Asymptotic bounds for $M^X/G/1$ processor sharing queues*

Hanhua Feng and Vishal Misra
Department of Computer Science
Columbia University

July 1, 2003

Abstract

This paper analyzed the asymptotic bounds of an $M/G/1$ processor sharing queue with bulk arrivals.

1 Introduction

We consider an $M/G/1$ processor sharing queueing system with bulk arrivals. This kind of system is usually denoted by $M^X/G/1/PS$. Due to analytical difficulties, little discussions have been paid to this kind queueing systems in the literature.

This system has a Poisson arrival process, but jobs arrive in groups. We denote by random variable N the number of jobs that arrive simultaneously. The sizes of jobs, denoted by random variable X, are independent of each other, even within a same group, and have an identical distribution. For simplicity we define $G(x) \triangleq 1 - F(x)$, where $F(x)$ is the CDF of the job sizes.

Since the scheduling descipline is processor sharing, at any time point, all jobs in the system equally share the processor. Therefore, $T'(x)$, the derivative of the expected sojourn time $T(x)$ for a tagged job of size x, is as same as the expected number of jobs in the system when this tagged job has already been served for size x. In this way, $T'(x)$ was described by Kleirock et al. in [3] (also reported in page 184 of [5]), with an integral equation

$$T'(x) \quad = \quad \lambda\bar{a}\int_0^\infty T'(y)G(x+y)dy + \lambda\bar{a}\int_0^x T'(y)G(x-y)dy + 1 + bG(x) \qquad (1)$$

where $\bar{a} = E[N]$ is the expected number of jobs in a group, and $b = E[N^2]/E[N]-1$ is the expected number of jobs arriving with a tagged job (not including the tagged job itself). The right-hand side of (1) shows the total number of different kinds of jobs in the system when the tagged job is being served at size x. The first item indicates the expected number of jobs that arrive before the tagged job did, the second item indicates the expected number of jobs that arrive after the tagged job did, the third item indicates the tagged job itself, and the last item indicates the number of jobs that arrive within a same group of the tagged job.

This equation cannot be easily solved. Kleinrock et al. [3] solved this equation with a very special class of job size distributions; in a recent paper, Bansal [4] developed methods to solve this equation for generalized hyperexponential and Coxian [6] job size distributions.

*This technical report is part of an unpublished document titled "An analysis of scheduling disciplines for network flows." (Sections V,VI, and VII)

The general solution of $T(x)$ is still unknown; all previous methods deal with special job size distributions, and involve sub-problems that need to solve linear equations. The properties of $T(x)$ of such systems are therefore not as clear as some other queueing systems that have plain analytic solutions. In this paper, by analyzing this integral equation, we give some boundary properties of an $M^X/G/1/PS$ system.

2 Asymptotic bounds

Clearly, the load ρ of this $M^X/G/1$ is $\lambda\bar{a}E[X]$, and it must be less than 1 if the system is stable. Let us introduce an $M/G/1$ processor sharing system with non-bulk arrival rate $\lambda\bar{a}$. We denote by $\tilde{T}(x)$ the expected sojourn time of a job of size x in such system. These two systems have the same load, and it is well-known that

$$\tilde{T}(x) = \frac{x}{1-\rho} = \frac{x}{1-\lambda\bar{a}E[X]}.$$

We are interested in the difference of sojourn times between the two systems, i.e., $D(x) \triangleq T(x) - \tilde{T}(x)$. Replacing $T(x)$ by $D(x)$ in (1), we get

$$\begin{aligned}
&D'(x) + \frac{1}{1-\rho} \\
&= \lambda\bar{a}\int_0^\infty \left[D'(y) + \frac{1}{1-\rho}\right] G(x+y)dy + \lambda\bar{a}\int_0^x \left[D'(y) + \frac{1}{1-\rho}\right] G(x-y)dy + 1 + bG(x) \\
&= \lambda\bar{a}\int_0^\infty D'(y)G(x+y)dy + \lambda\bar{a}\int_0^x D'(y)G(x-y)dy + \frac{\lambda\bar{a}}{1-\rho}\int_0^\infty G(y)dy + 1 + bG(x)
\end{aligned}$$

