Finding the optimal quantum size: Sensitivity analysis of the *M/G/1* round-robin queue

Varun Gupta
Carnegie Mellon University
varun@cs.cmu.edu

ABSTRACT

We consider the round robin (RR) scheduling policy where the server processes each job in its buffer for at most a fixed quantum, q, in a round-robin fashion. The processor sharing (PS) policy is an idealization of the quantum-based round-robin scheduling in the limit where the quantum size becomes infinitesimal, and has been the subject of many papers. It is well known that the mean response time in an $M/G/1/PS$ queue depends on the job size distribution via only its mean. However, almost no explicit results are available for the round-robin policy. For example, how does the variability of job sizes affect the mean response time in an $M/G/1/RR$ queue? How does one choose the optimal quantum size in the presence of switching overheads? In this paper we present some preliminary answers to these fundamental questions.

1. INTRODUCTION

We consider a single server $M/G/1/RR$ queueing system. The jobs arrive according to a Poisson process with rate λ, and their sizes are assumed to be independent and identically distributed according to a random variable S with mean $\mathbf{E}[S] = \frac{1}{\mu}$. The server picks a job from the head of the queue and processes it for at most a time quantum of q units. If the job finishes service within this quantum it leaves the system, otherwise it rejoins at the end of the queue and waits till every job in front of it has received a quantum. Let $\rho = \frac{\lambda}{\mu}$ denote the load of this system. Let C^2 denote the squared coefficient of variation (SCV) of the job size distribution: $C^2 = \frac{\mathbf{E}\left[S^2\right]}{\mathbf{E}[S]^2} - 1$.

In the limit where the service quantum q approaches 0, the scheduling discipline becomes the idealized processor sharing (PS) policy. It is well known that the mean response time of an $M/G/1/PS$ system is given by

$$\mathbf{E}\left[T^{PS}\right] = \frac{1}{\mu - \lambda}$$

and is independent of any characteristic of the job size distribution beyond its mean. When $q \to \infty$, the round-robin system resembles a First-Come-First-Served queue for which the mean response time is given by

$$\mathbf{E}\left[T^{FCFS}\right] = \frac{1}{\mu - \lambda}\left(1 + \rho\frac{C^2 - 1}{2}\right)$$

In the presence of variable job sizes (high C^2), PS is desirable over FCFS, and hence one wants as small a quantum size as possible. However, in a real system one must pay some switching cost whenever the server finishes processing one quantum of a job and starts processing a different job. For example, in an operating system, at every preemption, the kernel data structures managing the run queues have to be modified. Also, switching to a job involves waiting for the cache to be filled with the relevant data and instructions.

Naturally, there is a tradeoff involved in choosing the optimal quantum size. A very small quantum size increases the load of the system due to overheads, whereas a big quantum exposes the effects of job size variability. The goal of this paper is to address the following question: *What is the optimal quantum size?* To be able to answer the question of optimal quantum size, we must first consider the question, *how does the sensitivity of the mean response time to the job size variability vary as one increases the quantum size?* While $M/G/1/RR$ has received some attention in the literature, no simple answers are yet available to these fundamental questions.

Outline

In Section 2, we look at an approximation of the $M/G/1/RR$ system to obtain insights into the interplay of variability and quantum size on the mean response time. In Section 3 we briefly outline the analysis of $M/G/1/RR$ and present simple and essentially tight bounds on the mean response time. For simplicity, we restrict ourselves to job size distributions with support on integral multiples of the quantum size q. Based on conjectured bounds in Section 3, we propose an expression for the optimal quantum size in Section 4. Finally, we conclude in Section 5.

2. EFFECT OF JOB SIZE VARIABILITY IN *M/G/1/RR* - AN APPROXIMATION

To gain intuition into the effect of job size variability (as represented by C^2) in an $M/G/1/RR$ queue, we begin by looking at an approximation of the $M/G/1/RR$. We make the following two approximations which allow modeling the system as a Markov chain:

1. We approximate the job size distribution by a degenerate hyperexponential distribution, H_2^*, with mean $1/\mu$ and SCV C^2. The desired distribution is given by [1]

$$H_2^* \sim \begin{cases} 0 & w.p.\ \frac{C^2-1}{C^2+1} \\ \text{Exp}\left(\frac{2\mu}{C^2+1}\right) & w.p.\ \frac{2}{C^2+1} \end{cases}$$

[1] $\text{Exp}(\mu)$ denotes an exponential random variable with mean $1/\mu$.

