A Note on the Effects of Service Time Distribution in the M/G/1 Queue

Alexandre Brandwajn[1] and Thomas Begin[2]

[1] Baskin School of Engineering, University of California Santa Cruz, USA
[2] Université Pierre et Marie Curie, LIP6, France
alexb@soe.ucsc.edu, thomas.begin@lip6.fr

Abstract. The M/G/1 queue is a classical model used to represent a large number of real-life computer and networking applications. In this note, we show that, for coefficients of variation of the service time in excess of one, higher-order properties of the service time distribution may have an important effect on the steady-state probability distribution for the number of customers in the M/G/1 queue. As a result, markedly different state probabilities can be observed even though the mean numbers of customers remain the same. This should be kept in mind when sizing buffers based on the mean number of customers in the queue. Influence of higher-order distributional properties can also be important in the M/G/1/K queue where it extends to the mean number of customers itself. Our results have potential implications for the design of benchmarks, as well as the interpretation of their results.

Keywords: performance evaluation, M/G/1 queue, higher-order effects, finite buffers.

1 Introduction

The M/G/1 queue is a classical model used to represent a large number of real-life computer and networking applications. For example, M/G/1 queues have been applied to evaluate the performance of devices such as volumes in a storage subsystem [1], Web servers [13], or nodes in an optical ring network [3]. In many applications related to networking, the service times may exhibit significant variability, and it may be important to account for the fact that the buffer space is finite. It is well known that, in the steady state, the mean number of users in the unrestricted M/G/1 queue depends only on the first two moments of the service time distribution [11]. It is also known [4] that the first three (respectively, the first four) moments of the service time distribution enter into the expression for the second (respectively, the third) moment of the waiting time. In this note our goal is to illustrate the effect of properties of the service time distribution beyond its mean and coefficient of variation on the shape of the stationary distribution of the number of customers in the M/G/1 queue. In particular, we point out the risk involved in dimensioning buffers based on the mean number of users in the system.

D. Kaeli and K. Sachs (Eds.): SPEC Benchmark Workshop 2009, LNCS 5419, pp. 138–144, 2009.

2 M/G/1 Queue

Assuming a Poisson arrival process, a quick approach to assess the required capacity for buffers in a system is to evaluate it as some multiplier (e.g. three or six) times the mean number of customers in an open M/G/1 queue (e.g. [12]). From the Pollaczek-Khintchine formula [11], this amounts to dimensioning the buffers based on only the first two moments of service time distribution. Unfortunately, the steady-state distribution of the number of customers in the M/G/1 queue can exhibit a strong dependence on higher-order properties of the service time distribution.

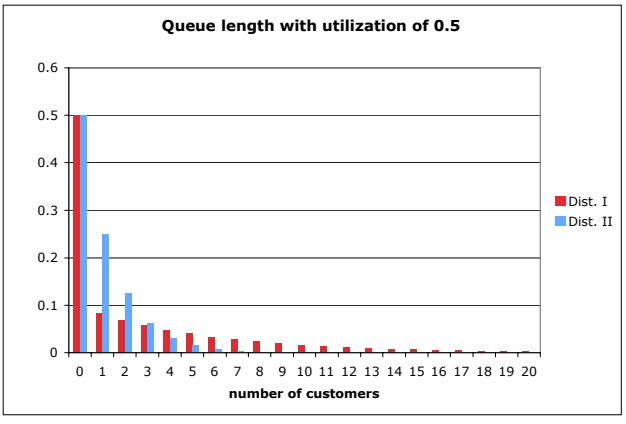
This is illustrated in Figure 1, which compares the distribution of the number of customers for two different Cox-2 service time distributions with the same first two moments, and thus yielding the same mean number of customers in the system. The parameters of these distributions are given in Table 1. Note that both distributions I and II correspond to a coefficient of variation of 3 but have different higher-order properties such as skewness and kurtosis [14]. Similarly, distributions III and IV both correspond to a coefficient of variation of 5 but again different higher-order properties. The stationary distribution of the number of customers in this M/G/1 queue was computed using a recently published recurrence method [2]. We observe that, perhaps not surprisingly, the effects of the distribution tend to be more significant as the server utilization and the coefficient of variation of the service time distribution increase. It is quite instructive to note, for instance, that with a coefficient of variation of 3 and server utilization of 0.5, the probability of exceeding 20 users in the queue (a little over 6 times the mean) is about 0.1% in one case while it is an order of magnitude larger for another service time distribution with same first two moments.

