M/M/c Queues and the Poisson Clumping Heuristic

STEVEN FINCH

April 8, 2019

Abstract. In continuous time, customers arrive at random. Each waits until one of c servers is available; each thereafter departs at random. The distribution of maximum line length of idle customers was studied over 25 years ago. We revisit two good approximations of this, employing a discrete Gumbel formulation and detailed graphics to describe simulation outcomes.

Consider an M/M/c queue with arrival rate λ and service rate µ. Let \( M_n \) denote the maximum queue length over the time interval \([0, n]\). For integer \( k \), we could study \( \mathbb{P}\{M_n < k\} \) asymptotically as a function of \( n \), as was done in [1] for the case \( c = 1 \). We prefer, however, to suppress the dependence on \( n \) somewhat, separating (in essence) signal from noise. Let \( k = \log_{c\mu/\lambda}(n) + h + 1 \) as defined in [2]. The Poisson clumping heuristic asserts that, if \( \lambda < c\mu \), then

\[
\mathbb{P}\{M_n \leq \log_{c\mu/\lambda}(n) + h\} \sim \exp\left[ -\frac{c^{-2} \lambda \mu^{c^{-3}}(c\mu - \lambda)^2}{\sum_{j=1}^{c} j!(c^{-1})\lambda^{c^{-j}}\mu^{j-1}} \left( \frac{\lambda}{c\mu} \right)^{h+1} \right]
\]

as \( n \to \infty \). The finite sum involving binomial coefficients is explained in Section 1. In particular,

\[
\mathbb{P}\{M_n \leq \log_{\mu/\lambda}(n) + h\} \sim \exp\left[ -\frac{\lambda(\mu - \lambda)^2}{\mu^2} \left( \frac{\lambda}{\mu} \right)^{h+1} \right]
\]

for \( c = 1 \),

\[
\mathbb{P}\{M_n \leq \log_{2\mu/\lambda}(n) + h\} \sim \exp\left[ -\frac{\lambda(2\mu - \lambda)^2}{\mu(2\mu + \lambda)} \left( \frac{\lambda}{2\mu} \right)^{h+1} \right]
\]

for \( c = 2 \),

\[
\mathbb{P}\{M_n \leq \log_{3\mu/\lambda}(n) + h\} \sim \exp\left[ -\frac{3\lambda(3\mu - \lambda)^2}{6\mu^2 + 4\lambda\mu + \lambda^2} \left( \frac{\lambda}{3\mu} \right)^{h+1} \right]
\]
for $c = 3$,
\[
\mathbb{P} \{ M_n \leq \log_{4\mu/\lambda}(n) + h \} \sim \exp \left[ -\frac{16\lambda\mu(4\mu - \lambda)^2}{24\mu^3 + 18\lambda\mu^2 + 6\lambda^2\mu + \lambda^3} \left( \frac{\lambda}{4\mu} \right)^{h+1} \right]
\]
for $c = 4$ and
\[
\mathbb{P} \{ M_n \leq \log_{5\mu/\lambda}(n) + h \} \sim \exp \left[ -\frac{125\lambda\mu^2(5\mu - \lambda)^2}{120\mu^4 + 96\lambda\mu^3 + 36\lambda^2\mu^2 + 8\lambda^3\mu + \lambda^4} \left( \frac{\lambda}{5\mu} \right)^{h+1} \right]
\]
for $c = 5$. Also,
\[
\mathbb{E}(M_n) \approx \frac{\ln(n)}{\ln(\frac{\lambda}{\mu})} + \frac{\gamma + \ln \left( \frac{\lambda^2(\mu - \lambda)^2}{\mu^3} \right)}{\ln(\frac{\lambda}{\mu})} + \frac{1}{2}
\]
\[
\approx (2.4663034623\ldots \ln(n) - (7.2049448811\ldots)
\]
for $(c, \lambda, \mu) = (1, 1/3, 1/2)$,
\[
\mathbb{E}(M_n) \approx \frac{\ln(n)}{\ln(\frac{2\mu}{\lambda})} + \frac{\gamma + \ln \left( \frac{\lambda^2(2\mu - \lambda)^2}{2\mu^2(2\mu + \lambda)} \right)}{\ln(\frac{2\mu}{\lambda})} + \frac{1}{2}
\]
\[
\approx (2.4663034623\ldots \ln(n) - (6.7552845943\ldots)
\]
for $(c, \lambda, \mu) = (2, 1/3, 1/4)$,
\[
\mathbb{E}(M_n) \approx \frac{\ln(n)}{\ln(\frac{3\mu}{\lambda})} + \frac{\gamma + \ln \left( \frac{\lambda^2(3\mu - \lambda)^2}{\mu(6\mu^2 + 4\lambda\mu + \lambda^2)} \right)}{\ln(\frac{3\mu}{\lambda})} + \frac{1}{2}
\]
\[
\approx (2.4663034623\ldots \ln(n) - (6.2049448811\ldots)
\]
for $(c, \lambda, \mu) = (3, 1/3, 1/6)$,
\[
\mathbb{E}(M_n) \approx \frac{\ln(n)}{\ln(\frac{4\mu}{\lambda})} + \frac{\gamma + \ln \left( \frac{4\lambda^2(4\mu - \lambda)^2}{24\mu^3 + 18\lambda\mu^2 + 6\lambda^2\mu + \lambda^3} \right)}{\ln(\frac{4\mu}{\lambda})} + \frac{1}{2}
\]
\[
\approx (2.4663034623\ldots \ln(n) - (5.6015876099\ldots)
\]
for $(c, \lambda, \mu) = (4, 1/3, 1/8)$ and
\[
\approx (2.4663034623\ldots \ln(n) - (4.9642624490\ldots)
\]
for \((c, \lambda, \mu) = (5, 1/3, 1/10)\), where \(\gamma\) denotes Euler’s constant \([3]\). With regard to expected maximums, in a hospital emergency room \((\lambda = 1/3)\), one fast doctor \((\mu = 1/2)\) outperforms \(c\) very slow doctors \((\mu = 1/(2c))\).

