# Estimating Queue Length Distributions for Queues with Random Arrivals

Daniel S. Myers
University of Wisconsin-Madison
dsmyers@cs.wisc.edu

Mary K. Vernon
University of Wisconsin-Madison
vernon@cs.wisc.edu

## 1. ABSTRACT

This work develops an accurate and efficient two-moment approximation for the queue length distribution in the M/G/1 queue. Queue length distributions can provide insight into the impact of system design changes that go beyond simple averages, but conventional queueing theory lacks efficient techniques for estimating the long-run queue length distribution when service times are not exponential. The approximate queue lengths depend on only the first and second moments of the service time rather than the full service time distribution, resulting in a model that is applicable to a wide variety of systems. Validation results show that the new approximation is highly accurate for light-tailed service time distributions. Work in progress includes developing accurate approximations for multi-server queues and heavy-tailed service distributions.

## 2. INTRODUCTION

Modern computer systems can be viewed as highly complex collections of shared resources. Customized simulations and analytic models quantify the behavior of these systems and provide insight into system bottlenecks, leading to improved designs. These models typically provide accurate approximations in cases where the exact solution is intractable.

One limitation in the queueing theory that customized analytic models draw on, to date, is the lack of efficient, accurate calculations of the queue length distribution in the M/G/1 queue. Queue length distributions can give further insight into the impact of system design changes than mean queue length and mean residence time alone. For example, the queue length distributions at disk queues provides valuable design insight for large-scale storage systems.

Closed form expressions for the queue length distribution are available only for service time distributions for which the Pollaczek-Khinchin z-transform of the queue length distribution can be inverted to create such solutions, notably including systems with exponential or hyperexponential service time distributions [4]. In general, inversion of the transform is computationally infeasible [3].

This paper develops a simple and efficient two-moment approximation for the M/G/1 queue length distribution that has three key features. First, its derivation uses only Little's result, the random arrival property of the queue, and the well-known formula for mean residence time in the M/G/1 queue. Second, the new approximate queue length distribution depends only on the first and second moments of the service time distribution, which can easily be obtained for most real-world systems. Finally, the approximation is accurate for a wide range of light-tailed service time distributions with high and low coefficients of variation.



## 3. DERIVATION

Consider an M/G/1 queue with average service time $\overline{x}$, Poisson arrival rate $\lambda$, and a squared coefficient of variation of service time $c_x^2$. Number the positions of the queue, beginning with number one for the server, two for the first position in the waiting line, and so forth. Let $U_k$ be the utilization of position $k$, which is the fraction of time that position $k$ is occupied. Let $N$ denote the number of customers in the queue at a random point in time after the queue has reached equilibrium. Note that

$$P[N > k] = U_{k+1} \tag{1}$$

and

$$\begin{aligned} P[N = k] &= P[N > k-1] - P[N > k] \\ &= U_k - U_{k+1} \end{aligned} \tag{2}$$

For position one, $U_1$ is the server utilization, $U = \lambda\overline{x}$. More generally, for position $k+1 > 1$, we let $\overline{x}_{k+1}$ denote the expected time a customer spends in position $k+1$ and use Little's result applied to position $k+1$ to obtain

$$P[N > k] = U_{k+1} = \lambda\overline{x}_{k+1} \tag{3}$$

Given $\overline{x}_{k+1}$ for $k+1 > 1$, we can estimate $P[N > k]$ for $k > 1$.

To derive the average time at position $k+1 > 1$, consider an arbitrary "tagged customer" that arrives to the queue in equilibrium. Due to the PASTA property, the queue length probability distribution at the arrival instant is equal to the equilibrium queue length distribution at a random point in time. There are three cases to consider:

1. The tagged customer arrives when there are fewer than $k$ customers in the queue. In this case, the customer spends no time at position $k+1$ and $\overline{x}_{k+1} = 0$.
2. With probability $P[N = k]$, the tagged customer arrives when there are exactly $k$ customers in the queue, in which case the customer arrives directly to position $k+1$ and $\overline{x}_{k+1} = \overline{r}_{k+1}$, where $\overline{r}_{k+1}$ is the average remaining service time for the customer in service conditioned on the knowledge that there are $k$ customers in the queue at the arrival instant, discussed further below.