Table 1. Parameters and properties of the service time distributions used in Figure 1

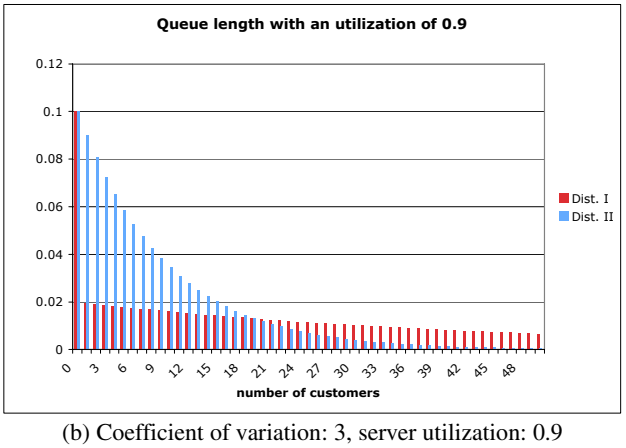
Distribution	Mean service time	Coefficient of variation	Skewness	Kurtosis	Rate of service at stage 1	Probability to go to stage 2	Rate of service at stage 2
Dist. I	1	3	4.5	27.3	10000.0	$2.00*10^{-1}$	$2.00*10^{-1}$
Dist. II	1	3	3557.4	$1.90*10^{7}$	1.0	$2.50*10^{-7}$	$2.50*10^{-4}$
Dist. III	1	5	7.5	75.1	10000.0	$7.69*10^{-2}$	$7.69*10^{-2}$
Dist. IV	1	5	6913.2	$6.63*10^{7}$	1.0	$8.33*10^{-8}$	$8.33*10^{-5}$

3 M/G/1/K Queue

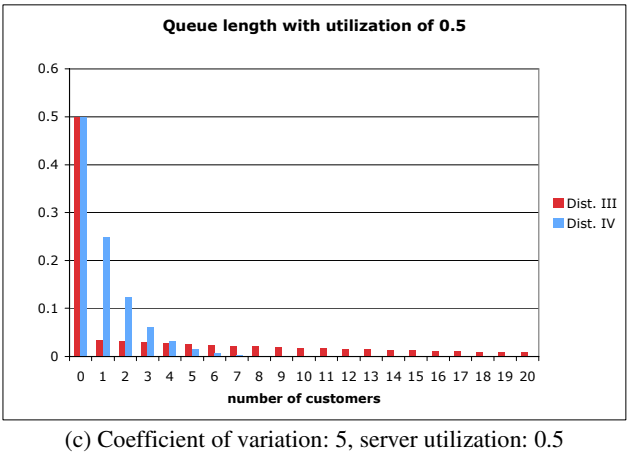
Clearly, using the M/G/1/K, i.e., the M/G/1 queue with a finite queueing room would be a more direct way to dimension buffers. There seem to be fewer theoretical results for the M/G/1/K queue than for the unrestricted M/G/1 queue, but it is well known that the steady-state distribution for the M/G/1/K queue can be obtained from that for the unrestricted M/G/1 queue after appropriate transformations [10, 7, 4]. Clearly, this approach can only work if the arrival rate does not exceed the service rate since otherwise the unrestricted M/G/1 would not be stable.



(a) Coefficient of variation: 3, server utilization: 0.5

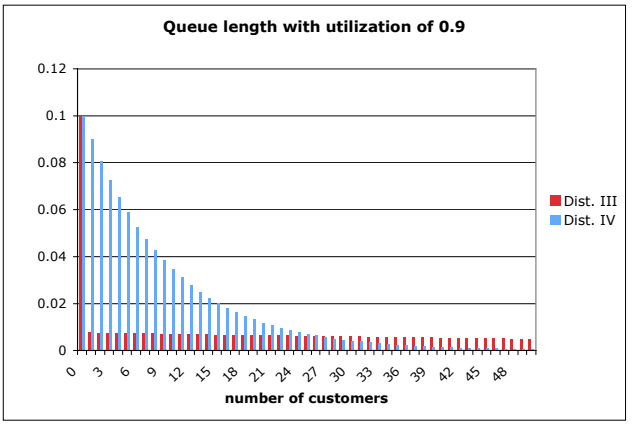


(b) Coefficient of variation: 3, server utilization: 0.9



(c) Coefficient of variation: 5, server utilization: 0.5

Fig. 1. Effect of service time distributions on the number of customers in the M/G/1 queue



