The RESET Technique for Multiserver-Job Analysis

Isaac Grosof

1 Introduction

Multiserver queueing theory emphasizes one-server-per-job models, such as the M/G/k. Such models were popular for decades in the study of computing systems. However, one-server-per-job models no longer reflect the behavior of many modern computing systems.

Modern datacenters, such as those of Google, Amazon, and Microsoft, no longer resemble traditional one-server-per-job models such as the M/G/k. Each job now may request many servers, which the job holds simultaneously. For instance, in Google's recently published trace of its "Borg" computation cluster [7, 13], the number of CPUs requested per job varied by a factor of 100,000. We focus on the corresponding theoretical model, the "multiserver-job model" (MSJ), and specifically the first-come first-served (FCFS) service ordering, a natural policy that production systems often use by default. [4, 12].

Currently, little is known about FCFS service in MSJ models. Some papers characterize the stability region of such systems, under restrictive assumptions on the job duration distributions [1, 11, 10, 9]. A key technique for characterizing the stability region of these systems is the *saturated system* approach [6, 2]. However, almost nothing is known about mean response time $E[T]$ in FCFS MSJ systems. The only FCFS MSJ system in which mean response time has been analytically characterized is the ultra-simple case of 2 servers and exponentially distributed durations [3, 5]. Mean response time is much better understood under more complex scheduling policies such as ServerFilling and ServerFilling-SRPT [7, 8], but these policies require both preemption and additional assumptions.

We invent a novel technique to characterize mean response time in the FCFS MSJ model, which we call "Reduction to Saturated for Expected Time" (RESET). The saturated system is a closed MSJ system in which completions trigger new arrivals, ensuring that there are always k jobs in the system, where k is the number of servers. The saturated system is a finite-state Markov chain, making it far easier to characterize. Reductions to the saturated system have long been employed in the study of MSJ models, but only for studying the *stability region* of MSJ models [6, 2]. We reduce to the same saturated system, but use it to characterize the mean response time of the original system for the first time.

In Theorem 3.1, we employ the RESET technique to prove the first characterization of mean response time $E[T]$ in the FCFS MSJ model. To do so, we invent the concept of "relative completions", which measures the extent to which a given state is expected to have more or less completions in the near future, relative to another state. Our characterization is tight in the heavy traffic limit, as the arrival rate λ approaches λ^*, the boundary of stability.

2 Relative Completions

Our key new concept is that of *relative completions*. Let $C_1(t)$ denote the number of completions up to time t of a saturated system initialized in state y_1 at time $t = 0$, and let $C_2(t)$ be defined similarly for an independent saturated system initialized in state y_2. Then let $\Delta(y_1, y_2)$ denote the relative completions between states y_1 and y_2.

$$\Delta(y_1, y_2) = \lim_{t\to\infty} E[C_1(t) - C_2(t)]$$

We also allow y_1 and/or y_2 to be distributions over states, rather than single states.

3 Result

Let Y^{Sat} be the time-average steady state of the saturated system, and let Y_d^{Sat} be the departure-average steady state, sampled just after each completion.

THEOREM 3.1. *In a multiserver-job system with i.i.d. arrivals and Poisson(λ) arrival process, the expected response time satisfies*

$$E[T^{MSJ}] = \frac{1}{\lambda^*}\frac{1 + \Delta(Y_d^{Sat}, Y^{Sat})}{1 - \frac{\lambda}{\lambda^*}} + O_\lambda(1)$$

To prove Theorem 3.1, we make use of the "Always-k" system, an intermediate system that bridges the gap between the MSJ system and the saturated system. The Always-k system mirrors the MSJ system, except that whenever a completion occurs when there are exactly k jobs in the system, an extra i.i.d. arrival is inserted. The k oldest jobs in the Always-k system form a sub-Markov chain which exactly matches the behavior of the saturated system. We prove that the Always-k system has the same mean response time behavior as the MSJ system, up to $O_\lambda(1)$ terms.

We employ an instantaneous generator approach to characterize the mean response time of the Always-k system, and a novel busy period analysis to prove that the Always-k and MSJ systems have identical asymptotic behavior.

4 Acknowledgements

This abstract is based on research performed with Yige Hong, Mor Harchol-Balter, and Alan Scheller-Wolf.

5 References

[1] L. Afanaseva, E. Bashtova, and S. Grishunina. Stability analysis of a multi-server model with simultaneous service and a regenerative input flow. *Methodology and Computing in Applied Probability*, pages 1–17, 2019.

[2] F. Baccelli and S. Foss. On the saturation rule for the stability of queues. *Journal of Applied Probability*, 32(2):494–507, 1995.

[3] P. H. Brill and L. Green. Queues in which customers receive simultaneous service from a random number of servers: A system point approach. *Management Science*, 30(1):51–68, 1984.

[4] Y. Etsion and D. Tsafrir. A short survey of commercial cluster batch schedulers. *School of Computer Science and Engineering, The Hebrew University of Jerusalem*, 44221:2005–13, 2005.

[5] D. Filippopoulos and H. Karatza. An M/M/2 parallel system model with pure space sharing among rigid jobs. *Mathematical and Computer Modelling*, 45(5):491 – 530, 2007.

[6] S. Foss and T. Konstantopoulos. An overview of some stochastic stability methods. *Journal of the Operations Research Society of Japan*, 47(4):275–303, 2004.

[7] I. Grosof, M. Harchol-Balter, and A. Scheller-Wolf. WCFS: A new framework for analyzing multiserver systems. *Queueing Systems*, 2022.

[8] I. Grosof, Z. Scully, M. Harchol-Balter, and A. Scheller-Wolf. Optimal scheduling in the multiserver-job model under heavy traffic. *Proc. ACM Meas. Anal. Comput. Syst.*, 6(3), dec 2022.

[9] E. Morozov and A. Rumyantsev. Stability analysis of a MAP/M/s cluster model by matrix-analytic method. In D. Fiems, M. Paolieri, and A. N. Platis, editors, *Computer Performance Engineering*, pages 63–76, Cham, 2016. Springer International Publishing.

[10] A. Rumyantsev, R. Basmadjian, S. Astafiev, and A. Golovin. Three-level modeling of a speed-scaling supercomputer. *Annals of Operations Research*, pages 1–29, 2022.

[11] A. Rumyantsev and E. Morozov. Stability criterion of a multiserver model with simultaneous service. *Annals of Operations Research*, 252(1):29–39, 2017.

[12] L. Sliwko. A taxonomy of schedulers–operating systems, clusters and big data frameworks. *Global Journal of Computer Science and Technology*, 2019.

[13] M. Tirmazi, A. Barker, N. Deng, M. E. Haque, Z. G. Qin, S. Hand, M. Harchol-Balter, and J. Wilkes. Borg: The next generation. In *Proceedings of the Fifteenth European Conference on Computer Systems*, EuroSys '20, New York, NY, USA, 2020. Association for Computing Machinery.