3. With probability $P[N > k]$, the tagged customer arrives when there are at least $k+1$ customers in the queue. As customers depart the queue, the tagged customer advances to position $k+1$, waits in position $k+1$ for a service time, and then moves to position $k$. In this case, $\overline{x}_{k+1} = \overline{x}$.

Combining the above cases yields the following:

$$\begin{aligned} P[N > k] &= \lambda \overline{x}_{k+1} \\ &= \lambda \left( P[N = k]\, \overline{r}_k + P[N > k]\, \overline{x} \right) \end{aligned} \tag{4}$$

Solving for $P[N > k]$ yields

$$P[N > k] = \frac{\lambda \overline{r}_k}{1 + \lambda \overline{r}_k - U} P[N > k-1] \tag{5}$$

As noted in [1], the position-dependent expected residual service times $\overline{r}_k$ are, in general, not equal to the unconditional mean residual life. In a service period with more than one arrival, the arrivals that occur later in the period will find a larger average queue length. Hence, for service time distributions with increasing failure rate, the expected residual service time is *negatively* correlated with arrival queue length. Similarly, for service time distributions with decreasing failure rate, residual service time is *positively* correlated with arrival queue length.

It is possible to use the full service time density function to calculate exact values for $\overline{r}_k$. To illustrate the dependence on $k$, Fig. 1 plots the values of $\overline{r}_k$ vs. $k$ in an M/$H_2$/1 queue with a two-stage hyperexponential service time distribution, $\overline{x} = 1.0$, $c_x^2 = 5$ and 70% utilization [4]. In this case, $\overline{r} = 3.0$ and $\overline{r}_k$ increases with $k$. Note that for $k > 1$, the correlation is relatively weak, and the individual residuals deviate from the mean by less than 10%. This is reasonable, as, after a queue build-up, an average of $\lambda \overline{x} < 1$ customers arrive during each service period. Some of these customers will observe a large queue size, but will be the first arrivals during their respective service intervals.

To obtain an efficient approximation, we assume that $\overline{r}_k \approx \overline{r}$, where $\overline{r}$ is the unconditional expected residual service time for the M/G/1 queue given by $\overline{r} = \frac{\overline{x}}{2}(1 + c_x^2)$. This approximation yields a geometric queue length distribution, derived below, which is consistent with a previous large-deviation theory result that the distribution of work in the G/G/1 queue is bounded by an exponential [6].

Substituting $\overline{r}$ for $\overline{r}_k$ in equation 5, we have

$$P[N > k] \approx \frac{\lambda \overline{r}}{1 + \lambda \overline{r} - U} P[N > k-1], \quad k = 1, 2, 3, \ldots \tag{6}$$

This equation provides a recursive calculation for $P[N > k]$. To remove the recursion, observe that the base case is simply $P[N > 0] = U$. Hence,

$$P[N > k] \approx U \left( \frac{\lambda \overline{r}}{1 + \lambda \overline{r} - U} \right)^k, \quad k = 0, 1, 2, \ldots. \tag{7}$$

To arrive at the final queue length probabilities, we substitute equation 7 in equation 2 and factor the common terms. Using the defintion of $\overline{r}$ and letting

$$q = \frac{\lambda \overline{r}}{1 + \lambda \overline{r} - U} = \frac{U(c_x^2 + 1)}{2 + U(c_x^2 - 1)}, \tag{8}$$

we obtain the following approximation:

$$P[N = k] \approx U q^{k-1} (1 - q), \quad k = 1, 2, 3, \ldots \tag{9}$$

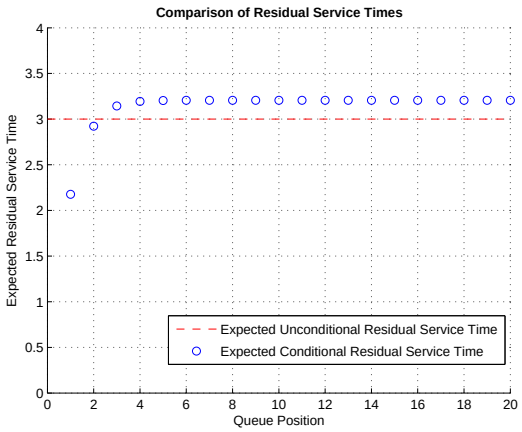



Figure 1: M/$H_2$/1 queue: $\overline{x} = 1.0$, $c_x^2 = 5.0$, $U = .70$

Note that the form of the approximation for queue size distribution yields insight into the input measures that are *principal determinants* of the queue size distribution. In particular, to the extent that the approximation is accurate, the queue length probabilities are principally determined by $U$ and $c_x$, rather than the specific values of $\lambda$ and $\overline{x}$ or higher moments of the service time distribution.

## 4. RELATED WORK

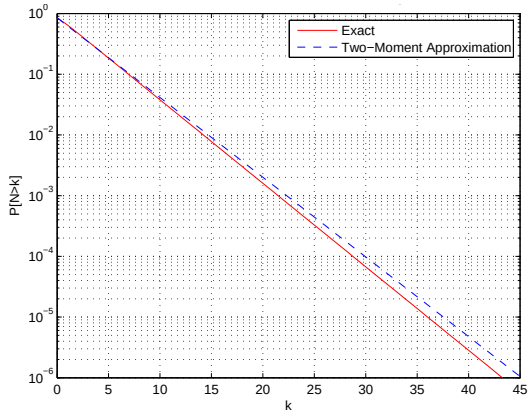
The technique of applying Little's result to individual queue positions has been used in [2] to derive the average residence time in the M/G/1 queue. Texts by Tijms [7] and Neuts [5] provide methods of computing queue length distributions for phase-type service times, but their methods are intractable in the context of large system models. In contrast, equations 8 and 9 are simple to calculate. Nelson [4] provides closed-form expressions for exact queue length distributions for deterministic, Erlang, and a special case of the two-stage hyperexponential distribution. Our two-moment approximation is simpler and more general, and provides more direct insight into the relationship between system parameters and the queue length probabilities.

## 5. VALIDATION

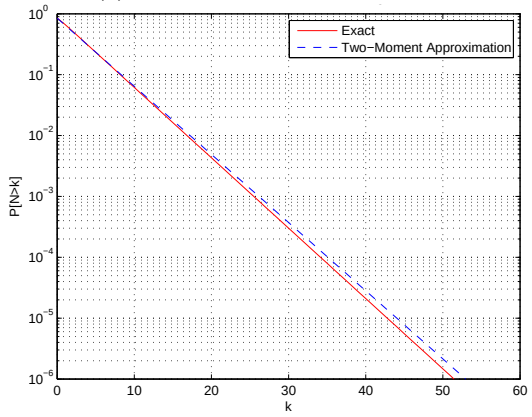
Equations 8 and 9 have been validated for queues with light-tailed service times. We have compared the two-moment approximation against exact values for deterministic, several Erlang, and several unrestricted two-stage hyperexponential distributions to investigate accuracy over a wide range of coefficients of variation. The exact queue length distributions were derived independently, and are significantly more complex than the two-moment approximation.

Figure 2 provides the results for three cases with $c_x^2 = 0$, .2, and 25. Results for other values of $c_x^2$ showed qualitatively similar accuracy. In each example, the average service time was fixed at $\overline{x} = 1$ and the arrival rate set to .85 achieve 85% utilization.