Note that $\int_0^\infty [1-F(x)]dx = E[X]$, therefore

$$\frac{\lambda\bar{a}}{1-\rho}\int_0^\infty G(y)dy = \frac{\lambda\bar{a}E[X]}{1-\rho} = \frac{\rho}{1-\rho},$$

then we get

$$D'(x) = \lambda\bar{a}\int_0^\infty D'(y)G(x+y)dy + \lambda\bar{a}\int_0^x D'(y)G(x-y)dy + bG(x). \tag{2}$$

Therefore,

$$\begin{aligned}
D'(x) &\leq \lambda\bar{a}\int_0^\infty \sup_{y\geq 0}\{D'(y)\}G(x+y)dy + \lambda\bar{a}\int_0^x \sup_{y\geq 0}\{D'(y)\}G(x-y)dy + bG(x) \\
&= \lambda\bar{a}\sup_{y\geq 0}\{D'(y)\}\int_x^\infty G(x)dx + \lambda\bar{a}\sup_{y\geq 0}\{D'(y)\}\int_0^x G(x)dx + bG(x) \\
&= \rho\sup_{y\geq 0}\{D'(y)\} + bG(x).
\end{aligned}$$

Since $G(y) \leq 1$, we have

$$\sup_{x\geq 0}\{D'(x)\} \leq \rho\sup_{x\geq 0}\{D'(x)\} + b,$$

then

$$\sup_{x\geq 0}\{D'(x)\} \leq \frac{b}{1-\rho},$$

which leads to an upper bound of $D'(x)$:

$$D'(x) \leq \frac{b}{1-\rho}. \tag{3}$$

Similarly,

$$D'(x) \geq \rho \inf_{y\geq 0}\{D'(y)\} + bG(x),$$

then we can get

$$\inf_{x\geq 0}\{D'(x)\} \geq \rho \inf_{x\geq 0}\{D'(x)\},$$

which leads to a lower bound of $D'(x)$:

$$D'(x) \geq 0 \tag{4}$$

Since $T(0) = \tilde{T}(0) = 0$, this lower bound means that the expected sojourn time of a job in an $M^X/G/1/PS$ system is always greater than that in an $M/G/1/PS$ system with the same load; in other words, for any job size distribution, systems with bulk arrivals always have bad performace: longer response time and more jobs in the system. Moreover, $D'(x) \geq 0$ also means that the difference of sojourn times between these two systems would monotonically increase as the job size increases; however, the difference is bounded, as the later analysis shows.

The upper bound in (3) indicates upper bounds of the difference of expected response time of a job of size x between bulk and non-bulk systems

$$D(x) \leq x \sup\{D'(x)\} \leq \frac{bx}{1-\rho}, \tag{5}$$

the expected response time of the bulk system for a job of size x

$$T(x) = D(x) + \tilde{T}(x) \leq \frac{(b+1)x}{1-\rho}, \tag{6}$$

and the expectation of sojourn time of any job

$$E[T] \leq \frac{b+1}{1-\rho}E[X].$$

This upper bound is actually tight for an $M/D/1$ system. In an $M/D/1$ system that all groups have the same number of jobs and all jobs have the same size, by equivalently combining all jobs in a group into a single job, we can easily see that $T(x) = \bar{a}x/(1-\rho) = (b+1)x/(1-\rho)$.

Since $D'(x) \geq 0$, we can get other upper bounds of $D(x)$ and $T(x)$ by estimating their value as $x \to \infty$.