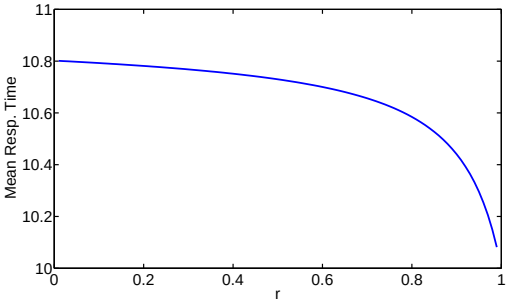

Figure 1: Illustration of the effect of varying the H_2 distribution on the mean response time in an $M/H_2/1/RR$ system with exponential quanta. The parameters of the job size distribution held constant were $\mu = 1$ and $C^2 = 19$ with load $\rho = 0.9$ and $\nu = 10$. The X-axis denotes the load made up the branch with the smaller mean of the two branches of the H_2 distribution.

2. The quantum size is picked in an i.i.d. fashion from an exponential distribution with mean $\frac{1}{\nu}$.

It is an easy exercise to verify that the mean response time for the system under the above assumptions is given by:

$$\mathbf{E}\left[T^{RR^*}\right] = \mathbf{E}\left[T^{PS}\right]\left[1 + \frac{C^2-1}{C^2+1} \cdot \frac{\lambda}{\nu + \frac{2}{C^2+1}\mu}\right] \qquad (1)$$

We make the following observations:

1. For a fixed ν, the mean response time monotonically increases from $\frac{1}{\mu-\lambda}$ to $\frac{1}{\mu-\lambda}\left(1+\frac{\lambda}{\nu}\right)$ as C^2 increases from 1 to ∞. Further, the mean response time asymptotes to the upper limit for relatively low values of C^2.
2. For a fixed C^2, the mean response time increases monotonically from $\mathbf{E}\left[T^{PS}\right]$ to $\mathbf{E}\left[T^{FCFS}\right]$ as the mean quantum size $1/\nu$ increases.

Our numerical results indicate that among all two-phase hyperexponential distributions (denoted by H_2 and defined as mixture of two independent exponential distributions) with a given mean and SCV, the H_2^* distribution yields the maximum response time. Figure 1 shows the mean response time in an $M/H_2/1/RR$ with exponential quanta while fixing $\lambda = 0.9$, $\mu = 1$, $\nu = 10$ and $C^2 = 19$ and varying the remaining degree of freedom of the H_2 distribution. The X-axis denotes r, the load made up by the branch with the smaller mean of the two branches of H_2 ($r = 0$ represents the H_2^* distribution). We will use this intuition to conjecture bounds on the mean response time in an $M/G/1/RR$ in Section 3. The expression in (1) is also important because it is a very simple and accurate approximation to the conjectured upper bound (Conjecture 1).

3. *M/G/1/RR* ANALYSIS AND BOUNDS

Let q denote the quantum size. To keep the analysis clean we restrict our attention to distributions with support on integral multiples of q, in particular, on $\{0, q, 2q \cdots, Kq\}$ for some positive integer K. Let p_i be the probability mass on iq and define

$$P_{\geq i} = \sum_{j=i}^{K} p_j$$

Let D_i be the expected delay experienced by a job while waiting to receive its ith quantum (given its size is at least iq). That is, the time spent in queue from the time a job finishes its $(i-1)$st quantum to the time it is served next. Under our assumptions, the analysis in [3] simplifies so that D_i satisfy the following linear system of equations:

$$D_1 = \lambda q\left[\left(\sum_{j=1}^{K} D_j P_{\geq j}\right)\right] + q\frac{\rho}{2}$$

and for $2 \leq i \leq K$,

$$D_i = \lambda q\left[\left(\sum_{j=1}^{K-i+1} D_j P_{\geq j+i-1}\right) + \left(\sum_{j=1}^{i-1} D_j P_{\geq i-j}\right)\right] + q\rho$$

The mean delay (time in queue) is then given by

$$\mathbf{E}[T_Q] = p_0 D_1 + \sum_{i=1}^{K} D_i P_{\geq i}$$

Theorem 1 gives bounds on D_i and the mean response time under the above assumptions.