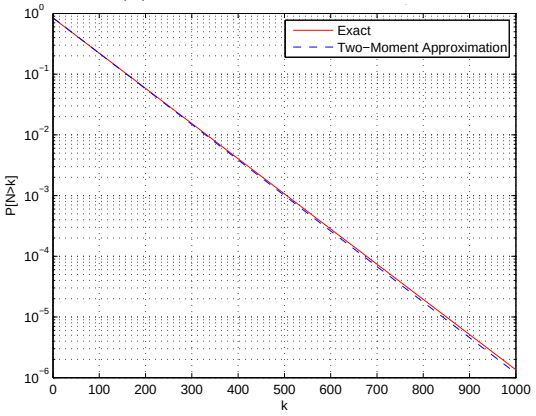
Fig. 2(a) plots the comparison for a queue with deterministic service times. The approximate queue length probabilities are sufficiently accurate to guide initial system design. Furthermore, the approximation overpredicts the probabilities at the tail of the queue length distribution, which is




(a) Deterministic, $U = .85$, $c_x^2 = 0$




(b) Erlang-5, $U = .85$, $c_x^2 = .2$




(c) Two-stage hyperexponential, $U = .85$, $c_x^2 = 25$

Figure 2: Two-moment approximation validation

generally preferable for system design decisions.

For example, consider selecting the buffer size for a given target value of $P[N > k]$, say $10^{-5}$, using the approximated probabilities vs. the exact probabilities. The exact results show that the probability of exceeding a buffer size of 36 is approximately .00001. Using the approximate probabilities, a length of 38 is sufficient to achieve the same level of service.

Fig. 2(b) plots the comparison for an Erlang-5 distribution. As in Fig. 2(a), the approximate probabilities are larger than the true values, but the approximations are again close enough to provide accurate buffer size estimations. Observations show that the two-moment approximation overpredicts the true queue length distribution when $c_x^2 < 1$.

Fig. 2(c) provides a comparison using the class of two-stage hyperexponential distributions described in [4]. The example has exponential service rates $\mu_1 = 13.9282$ and $\mu_2 = .0718$, with customers receiving service at rate $\mu_1$ with probability $p = .9282$. In contrast to Fig. 2(a) and 2(b), the approximate probabilities *underpredict* the true results in this case, although the degree of underprediction is small in this case. In general, system designers should bear in mind that the true queue length probability will be larger than the value predicted by the two-moment approximation when $c_x^2 > 1$.

As Fig. 2 shows, the two-moment approximation is accurate for light-tailed service time distributions. This includes many systems of practical interest, such as disk access times in storage systems. However, formula 9 has a modified geometric form, which gives less accurate estimates of the probability of large queue lengths (omitted due to space constraints) when the service times have a heavy-tailed distribution, such as the Pareto. Investigating accurate approximations for non-phase-type distributions is an element of our ongoing work.

## 6. CONCLUSION

We have derived a new and efficient method for computing queue length probability distributions for shared resources with random request arrivals and general service time distributions. The two-moment approximation is simple to implement and requires only the arrival rate and the first two moments of the service times. Validation results show that the approximation is accurate for light-tailed service time distributions, such as occur at resources with implicit or explicit service level objectives.

Current and future work includes developing queue-length distribution approximations for multiple server queues, investigation of more accurate appproximate queue length estimates for heavy-tailed service time distributions, and employing the new approximations in actual system design.

## 7. REFERENCES

[1] I. Adan and M. Haviv. Conditional ages and residual service times in the M/G/1 queue. *Stochastic Models*, 25(1), 2009.

[2] D. Fakinos. The expected remaining service time in a single server queue. *Operations Research*, 30(5), Sep.-Oct. 1982.

[3] L. Kleinrock. *Queueing Systems*, volume 1. Wiley, New York, 1975.

[4] Randolph Nelson. *Probability, Stochastic Processes, and Queueing Theory*. Springer-Verlag, New York, 1995.

[5] M.F. Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Johns Hopkins University Press, Baltimore, 1981.

[6] S. Ross. *Stochastic Processes*. John Wiley and Sons, New York, 1983.

[7] H.C. Tijms. *Stochastic Modelling and Analysis: A Computational Approach*. John Wiley and Sons, Great Britain, 1986.