A higher-order approximation for the probability is \([4, 5, 6]\)

\[
P \{ M_n \leq \log_{c\mu/\lambda}(n) + h \} \sim \exp \left[ -\frac{n \frac{c^{-2}}{c^{-1} + \lambda + c^{-3}(c\mu - \lambda)^2}}{\{ n \lambda c^{-1} (c\mu)^{h+1} - \lambda^{h+1} (c\mu)^{c-1} \} \sum_{j=1}^{c} j! \left( \frac{c-1}{c} \right) \lambda^{c-j} \mu^j - 1} \right]
\]

with the same relation between (real) \(h\) and (integer) \(k\) as earlier. The dependence of the probability on \(n\) is more visible here; allowing \(n \to \infty\) within the square brackets yields exactly the same expression as before.

In greater detail, Serfozo \([4]\) examined the distribution of \(M_\nu\), the maximum queue length over \(\ell\) busy cycles, where \(\nu\) is the \(\ell\)th time the system becomes empty. Very important corrections to Serfozo’s second table appeared in \([6]\); observe the existence of exact probabilistic results in this special case. McCormick & Park \([5]\) examined the distribution of \(M_n\) for arbitrary \(n\), following \([4]\). A missing coefficient \(c! / c^c\) in McCormick & Park’s formula (2.22) was uncovered in \([6]\); no exact results are generally available here.

We simulated \(10^6\) M/M/c queues for each of the following choices of \((n, c, \lambda, \mu)\):

- \(n = 1000\) or \(2500\)
- \(c = 1, 2, 3, 4\) or \(5\)
- \(\lambda = 1/3\)
- \(\mu = 1/(2c)\)

and indicate both low-order approximation (red) and high-order approximation (green) histograms against empirical outcomes (blue). The green segments are always closer to the blue segments than the red segments. Also, the discrepancies become smaller as \(n\) grows larger. We used clumping heuristic-based estimates for the mean (from earlier) and likewise

\[
\mathbb{V}(M_n) \approx \frac{\pi^2}{6} \frac{1}{\ln \left( \frac{c\mu}{\lambda} \right)^2} + \frac{1}{12}
\]

for the variance. The fit between moments is surprisingly good; no attempt was made to employ a more sophisticated underlying formula.
M/M/c Queues and the Poisson Clumping Heuristic

M/M/1 queue maxima: n=1000, lambda=1/3, mu=1/2
Experimental mean is 9.971; mean square is 108.393
Theoretical mean is 9.832; mean square is 106.751

M/M/1 queue maxima: n=2500, lambda=1/3, mu=1/2
Experimental mean is 12.155; mean square is 157.275
Theoretical mean is 12.052; mean square is 156.294
M/M/c Queues and the Poisson Clumping Heuristic

M/M/2 queue maximums: n=1000, lambda=1/3, mu=1/4
Experimental mean is 10.416; mean square is 117.474
Theoretical mean is 10.281; mean square is 115.795

M/M/2 queue maximums: n=2500, lambda=1/3, mu=1/4
Experimental mean is 12.610; mean square is 168.591
Theoretical mean is 12.541; mean square is 167.970
M/M/c Queues and the Poisson Clumping Heuristic

M/M/3 queue maximums: n=1000, lambda=1/3, mu=1/6
Experimental mean is 10.963; mean square is 129.171
Theoretical mean is 10.832; mean square is 127.414

