that it is not difficult to identify these paths and that the required technical conditions tend to be fairly straightforward to verify. In this paper, we hope to give an indication of what kind of results may be achieved for a system with two classes of customers served under the k_i-limited service policy, where the service requirements for both classes of customers are the same. This model will be made more precise in the next section.

The layout of the paper is as follows. In Section 2, we present the model and the appropriate Markov chain that describes its behavior. Section 3 gives the main results, while Section 4 discusses McDonald's methodology and its application to our system (thus providing the justification of the results in the previous section).

2 The Model

We consider a system with 2 distinct classes of customers. Customers of class 1 and class 2 arrive to the system according to independent Poisson processes with rates λ_1 and λ_2, respectively. Without loss of generality, we will assume that the classes have been labelled such that $\lambda_2 > \lambda_1$. There is one server and the service times for a customer (regardless of class) are exponentially distributed with rate μ. The service time for a customer is independent of all other service times as well as the arrival processes. Service is performed in a *k_i-limited* manner. That is, nominally when the server is visiting class i, it serves exactly k_i customers of class i (if there are at least k_i customers available) before visiting the other class. The service policy is completely specified upon noting that it is non-idling, i.e. if there are no customers of class 1 (class 2) in the system, the server continues to serve class 2 (class 1) until it can return to nominal operation. Service is non-preemptive.

The process $Q(t) = (Q_1(t), Q_2(t), Z(t), I(t))$ is a continuous time Markov chain, where $Q_i(t)$ is the number of customers of class i in the system at time t (including the customer in service, if any), $Z(t)$ is the class that is currently being served by the system and $I(t)$ is the number of service completions during the current server visit (if this reaches $k_i - 1$ and the other queue is empty, subsequent services will not increment this value). We set $Z(t) = 0$ if no customers are in the system. The Markov chain $Q(t)$ thus evolves on $S = Z_+ \times Z_+ \times \{0, 1, 2\} \times \{0, \cdots, max(k_1, k_2) - 1\}$. For the rest of this paper, we will look at the equivalent *uniformized* chain, $Q[n]$, where we assume without loss of generality that we have rescaled time such that $\lambda_1 + \lambda_2 + \mu = 1$. Let the transition kernel for this chain be denoted by K, where $K(x, y)$ gives the probability of moving to state y, given the chain is currently in state x. For example, if $x = (i, j, 2, k)$, with $i \geq 0$, $j > 0$, $k_2 > 1$ and $k \leq k_2 - 2$,

$$\begin{aligned} K(x, (i+1, j, 2, k)) &= \lambda_1 \\ K(x, (i, j+1, 2, k)) &= \lambda_2 \\ K(x, (i, j-1, 2, k+1)) &= \mu. \end{aligned}$$

By comparing this system with an M/M/1 queue, the existence of an invariant distribution π_Q is guaranteed if

$$\rho := (\lambda_1 + \lambda_2)/\mu < 1.$$

From this point on, we shall assume that this stability condition holds.

The event that we will be interested in in this paper is that the total system size is ℓ, i.e. the performance measure of interest is $F_\ell = \mathrm{P}\{Q_1[n] + Q_2[n] = \ell\}$. The procedure is similar if we were interested in related events such as the probability that one particular queue length is large.

3 Main Results

We will find that there are two distinct types of behaviour, which depend on the following condition

$$\frac{(k_1+k_2)\lambda_1\mu}{k_1(\lambda_1+\lambda_2)^2}. \tag{1}$$

If (1) is less than 1, we call the behaviour *single-class dominant*, for reasons that we will later see. If (1) is greater than or equal to 1, we will call the behaviour *non single-class dominant*.

We say $f \sim g$ if $f(\ell)/g(\ell) \to 1$ as $\ell \to \infty$. Let T_ℓ denote the first time that there are ℓ or more customers in the system.