THEOREM 1. *Let the job size distribution have support $\{0, q, 2q, \cdots, Kq\}$. Then,*

$$\frac{\rho(1+\rho)}{2(1-\rho)}\left[\frac{q}{1+\lambda q}\right] \leq D_1 \leq \frac{\rho(1+\rho)}{2(1-\rho)}q$$

$$\frac{\rho}{(1-\rho)}\left[\frac{q}{1+\lambda q}\right] \leq D_i \leq \frac{\rho}{1-\rho}q \qquad \ldots i \geq 2$$

This gives the following bounds on the mean response time:

$$\left[\frac{1-\rho}{2} + \frac{1}{2}\cdot\frac{1+\rho}{1+\lambda q}\right] \leq \frac{\mathbf{E}[T(K,q)]}{\mathbf{E}\left[T^{PS}\right]} \leq \left[1 + \frac{(1+\rho)\lambda q}{2} - \frac{\rho(1+\rho)}{2K}\right]$$

PROOF. We briefly outline the proof here. The solution for $\mathbf{D} = [D_1 \ \cdots \ D_K]$ can be seen as solving the fixed point of a monotone linear system of equations:

$$\mathbf{D} = \mathbf{D}A_P + b = f_P(\mathbf{D})$$

where we have subscripted the function with P to indicate the dependence on the job size distribution. Let $\mathbf{D}' = [D_1' \ \cdots \ D_K']$ and $\mathbf{D}^* = [D_1^* \ \cdots \ D_K^*]$ where D_i' and D_i^* denote the lower and upper bounds, respectively, on D_i mentioned in the theorem statement. It is straightforward to verify that,

$$f_P(\mathbf{D}^*) \preceq \mathbf{D}^* \quad , \quad f_P(\mathbf{D}') \succeq \mathbf{D}'.$$

Since f_P is a monotone linear function, it follows

$$\mathbf{D}' \preceq \mathbf{D} \preceq \mathbf{D}^*$$

where we use $\preceq$ and $\succeq$ to imply componentwise ordering. Bounds on mean response time follow by observing

$$0 \leq p_0 \leq 1 - \frac{1}{\mu K q}.$$

□

The lower bound is tight due to Proposition 1, and the upper bound is tight within a factor of $(1 + \rho/K)$ due to Proposition 2.

PROPOSITION 1. *Let $1/\mu = iq$ for some $i \in \{1, \cdots, K\}$. For the deterministic distribution with mean $1/\mu$, the mean response time is given by:*

$$\mathbf{E}[T_1(K,q)] = \mathbf{E}\left[T^{PS}\right]\left[\frac{1-\rho}{2} + \frac{1}{2}\cdot\frac{1+\rho}{1+\lambda q}\right].$$

PROPOSITION 2. *For the distribution with support* only *on 0 and Kq, the mean response time is given by:*

$$\mathbf{E}[T_2(K,q)] = \frac{\mathbf{E}\left[T^{PS}\right]}{(1+\rho/K)}\left[1 + \frac{(1+\rho)\lambda q}{2} - \frac{\rho^2}{K}\right].$$

Note that the upper bound in Theorem 1 is increasing in K. Taking the limit $K \to \infty$, we obtain the following upper bound on the mean response time in an $M/G/1/RR$ queue:

$$\mathbf{E}\left[\overline{T}(q)\right] = \frac{1}{\mu-\lambda}\left[1 + \frac{(1+\rho)\lambda q}{2}\right] \qquad (2)$$

We will use the expression in (2) to obtain the optimal quantum size in Section 4.

Based on observations made in Section 2, and many numerical experiments, we also conjecture the following stronger statement:

CONJECTURE 1. *Among the job size distributions with a given mean $1/\mu$ and SCV C^2, the mean response time in an $M/G/1/RR$ queue is maximized by the distribution with support* only *on 0 and $(C^2+1)/\mu$.*

Note that we do not restrict the distribution to have support on integral multiples of q in Conjecture 1. Besides displaying many extremal properties, the distribution mentioned in Conjecture 1 has been shown to maximize the mean response time in a $GI/M/1$ queue given the first two moments of the interarrival time distribution [2], and conjectured to maximize the mean response time in an $M/G/K$ queue given the first two moments of the job size distribution [1].