M/M/3 queue maximums: n=2500, lambda=1/3, mu=1/6
Experimental mean is 13.158; mean square is 182.692
Theoretical mean is 13.092; mean square is 191.477
M/M/c Queues and the Poisson Clumping Heuristic

M/M/4 queue maximums: n=1000, lambda=1/3, mu=1/8
Experimental mean is 11.555; mean square is 142.516
Theoretical mean is 11.435; mean square is 140.849

M/M/4 queue maximums: n=2500, lambda=1/3, mu=1/8
Experimental mean is 13.761; mean square is 198.879
Theoretical mean is 13.695; mean square is 197.639
M/M/c Queues and the Poisson Clumping Heuristic

M/M/5 queue maximums: n=1000, lambda=1/3, mu=1/10
Experimental mean is 12.185; mean square is 157.506
Theoretical mean is 12.072; mean square is 155.831

M/M/5 queue maximums: n=2500, lambda=1/3, mu=1/10
Experimental mean is 14.393; mean square is 216.700
Theoretical mean is 14.332; mean square is 215.501
1. **Erlang C**

Letting $\pi$ denote the stationary distribution of $\text{M/M/}c$, the probability that all $c$ servers are busy is \[ \sum_{k=c}^{\infty} \pi_k = \frac{c^c}{c!} \left( \sum_{k=c}^{\infty} \rho^k \right) \pi_0 = \frac{(cp)^c}{c!(1-\rho)} \pi_0 \]

where $\rho = \lambda/(c\mu)$ and

\[ \frac{1}{\pi_0} = \sum_{j=0}^{c-1} \frac{(cp)^j}{j!} + \frac{(cp)^c}{c!(1-\rho)} \]

We wish to demonstrate that

\[ \frac{1}{\pi_0} = \frac{\sum_{i=1}^{c} i! (\frac{c-1}{i-1}) \lambda^{c-i} \mu^{i-1}}{(c-1)! \mu^{c-2}(c\mu-\lambda)} \]

As a preliminary step, note that

\[ \frac{(cp)^c}{c!(1-\rho)} = \frac{\rho(c\rho)^{c-1}}{(c-1)!(1-\rho)} = \frac{-(1-\rho)(c\rho)^{c-1} + (cp)^{c-1}}{(c-1)!(1-\rho)} \]

\[ = -\frac{(c\rho)^{c-1}}{(c-1)!} + \frac{(cp)^{c-1}}{(c-1)! (1-\rho)} \]

The new $(1/\pi_0)$-formula is equal to

\[ \frac{1}{c\mu} \sum_{i=1}^{c} \frac{i \lambda^{c-i} \mu^{i-1}}{(c-i)! 1-\lambda/(c\mu)} = \frac{1}{c} \sum_{i=1}^{c} \frac{i (cp)^{c-i}}{(c-i)! 1-\rho} \]

As index $i$ runs from 1 to $c$, index $j = c - i$ runs from $c - 1$ to 0, giving

\[ \frac{1}{c} \sum_{c-j=0}^{c-1} \frac{c-j}{j!} \frac{(cp)^j}{1-\rho} = \frac{1}{c} \sum_{j=0}^{c-1} \frac{(cp)^j}{j!(1-\rho)} - \rho \sum_{j=1}^{c-1} \frac{(cp)^{j-1}}{(j-1)!(1-\rho)} \]

which telescopes to

\[ \sum_{j=0}^{c-2} \frac{(cp)^j}{j!} + \frac{(cp)^{c-1}}{(c-1)!(1-\rho)} = \sum_{j=0}^{c-1} \frac{(cp)^j}{j!} - \frac{(cp)^{c-1}}{(c-1)!(1-\rho)} + \frac{(cp)^{c-1}}{(c-1)!(1-\rho)} \]

By the preliminary step, this collapses to the old $(1/\pi_0)$-formula and we are done.
2. **Erlang B**

In an M/M/c/c queue, if a customer arrives when all c servers are busy, the customer leaves the system immediately (with no effect on the queue). The probability that all c servers are busy is \[ \pi_c = \frac{(cp)^c}{c!} \pi_0 \]

where \( \rho = \lambda/(c\mu) \) and

\[
\frac{1}{\pi_0} = \sum_{j=0}^{c} \frac{(cp)^j}{j!}.
\]