Theorem 1 *For a single-class dominant system,*
(i)

$$\mathrm{E}[T_\ell | Q(0) = (0,0,0,0)] \sim g_1^{-1}\rho^{-\ell}$$

where the constant g_1 may be obtained by a fast simulation and is given in (5) below.
(ii) For non-negative integers ℓ and j, $k \in \{1,2\}$, and $m \in \{0, \cdots, max(k_1,k_2)-1\}$

$$\mathrm{P}\{Q_2(t) = \ell - j, Q_1(t) = j, Z(t) = k, I(t) = m\} \tag{2}$$
$$\sim \frac{g_1}{\mu - (\lambda_1+\lambda_2)}\rho^\ell \varphi(j,k,m)$$

We can get the required values of φ by (fast) simulation. When $k_1 = k_2 = 1$, we can compute φ exactly. (Note that this expression is somewhat complicated and will not be given here. For this and details of proofs throughout, as well as numerical work on a buffer allocation problem, a full paper is in preparation and will be available by the time these proceedings appear. Please contact the second author for a copy.)

Theorem 2 *For a non single-class dominant system*
(i)

$$\mathrm{E}[T_\ell | Q(0) = (0,0,0,0)] \sim g_2^{-1}\rho^{-\ell}$$

where the constant g_2 may be obtained by a fast simulation and is given in (6) below.
(ii) For any non-negative integer ℓ, $k \in \{1,2\}$, and $m \in \{0, \cdots, max(k_1,k_2)-1\}$

$$\mathrm{P}\{Q_1(t) + Q_2(t) = \ell, Z(t) = k, I(t) = m\} \tag{3}$$
$$\sim \frac{g_2}{(k_1+k_2)(\mu - (\lambda_1+\lambda_2))}\rho^\ell.$$

(iii)

$$\lim_{\ell\to\infty}\left(\frac{Q_1(T_\ell)}{\ell}, \frac{Q_2(T_\ell)}{\ell}\right) \tag{4}$$
$$\longrightarrow \left(\frac{\lambda_1\rho^{-1} - \frac{k_1(\lambda_1+\lambda_2)}{k_1+k_2}}{\mu-(\lambda_1+\lambda_2)}, \frac{\lambda_2\rho^{-1} - \frac{k_2(\lambda_1+\lambda_2)}{k_1+k_2}}{\mu-(\lambda_1+\lambda_2)}\right)$$

At this point, several remarks are in order. We have provided exact asymptotics for the k_i-limited system (they are not quite closed form, but g_1 and g_2 may be obtained by fast simulations). The fact that the rate in (2) and (3) is simply ρ^ℓ should not be surprising, as $Q_1(t) + Q_2(t)$ acts like the queue length in an M/M/1 system with arrival rate $\lambda_1 + \lambda_2$ and service rate μ.

Given that the rate is easily calculated, the usefulness of our results is twofold. First, we have identified the region of the parameter space for which the two different types of asymptotic behavior occur. For the single-class dominant system, looking at (2), we can see that $Q_1(t) + Q_2(t)$ reaches some large value ℓ with $Q_1(t)$ staying "close" to zero. Of course, we have gone much further than this by explicitly giving the distributions of the two queue

lengths and the position of the server when this large level has been hit. This latter explicit solution is the second useful feature.

For the non single-class dominant system, we see that on the approach to F_ℓ both queue lengths get "large". Thus the asymptotics are essentially those of an M/M/1 queue, with the additional information contained in (4) which indicates where we expect to hit the line $Q_1(t)+Q_2(t)=\ell$.

We would expect these results to be useful for such problems as buffer allocation. Unfortunately, we have yet to find a physical interpretation for the condition (1) in terms of the original network. Rather, the interpretation is in terms of one of the constructed chains ($\mathcal{W}^\infty$). Such insight could be useful for system design.

4 Justification of Main Results

In the interests of space, we have chosen not to go into detail in describing the methodology of McDonald [6]. There is an excellent summary of the methodology given in Foley and McDonald [4], where it is applied to a particular Join the Shortest Queue system. It would probably be useful to the reader to have [4] in front of herself when reading this section.