4. CHOOSING THE OPTIMAL QUANTUM

In this section, we will address the question of choosing the optimal quantum size to balance the tradeoff between excessive overhead and high variability of the job size distribution. Based on Theorem 1, we have the following upper bound on the mean response time in an $M/G/1/RR$ queue when there are no switching overheads:

$$\mathbf{E}\left[\overline{T}(q)\right] = \frac{1}{\mu-\lambda}\left[1 + \frac{(1+\rho)\lambda q}{2}\right]$$

Let there be an overhead h at the end of every quantum. To incorporate the overhead into the mean response time, we make the following changes:

$$q \to q+h \quad , \quad \mu \to \frac{\mu}{1+\frac{h}{q}}$$

We still use $\rho = \lambda/\mu$ to denote the system load under no switching overhead. Let $\beta = \frac{h}{h+q}$, and

$$\beta^* = \arg\min_{\beta} \frac{1}{(1-\beta)\mu - \lambda}\left[1 + \frac{1}{2}\left(1 + \frac{\lambda}{(1-\beta)\mu}\right)\frac{\lambda h}{\beta}\right]$$

The optimal β is approximated by:

$$\beta^* \approx \frac{\rho(1-\rho)}{1+\sqrt{1+\frac{2}{1+\rho}\frac{1-\rho}{\mu h}}}$$

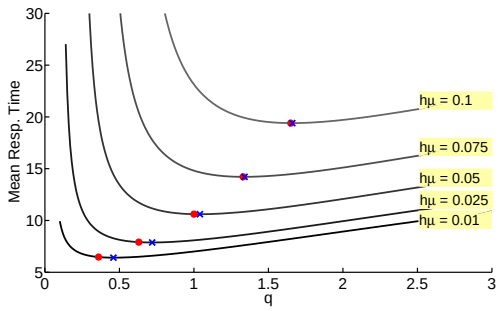

Figure 2: Comparison of q^* against the numerically obtained optimal quantum size for varying values of $h\mu$ from 1% to 10%. The job size distribution was chosen to be H_2 with mean 1, $C^2 = 19$ and $r = 0.5$ (balanced means) with load $\rho = 0.8$. The red circles denote our approximation q^* for the optimal quantum size and the blue crosses denote the numerically obtained optimal quantum size.

For the common case: $\frac{1-\rho}{h\mu} \gg 1$

$$q^* \approx \left(\frac{1}{\rho\sqrt{\frac{1-\rho^2}{2}}}\right)\sqrt{\frac{h}{\mu}}$$

Above, we approximated the optimal quantum size by minimizing the worst-case mean response time given the system load. The remarkable accuracy of this approximation is illustrated in Figure 2, where we compare our approximation for the optimal quantum size to the numerically obtained optimum for an H_2 job size distribution with $C^2 = 19$ and $r = 0.5$.

5. CONCLUSIONS

In this paper, we addressed the important question of how does one choose a good quantum size in a quantum-based round-robin system to balance the tradeoff between switching overhead and the effect of job size variability. This required us to first perform a sensitivity analysis to understand the effect of job size variability in a quantum-based round-robin system. We presented an approximate sensitivity analysis and tight bounds on the effect of variability in an $M/G/1/RR$ system. Based on these results, we provide simple and accurate expressions for choosing a good quantum size. It will be interesting to extend the results in this paper to other variants of processor sharing. For example, for a practical implementation of multi-level processor sharing (MLPS) our expression for q^* indicates that one might want to choose a different quantum size for each level.

6. REFERENCES

[1] Varun Gupta, Jim Dai, Mor Harchol-Balter, and Bert Zwart. The effect of higher moments of job size distribution on the performance of an $M/G/K$ queueing system. Technical Report CMU-CS-08-106, School of Computer Science, Carnegie Mellon University, 2008.

[2] Ward Whitt. On approximations for queues, I: Extremal distributions. *AT&T Bell Laboratories Technical Journal*, 63:115–138, 1984.

[3] Ronald W. Wolff. Time sharing with priorities. *SIAM J. Appl. Math.*, 19(3):566–574, 1970.