(d) Coefficient of variation:0.5, server utilization: 0.9

Fig. 1. (*continued*)

While the steady-state distribution for the M/G/1/K queue can be derived from the one for the unrestricted M/G/1 queue, and the mean number of users in the latter depends only on the first two moments of the service time distribution, this is not the case for the M/G/1/K queue. Table 2 shows that even the first three moments of the service time distribution do not generally suffice to determine the mean number of customers in the M/G/1/K queue. Here we illustrate the results obtained for two Cox-3 distributions sharing the first three moments but with different properties of higher-order.

Since the mean number of customers in the unrestricted M/G/1 queue depends only on the first two moments of the service time distribution, and in the M/G/1/K for K=1 there is no distributional dependence at all (since there is no queueing), it is interesting to see how the dependence on properties of higher-order varies with K, the size of the queueing room. This is the objective in Figure 2 where we have represented the relative difference in the probabilities of having exactly one customer in the system, as well as in the probabilities of the buffer being full, for distribution I and II of Table 1. We observe that, although the first two moments of the service time distribution are the same for both distributions, higher-order properties lead to drastically different values for the selected probabilities. Interestingly, for the probability of the buffer being full, although the relative difference between the distributions considered decreases as the size of the queueing room, K, increases, it remains significant even for large values of the latter.

To further illustrate the dependence on higher-order properties of the service time distribution, we consider read performance for two simple cached storage devices. When the information requested is found in the cache, a hit occurs and the service time is viewed as a constant (assuming a fixed record length). When the information is not in the cache, it must be fetched from the underlying physical storage device. In Table 3 we show simulation results [8] obtained for two different storage systems with the same first two moments of the service time (resulting from the combination

Table 2. Effect of properties beyond the third moment on the mean number in the M/G/1/K queue

	First Cox-3	Second Cox-3	Relative differences
Rate of arrivals	1	1	
Size of queueing room	30	30	
Mean service time	1		
Coefficient of variation	6.40		
Skewness	2331.54		
Kurtosis	$7.43*10^6$	$1.44*10^7$	
Mean number in the M/G/1/K	3.98	5.07	27.4 %

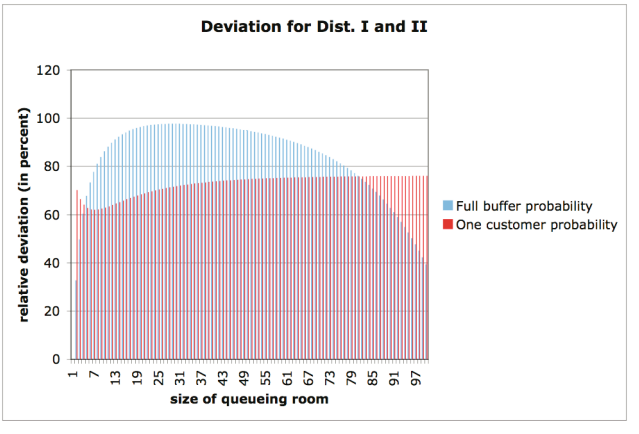


Fig. 2. Relative difference in selected probabilities for distributions I and II as a function of the queueing room in the M/G/1/K queue

of hit and miss service times), and queueing room limited to 10. In one case the service time of the underlying physical device (i.e. miss service time) is represented by a uniform distribution, and in the other by a truncated exponential [9]. We are interested in the achievable I/O rate such that the mean response time does not exceed 5 ms. We observe that the corresponding I/O rates differ by over 20% in this example (the coefficient of variation of the service time being a little over 1.6).

It has been our experience that the influence of higher-order properties tends to increase as the coefficient of variation and the skewness of the service time increase. It is interesting to note that this is precisely the case when one considers instruction execution times in programs running on modern processors where most frequent

Table 3. I/O rate for same mean I/O time in two storage subsystems

	Uniform miss service time	Truncated exponential miss service time	Relative differences
Mean service time	1.9		
Coefficient of variation	1.62		
Hit probability	0.9	0.985	
Hit service time	1	1.64	
Miss service time	Uniform [2,18]	Truncated exponential mean: 20, max: 100	
Attainable I/O rate for Mean I/O time of 5 ms	0.257	0.312	21.4 %

instructions are highly optimized, less frequent instructions can be significantly slower, and certain even less frequent instructions may be implemented as subroutine calls with order of magnitude longer execution times.