3. **Erlang A**

In an M/M/c+M queue, customers arrive with patience times \( \tau \) that are independent, exponentially distributed with mean \( 1/\theta \). The abandonment rate \( \theta \) is 0 for Erlang C and is \( \infty \) for Erlang B. If no service is offered before time \( \tau \) has elapsed, the customer leaves the system immediately. Define

\[
E = \frac{(cp)^c}{c!} \frac{1}{\sum_{j=0}^{c} \frac{(cp)^j}{j!}},
\]

the ratio from Section 2; and for \( x > 0, \ y \geq 0 \),

\[
A(x, y) = \frac{x \exp(y)}{y^x} \int_0^y t^{x-1} \exp(-t) dt = 1 + \sum_{k=1}^{\infty} \frac{y^k}{k! (x + \ell)} 
\]

an incomplete gamma function. The probability that all c servers are busy is \[ \sum_{k=c}^{\infty} \pi_k = \frac{(cp)^c}{c!} A \left( \frac{c\mu}{\theta}, \frac{\lambda}{\theta} \right) \pi_0 \]

and

\[
\frac{1}{\pi_0} = \frac{(cp)^c}{c!} \left[ \frac{1}{E} + A \left( \frac{c\mu}{\theta}, \frac{\lambda}{\theta} \right) - 1 \right].
\]

We wonder about the implications of work in [11], especially a result involving the constants \( \gamma \) and \( \pi^2/6 \). Might certain issues we’ve neglected here concerning asymptotic moments (\( h \) is real, not integer) be resolvable?
4. Acknowledgements

I am thankful to Guy Louchard for introducing me to the Poisson clumping heuristic and to Stephan Wagner for extracting discrete Gumbel asymptotics in [12] (a contribution leading to [2, 13] and the present work). Writing simulation code for M/M/c was facilitated by a theorem in [14] involving order statistics of iid Uniform rvs. The creators of R and Mathematica, as well as administrators of the MIT Supercloud Cluster, earn my gratitude every day.

References

[1] D. Aldous, Probability Approximations via the Poisson Clumping Heuristic, Springer-Verlag, 1989, pp. 1–8, 23–25, 30; MR0969362.

[2] S. Finch, Geo/Geo/2 queues and the Poisson clumping heuristic, arXiv:1902.09272.

[3] S. R. Finch, Euler-Mascheroni constant, Mathematical Constants, Cambridge Univ. Press, 2003, pp. 28–40; MR2003519.

[4] R. F. Serfozo, Extreme values of birth and death processes and queues, Stochastic Process. Appl. 27 (1988) 291–306; MR0931033.

[5] W. P. McCormick and Y. S. Park, Approximating the distribution of the maximum queue length for M/M/s queues, Queueing and Related Models, ed. U. Narayan Bhat and I. V. Basawa, Oxford Univ. Press, 1992, pp. 240–261; MR1210568.

[6] G. Hooghiemstra, L. E. Meester and J. Hüsler, On the extremal index for the M/M/s queue, Comm. Statist. Stochastic Models 14 (1998) 611–621; MR1621358.

[7] R. B. Cooper, Introduction to Queueing Theory, 2nd ed., North-Holland, 1981, pp. 79–101, 176–178; MR0636094.

[8] C. Palm, Research on telephone traffic carried by full availability groups, Tele 1 (1957) 1–107 [Engl. transl. of five papers published in Swedish in 1946].

[9] S. Zeltyn, Call Centers with Impatient Customers: Exact Analysis and Many-Server Asymptotics of the M/M/n+G Queue, Ph.D. thesis, Israel Institute of Technology, 2004; http://ie.technion.ac.il/~serveng/course2004/References/MMNGthesis.pdf
[10] L. Rozenshmidt, *On Priority Queues with Impatient Customers: Stationary and Time-Varying Analysis*, M.Sc. thesis, Israel Institute of Technology, 2007; [http://ie.technion.ac.il/~serveng/course2004/References/thesis_Luba_Eng.pdf](http://ie.technion.ac.il/~serveng/course2004/References/thesis_Luba_Eng.pdf).

[11] G. Pang and W. Whitt, Heavy-traffic extreme value limits for Erlang delay models, *Queueing Syst.* 63 (2009) 13–32; MR2576005; [http://www.columbia.edu/~ww2040/PangWhittExtremesErlang.pdf](http://www.columbia.edu/~ww2040/PangWhittExtremesErlang.pdf).

[12] S. Finch, The maximum of an asymmetric simple random walk with reflection, [arXiv:1808.01830](https://arxiv.org/abs/1808.01830).

[13] S. Finch and G. Louchard, Traffic light queues and the Poisson clumping heuristic, [arXiv:1810.12058](https://arxiv.org/abs/1810.12058).

[14] K. Sigman, Notes on the Poisson process, lecture notes (2009), [http://www.columbia.edu/~ks20/stochastic-I/stochastic-I-PP.pdf](http://www.columbia.edu/~ks20/stochastic-I/stochastic-I-PP.pdf).

Steven Finch  
MIT Sloan School of Management  
Cambridge, MA, USA  
steven_finch@harvard.edu