4.1 Single-class dominant (Theorem 1)

Recall that we defined the system to be class 2 dominant if $(k_1+k_2)\lambda_1\mu(\lambda_1+\lambda_2)^{-2} < k_1$. We expect F_ℓ to be reached only through the impact of class 2 customers. To this end, we first define a chain W such that $W[n] = (\tilde{W}[n], \hat{W}[n])$ such that the components of $\tilde{W}[n]$ diverge as ℓ gets large and $\tilde{W}_1[n] \geq \ell$ when the event F_ℓ occurs. In this case, we simply choose $\tilde{W}[n]$ to be $Q_1[n]+Q_2[n]$ and $\hat{W}[n]$ to be the ordered triple $(Q_1[n], Z[n], I[n])$. The transition kernel for this new chain is denoted K_W. An example of the structure of K_W is, if $x=(i,j,2,k_2-1)$, with $i,j>0$,

$$\begin{aligned} K_W(x,(i+1,j,2,k_2-1)) &= \lambda_1 \\ K_W(x,(i+1,j+1,2,k_2-1)) &= \lambda_2 \\ K_W(x,(i-1,j,1,0)) &= \mu. \end{aligned}$$

Next, we construct the "free" chain (W^∞) which simply means removing the boundaries corresponding to those components in $\tilde{W}$ and forcing W^∞ to have an additive structure. In our example, this boundary is denoted $\triangle$ and is all states of the form $(0,0,0,0)$, $(x+1,x,2,k)$, where $x \geq 0$ and $0 \leq k < k_2-1$, and $(x,x,1,k_1-1)$ where $x>1$. Perhaps this change is best described for our example by saying that W^∞ is constructed from W by removing the boundary that queue 2 must be nonnegative, i.e. the server *always* serves k_2 customers at queue 2 and a negative number of customers is allowed. After doing an appropriate change of measure, we will find that with positive probability that expanding the state space in this manner does not change the state trajectory. In this case, we let K^∞ be the transition kernel for W^∞ so for example, if $x=(i+1,i,2,k_2-3)$ with $i>0$, $k_2>3$

$$\begin{aligned} K^\infty(x,(i+2,i+1,2,k_2-3)) &= \lambda_1 \\ K^\infty(x,(i+2,i,2,k_2-3)) &= \lambda_2 \\ K^\infty(x,(i,i,2,k_2-2)) &= \mu. \end{aligned}$$

Note that $(i,i,2,k_2-2)$ is not even a valid state for the original W (or Q)-chain. The chain W^∞ evolves on the expanded state space $Z\times Z_+\times\{1,2\}\times\{0,\cdots,max(k_1,k_2)-1\}$ and has the following additive structure

$$\begin{aligned} &K^\infty((i,(k,w,\gamma)),(j,(l,v,\delta))) \\ &\quad = P^\infty((k,w,\gamma);(j-i,(l,v,\delta)). \end{aligned}$$

Next, we construct the *twisted free* ($\mathcal{W}^\infty$) chain. The idea here is that we wish to twist the underlying distributions such that our choice of the most likely path to F_ℓ is emphasized. To do this we need to find a harmonic function for the free process of the form $h(\tilde{x},\hat{x}) = \alpha^{\tilde{x}_1}\hat{a}(\hat{x})$, where $\tilde{x}$ corresponds to $\tilde{W}^\infty$, $\hat{x}$ corresponds to $\hat{W}^\infty$ and $\tilde{x}_1$ is the first component of $\tilde{x}$. In this case, it is straightforward to verify that $\alpha=\rho^{-1}$ and $\hat{a}(\hat{x})\equiv 1$ is an appropriate harmonic function by verifying that $\mathrm{E}[h(W^\infty[n+$

1])$|W^\infty[n]] = h(W^\infty[n])$. The transition kernel for $\mathcal{W}^\infty$, $\mathcal{K}^\infty$, is constructed by twisting K^∞ as follows: $\mathcal{K}^\infty(x,y) = K^\infty(x,y)h(x)/h(y)$.

Finally, we need to show that technical Conditions 0-7 in [4] hold.

Condition 0. Let $\tilde{d}_1 = \mathrm{E}[\tilde{\mathcal{W}}^\infty[n+1]] - \mathrm{E}[\tilde{\mathcal{W}}^\infty[n]]$. We must verify $\tilde{d}_1 > 0$, or in other words that the twisted free process drifts towards F_ℓ. In our case, $\tilde{d}_1 = \mu - (\lambda_1 + \lambda_2)$, which is positive by assumption.