As another example of the effects of higher-order properties of the service time in an M/G/1 queue, consider the probability that a small buffer of 10 messages at an optical network node is full. Incoming packets can be of three different lengths. In the fist case, abstracted from reported IP traffic, the packet lengths are 40, 300 and 1500 bytes with probabilities 0.5, 0.3 and 0.2, respectively. In the second case, longer packets are used: 150, 500 and 5000 bytes, with respective probabilities 0.426, 0.561 and 0.013. Both packet length distributions have the same mean of 410 bytes with a coefficient of variation of 1.36, but different higher order properties. With the average packet arrival rate at 1 per mean packet service time, simulation results indicate that the probability of the buffer being full differs by some 20% depending on the packet mix (12.5% in the first case vs. 10.5% in the second) even though both packet mixes have the same fist two moments.

4 Conclusion

In conclusion, we have shown that, for coefficients of variation of the service time in excess of one, higher-order properties of the service time distribution may have an important effect on the steady-state probability distribution for the number of customers in the M/G/1 queue. As a result, markedly different state probabilities can be observed even though the mean numbers of customers remain the same. This should be kept in mind when sizing buffers based on the mean number of customers in the queue. Influence of higher-order distributional properties can also be important in the M/G/1/K queue where it extends to the mean number of customers itself. The potentially significant impact of higher-order distributional properties of the service times should be kept in mind also when interpreting benchmark results for systems that may

be viewed as instances of the M/G/1 or M/G/1/K queue, in particular, transaction oriented systems. Our results imply that it may not be sufficient to look just at the mean or even the mean and the variance of the system execution times to correctly assess the overall system performance. Another implication relates to benchmark design since, unless one is dealing with a system that satisfies the assumptions of a product-form queueing network, it may not be sufficient to simply preserve the mean of the system load [6].

Acknowledgments. The authors wish to thank colleagues for their constructive remarks on an earlier version of this note.

References

1. Brandwajn, A.: Models of DASD Subsystems with Multiple Access Paths: A Throughput-Driven Approach. IEEE Transactions on Computers C-32(5), 451–463 (1983)
2. Brandwajn, A., Wang, H.: Conditional Probability Approach to M/G/1-like Queues. Performance Evaluation 65(5), 366–381 (2008)
3. Bouabdallah, N., Beylot, A.-L., Dotaro, E., Pujolle, G.: Resolving the Fairness Issues in Bus-Based Optical Access Networks. IEEE Journal on Selected Areas in Communications 23(8), 1444–1457 (2005)
4. Cohen, J.W.: On Regenerative Processes in Queueing Theory. Lecture Notes in Economics and Mathematical Systems. Springer, Berlin (1976)
5. Cohen, J.W.: The Single Server Queue, 2nd edn. North-Holland, Amsterdam (1982)
6. Ferrari, D.: On the foundations of artificial workload design. SIGMETRICS Perform. Eval. Rev. 12(3), 8–14 (1984)
7. Glasserman, P., Gong, W.: Time-changing and truncating K-capacity queues from one K to another. Journal of Applied Probability 28(3), 647–655 (1991)
8. Gross, D., Juttijudata, M.: Sensitivity of Output Performance Measures to Input Distributions in Queueing Simulation Modeling. In: Proceedings of the 1997 winter simulation conference, pp. 296–302 (1997)
9. Jawitz, J.W.: Moments of truncated continuous univariate distributions. Advances in Water Resources 27(3), 269–281 (2004)
10. Keilson, J.: The Ergodic Queue Length Distribution for Queueing Systems with Finite Capacity. Journal of the Royal Statistical Society 28(1), 190–201 (1966)
11. Kleinrock, L.: Queueing systems. Theory, vol. I. Wiley, Chichester (1974)
12. Mitrou, N.M., Kavidopoulos, K.: Traffic engineering using a class of M/G/1 models. Journal of Network and Computer Applications 21, 239–271 (1998)
13. Molina, M., Castelli, P., Foddis, G.: Web traffic modeling exploiting. TCP connections' temporal clustering through HTML-REDUCE 14(3), 46–55 (2000)
14. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman & Hall / CRC, Boca Raton (1986)