Integrating both sides of (2), we get

$$
\begin{aligned}
D(x) = & \int_0^x D'(t)dt \\
= & \lambda\bar{a}\int_0^x \int_0^\infty D'(y)G(t+y)dydt + \lambda\bar{a}\int_0^x\int_0^t D'(y)G(t-y)dydt + b\int_0^x G(t)dt \\
= & \lambda\bar{a}\int_0^x \left[[D(y)G(t+y)]|_{y=0}^{y=\infty} - \int_{y=0}^\infty D(y)\frac{\partial G(t+y)}{\partial y}dy \right] dt \quad \boxed{\text{integration by parts}} \\
& +\lambda\bar{a}\int_0^x \left[[D(y)G(t-y)]|_{y=0}^{y=t} - \int_{y=0}^t D(y)\frac{\partial G(t-y)}{\partial y}dy \right] dt + b\int_0^x G(t)dt \\
= & -\lambda\bar{a}\int_0^x \int_{y=0}^\infty D(y)\frac{\partial G(t+y)}{\partial y}dy\; dt \quad \boxed{G(0)=1, G(\infty)=0, D(0)=0} \\
& -\lambda\bar{a}\int_0^x D(t)dt - \int_0^x \int_{y=0}^t D(y)\frac{\partial G(t-y)}{\partial y}dydt + b\int_0^x G(t)dt \\
= & -\lambda\bar{a}\int_0^\infty D(y)\int_{t=0}^x \frac{\partial G(t+y)}{\partial t}dt\; dy \quad \boxed{\text{Fubini's theorem; } \partial y \to \partial t} \\
& +\lambda\bar{a}\int_0^x D(t)dt - \lambda\bar{a}\int_0^x D(y)\int_{t=y}^x \left[-\frac{\partial G(t-y)}{\partial t} \right] dt\; dy + b\int_0^x G(t)dt \\
= & \lambda\bar{a}\int_0^\infty D(y)[G(y) - G(x+y)]\; dy + \lambda\bar{a}\int_0^x D(y)G(x-y)\; dy + b\int_0^x G(y)dy
\end{aligned}
$$

Therefore, letting $x \to \infty$, we get

$$
\begin{aligned}
& D(\infty) \\
= & \lambda\bar{a}\int_0^\infty D(y)G(y)dy + \lambda\bar{a}\lim_{x\to\infty}\int_0^x D(y)G(x-y)dx + b\int_0^\infty G(y)dy \\
\leq & \lambda\bar{a}\int_0^\infty \frac{by}{1-\rho}G(y)dy + \lambda\bar{a}D(\infty)E[X] + bE[X]
\end{aligned}
$$

Note that $\int_0^\infty yG(y)dy = \frac{1}{2}E[X^2]$ and $\rho = \lambda\bar{a}E[X]$, we have

$$
D(\infty) \leq \frac{b\rho}{(1-\rho)^2}\left(\frac{E[X^2]}{2E[X]}\right) + \frac{b}{1-\rho}E[X]
$$

Since $D(x)$ is monotonically increasing, if X has a finite second-order moment, then we have a constant upper bound

$$
D(x) \leq \frac{b\rho}{(1-\rho)^2}\left(\frac{E[X^2]}{2E[X]}\right) + \frac{b}{1-\rho}E[X] \tag{7}
$$

3 Bounds of $T(x)/x$ for an $M^X/G/1/PS$ queue

Sometimes we are interested in the ratio of $T(x)/x$. This ratio gives the average time to serve a unit size for a job of size x. Equation (7) tells us that for large jobs,

$$
\lim_{x\to\infty}\frac{T(x)}{x} = \frac{\tilde{T}(x)}{x} = \frac{1}{1-\rho} \tag{8}
$$