Condition 1. Here, we must show that $\hat{\mathcal{W}}^\infty$ is irreducible and has an invariant distribution φ. Irreducibility is obvious, so we will now sketch why φ exists. In preparation to doing this, let $\tilde{\lambda}_i = \lambda_i/\rho$ and $\tilde{\mu} = \mu\rho$. By sampling the chain when the server returns to queue 1, we have the following results for $\mathcal{W}^\infty$. First, let $\sigma(n)$ denote the time of the nth visit of the server to class 1 and $\hat{X}_n(1)$ be the number of class 1 customers in the queue at time $\sigma(n)$, for the chain $\tilde{\mathcal{W}}^\infty$. Also, let $A_n(i)$ be the number of arrivals at time $\sigma(n)$ to class i.

$$\begin{aligned}\mathrm{E}[\sigma(n+1) - \sigma(n) | \hat{X}_n(1) \geq k_1] &= \frac{(k_1+k_2)}{\tilde{\mu}} \\ \mathrm{E}[A_{n+1}(1) - A_n(1) | \hat{X}_n(1) \geq k_1] &= \frac{(k_1+k_2)}{\tilde{\mu}}\tilde{\lambda}_1\end{aligned}$$

Lemma 1 *The chain $\hat{\mathcal{W}}^\infty$ is positive recurrent (and thus has an invariant distribution φ) if*

$$\frac{(k_1+k_2)}{\tilde{\mu}}\tilde{\lambda}_1 < k_1.$$

Remark. The proof will not be given here, but a sketch may be quickly given. First note that on substituting for $\tilde{\lambda}_1$ and $\tilde{\mu}$, this is exactly the single-class dominance condition. Also, this result should be intuitively clear, as between times $\sigma(n+1)$ and $\sigma(n)$, k_1 customers of class 1 may be served, while the expected number of class 1 arrivals is at most $\frac{k_1+k_2}{\tilde{\mu}}\tilde{\lambda}_1$. The proof consists of two parts. First, that this condition is indeed the right condition for the sampled chain, then connecting positive recurrence of the sampled chain to that of $\hat{\mathcal{W}}^\infty$.

Condition 2. It is easy to verify that $\mathcal{W}^\infty$ is aperiodic.

Condition 3. This is subsumed by Condition 0.

Condition 4. Here we wish to verify that with positive probability $\mathcal{W}^\infty$ wanders away from Δ without returning to Δ. This follows from the fact that $\tilde{d}_1$ is positive and $\hat{\mathcal{W}}^\infty$ drifts towards 0.

Condition 5. Verify

$$\sum_{\hat{z}\in Z_+\times\{1,2\}\times\{0,\ldots,\max(k_1,k_2)-1\}} \varphi(\hat{z})/\hat{a}(\hat{z}) < \infty.$$

This is trivial as $\hat{a}(\cdot) \equiv 1$.

Condition 6. We need to verify

$$\lambda(x) = \sum_{z\in\Delta} \pi_W(z)K_W(z,x)h(x)\chi\{x \in S/\Delta\} < \infty.$$

This tends to be the most difficult condition to check. In our case,

$$\begin{aligned}\lambda(x) &= \pi_W(0,0,0,0)(\lambda_2\alpha\mathbf{1}\{k_2=1\} + \lambda_1\alpha) \\ &\quad + \sum_{k=0}^{k_2-2}\sum_{x>0}\pi_W(x+1,x,2,k)\lambda_2\alpha^{x+2} \\ &\quad + \sum_{x=2}^{\infty}\pi_W(x,x,1,k_1-1)\lambda_2\alpha^{x+1}\end{aligned}$$

which is equivalent to checking

$$\begin{aligned}\sum_{k=0}^{k_2-2}\sum_{x>0}\pi_Q(x,1,2,k)\alpha^x &< \infty \\ \sum_{x>0}\pi_Q(x,0,1,k_1-1)\alpha^x &< \infty\end{aligned}$$

where π_Q is the invariant distribution for the Markov chain Q. It turns out that this can be done by comparing this system with an appropriate M/M/1 queue.

Condition 7. This involves checking a uniform integrability condition, here we simply state that given $\hat{a}(\cdot) \equiv 1$, the condition is trivially satisfied.

Thus we can assert the results of Theorem 1. In its statement, there is a constant g_1 in several of the expressions. We have now developed enough notation to give its definition.

$$\begin{aligned} g_1 &= \pi_W(0,0,0,0)\lambda_1\rho^{-1}H(1,1,1,0) \\ &+ \pi_W(0,0,0,0)\lambda_2\rho^{-1}H(1,0,2,0)\mathbf{1}\{k_2=1\} \\ &+ \sum_{k=0}^{k_2-2}\sum_{x=0}^{\infty}\pi_W(x+1,x,2,k)\lambda_2\rho^{-(x+2)} \\ &\quad \cdot H(x+2,x,2,k) \\ &+ \sum_{x=2}^{\infty}\pi_W(x,x,1,k_1-1)\lambda_2\rho^{-(x+1)} \\ &\quad \cdot H(x+1,x,1,k_1-1) \end{aligned} \tag{5}$$

where $H(x,y,i,k)$ is the probability that starting at (x,y,i,k), $\mathcal{W}^\infty$ never hits Δ.

4.2 Non-single-class dominant (Theorem 2)

Now we expect the event F_ℓ to be reached through the effects of both classes of customer. Here we define $W[n] = (\tilde{W}[n], \hat{W}[n])$ as $\tilde{W}$ to be the ordered pair $(Q_1[n]+Q_2[n], Q_1[n])$ and $\hat{W}[n]$ to simply be $(Z[n], I[n])$. The transition kernel for this chain is the same as that for the single-class dominant case.

The main difference in the analysis comes from the change in structure of W^∞. We now effectively remove the zero boundary for both queues and always serve k_1 and k_2 customers at each queue respectively and allow a negative number of customers at both queues. This translates to Δ being $(0,0,0,0)$, $(x+1,x,2,k)$ for $x \geq 0$ and$0 \leq k < k_2 - 1$, $(x,x,1,k_1-1)$ for $x > 1$, $(z+1,1,1,\ell)$, for $z \geq 0$ and $0 \leq \ell < k_1 - 1$, and $(y,0,2,k_2-1)$ where $y > 1$.

The harmonic function remains unchanged, $h(x) = \rho^{-\tilde{x}_1}$ and we construct the twisted free chain in exactly the same manner as for the single-class dominant case. The fact that the harmonic function has remained unchanged means that in checking technical Conditions 0-7, the only condition which is different is Condition 1. (On the surface, Condition 6 looks different, but the earlier argument for the single-class dominant situation is symmetric in the classes.

Condition 1. It is not difficult to see that $\varphi(1,k) = 1/(k_1+k_2)$ with $k \in \{0,\cdots,k_1-1\}$, and $\varphi(2,k) = 1/(k_1+k_2)$ with $k \in \{0,\cdots,k_2-1\}$.

Finally, we give the constant g_2 for this case:

$$\begin{aligned} g_2 &= g_1 - \pi_W(0,0,0,0)\lambda_1\rho^{-1}H(1,1,1,0) \\ &\quad \cdot(1-\mathbf{1}\{k_1=1\}) \\ &+ \sum_{\ell=0}^{k_1-2}\sum_{z=0}^{\infty}\pi_W(z+1,1,1,\ell)\lambda_1\rho^{-(z+2)} \\ &\quad \cdot H(z+2,2,1,\ell) \\ &+ \sum_{y=2}^{\infty}\pi_W(y,0,2,k_2-1)\lambda_1\rho^{-(y+1)} \\ &\quad \cdot H(y+1,1,2,k_2-1) \end{aligned} \tag{6}$$

References

[1] O.J. Boxma and H. Takagi (Eds.). Special issue on polling systems. *Queueing Systems*, 11, 1992.

[2] G.L. Choudhury and W. Whitt. Computing distributions and moments in polling models by numerical transform inversion. *Performance Evaluation*, 25:267–292, 1996.

[3] N.G. Duffield. Exponents for the tails of distributions in some polling models. *Queueing Systems*, 26:105–119, 1997.

[4] R.D. Foley and D. McDonald. Join the shortest queue: stability and exact asymptotics. *Annals of Applied Probability.* Submitted. An electronic version is available at http://www.isye.gatech.edu/faculty/Robert_Foley/pub.html.

[5] H. Levy and M. Sidi. Polling systems: Applications, modeling, and optimization. *IEEE Trans. Comm.*, 38:1750–1760, 1990.

[6] D. McDonald. Asymptotics of first passage times for random walk in a quadrant. *Annals of Applied Probability*, 9:110–145, 1999.