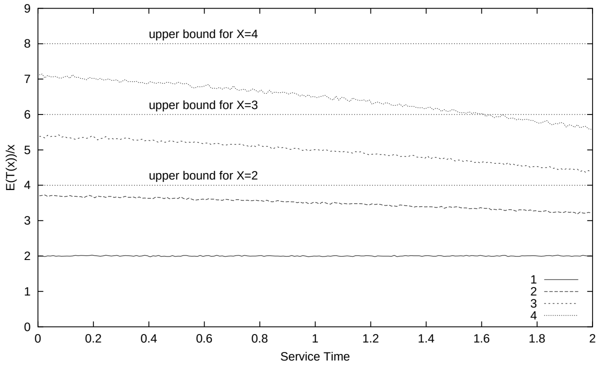

Figure 1: the sojourn time as a function of job sizes for uniformly distributed job sizes with $\rho = 0.5$, for bulk sizes being 1, 2, 3, and 4.

which means that large jobs would not experience much service slowdown with bulk arrivals. What about the small jobs? From (2) we can easily solve $D'(0)$,

$$\begin{aligned} D'(0) &= \lambda\bar{a}\int_0^\infty D'(y)G(y)dy + bG(0) \\ &= \lambda a\int_0^\infty D(y)dF(y) + b \\ &\geq b \end{aligned}$$

We can get an upper bound from (6). Because $T(x)/x \approx T'(0) = D'(0) + 1/(1-\rho)$ for small jobs, now we have both upper and lower bounds for small jobs

$$\frac{1}{1-\rho} + b \leq \lim_{x\to 0}\frac{T(x)}{x} \leq \frac{b+1}{1-\rho}. \tag{9}$$

These bounds do not depend on the job size distribution.

Simulation results show that if the job size is more deterministic, the sojourn time of small jobs is very close to the upper bound, as shown in Figure 1, in which case the job sizes are uniformly distributed between 0 and 2, with $\rho = 0.5$.

If the load ρ is small, these two bounds are quite close. If the load is close to 1, the upper bound is quite large. The simulation results show that, for heavy-tailed distributions, the sojourn time is quite close to the lower bound. Figure 2 shows the sojourn time of bounded-Pareto job size distribution with $\rho = 0.95$, for various bulk sizes.

References

[1] L. Kleinrock and R.R. Muntz, *Processor-sharing queueing models of mixed scheduling disciplines for time-shared systems*, Journal of the ACM, 19, 464-482, 1972

[2] L. Kleinrock, R.R. Muntz and J. Hsu, *Tight bounds on average response time for processor-sharing models of time-shared computer systems*, Information Processing 71, TA-2, 50-58, August 1971.

[3] L. Kleinrock, R. Muntz, and E. Rodemich, *The processor-sharing queueing model for time-shared systems with bulk arrivals*, Networks, 1, 1–13, 1971.

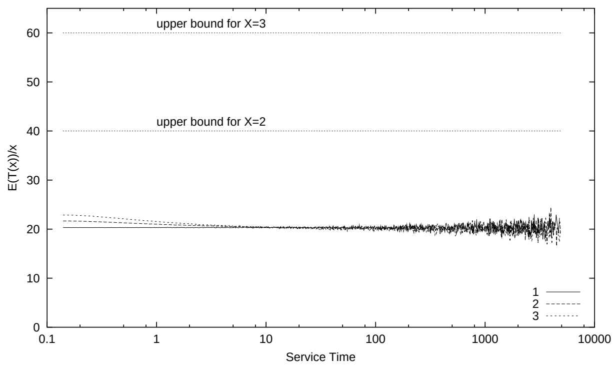

Figure 2: the sojourn time as a function of job sizes for bounded-Pareto distributed job sizes with $\rho = 0.95$, for bulk sizes being 1, 2, and 3.

[4] N. Bansal, *Analysis of the $M/G/1$ processor-sharing queue with bulk arrivals, Operations Research Letters*, In press.

[5] L. Kleinrock, *Queueing systems, volumn II: computer applications*, John Wiley & Sons, 1976.

[6] D. R. Cox, *A use of complex probabilities in the theory of stochastic processes*, Proceedings of Cambridge Philosophical Society, pages 313–319, 1955.