The Relationship between Bernoulli and Fixed Feedback Policies for the M/G/1 Queue

Vikram S. Adve
Computer Science Department
University of Wisconsin-Madison
Madison, WI 53706

Randolph Nelson
IBM Research Division
T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Abstract

We consider an M/G/1 queue with feedback, in which customers, after receiving service, either return to the tail of the queue or depart the system, according to some feedback policy. We derive simple expressions for the expected response time for feedback policies that include Bernoulli feedback and feeding back a fixed number of times. Our results reveal some interesting and non-intuitive properties of the behavior of such feedback policies when the coefficient of variation of service time is varied. One result shows that for the Bernoulli feedback and fixed feedback policies with equal mean number of visits to the queue, the expected response time for the Bernoulli policy is smaller than for the fixed policy if the coefficient of variation of service time is greater than 1. The relationship reverses if the coefficient of variation is less than 1.

We consider a queueing model consisting of a single server queue having infinite capacity. We assume that customer arrivals come from a Poisson point source with intensity λ. Customer service times at the queues are independent and identically distributed random variables (r.v.'s). We let X_i be the r.v. denoting the service time of a customer on its i^{th} visit, and we denote the mean, variance and squared coefficient of variation of X_i by $\overline{x}_i$, $\sigma^2_{x_i}$, and $C^2_{x_i} \equiv \sigma^2_{X_i}/\overline{x}^2_i$, respectively. After the i^{th} service completion, a customer returns to the tail of the queue with probability q_i and leaves the system with probability $(1 - q_i)$, $i \geq 1$. Customers in the queue, both newly arrived and those that are fed back, are served in the order in which they joined the tail of the queue. The customer response time is defined to be the duration of time from customer arrival until departure including all intermediate feedbacks. An analysis of a more general model can be found in [Nelson 1987] and, more recently, in [van den Berg 1990]. Takács [Takács 1963] also derives the queue length distribution for the M/G/1 queue with Bernoulli feedback and uses that to derive the first and second moments of queue length. Disney [Disney 1981] considered issues relating the distribution of sojourn time in a Bernoulli feedback system and in an M/G/1 system without feedback. His results show that if the two systems have identical first three moments of total service time then the limiting distributions of sojourn times of the two systems are not identical. Other results regarding Bernoulli feedback, including delayed feedback, can be found in [Disney, König and Schmidt 1983] and [Disney and Kiessler 1987].

Queueing structures of this type can model systems in which customers receive service in stages. In timeshared scheduling, for example, jobs return to the tail of the queue after each quantum of service until their total service requirement is satisfied. In modeling such a system we would adjust the feedback probabilities, q_i, to match job service requirements.

In Section 1 we present equations for the expected response time for Bernoulli feedback, in

which customers return to the queue with a fixed probability after each service completion, and Fixed feedback, in which customers visit the queue a fixed number of times before departure. (The Bernoulli policy implies a geometric distribution for the number of visits to the queue by a customer.) An important motivation for studying these feedback policies is that, for many queueing systems with feedback, the Bernoulli policy is much simpler to analyze than the Fixed policy. For example, a fork/join queueing system with Bernoulli feedback has an exact analysis [Nelson 1990] whereas such a system with Fixed feedback appears to be intractable. It is then natural to ask whether knowledge of the response time for the Bernoulli policy can be used to shed light on the response time for the Fixed policy.

In Section 2 we present our results comparing these and other feedback policies. For the Bernoulli and Fixed policies with equal expected number of visits, one would expect the increased randomness of the Bernoulli policy to imply higher expected response times. We show, however, that the relationship between these policies depends strongly on the coefficient of variation of service time. In particular, for $C_x^2 > 1$ we show that the Fixed policy leads to higher expected response times. We provide tight bounds for the ratio of the response times of these two policies. Generalizing these results to other distributions of the number of visits, we show that the first and second moments of the number of visits seem to dominate in determining the expected response time. With a fixed mean number of visits to the queue, we show that in almost all cases the response time is a decreasing (increasing) function of the variance of the number of visits when $C_x^2 > 1$ ($C_x^2 < 1$).

1 Analysis of the Models

1.1 Bernoulli Feedback Policy

Takács [Takács 1963] has derived the queue length distribution and the first and second moments of queue length for the M/G/1 queue with Bernoulli feedback. The mean queue length is given by:

$$Q_B(\rho, q) = \frac{\rho}{1-q} + \frac{\rho}{1-q}\left(\frac{\rho(1+{C_x}^2) + \frac{2q\rho}{1-q}}{2(1-\frac{\rho}{1-q})}\right), \tag{1}$$

where $\rho \equiv \lambda\overline{x}$ is the utilization of the queue and $\overline{x}$ is the expected service time. One can also derive the above expression for $Q_B(\rho, q)$ by observing that an M/G/1 queue with Bernoulli feedback to the tail of the queue, as studied here, and an M/G/1 queue with Bernoulli feedback to the head of the queue have identical steady state queue length distributions. Feedback to the head of the queue implies that a customer who is fed back after receiving service immediately goes into service again. This system is, in fact, an M/G/1 system where the service time consists of a random number n of service intervals. Furthermore, n is geometrically distributed with parameter q, and its squared coefficient of variation, denoted by C_n^2, is given by $C_n^2 = q$. The squared coefficient of variation of the total service time then is $C_n^2 + \frac{C_x^2}{n}$. Using this expression in the Pollaczek-Khinchin formula for the expected queue length of an M/G/1 queue directly gives (**??**).

From Little's result [Little 1961] we can use (1) to derive the mean response time, $T_B(\rho, q)$, as:

$$T_B(\rho, q) = \frac{\overline{x}}{1-q} + \frac{\overline{x}}{1-q}\left(\frac{\rho(1+{C_x}^2) + \frac{2q\rho}{1-q}}{2(1-\frac{\rho}{1-q})}\right). \tag{2}$$

1.2 Fixed Cycling Policy

We next use the results of [Nelson 1987] to derive a closed-form expression for $T_F(\rho, N)$, the average response time in the fixed cycling case, The model analyzed there is the generalized model for M/G/1 with feedback. A job enters the queue as a class 1 customer, and is defined to be a class i customer on its i^{th} visit, $1 \leq i \leq N$. The utilization of the server by jobs of class i is given by $\rho_i = \lambda \overline{x}_i \prod_{j=1}^{i-1} q_j$. Let T_j denote the expected time that a job spends in the system as a customer of class j, if it visits the queue at least j times. The expected response time for a job is $R = \sum_{i=1}^{N} P[N_v = i] \sum_{j=1}^{i} T_j$, where $P[N_v = i]$, the probability that a job leaves the system after exactly i cycles, is calculated as $P[N_v = i] = (1 - q_i) \prod_{j=1}^{i-1} q_j$. As shown in [Nelson 1987], the values of T_i, $i = 1 \ldots N$ can be derived from the following set of linear recurrence relations:

$$T_j = \begin{cases} \overline{x}_1 + \sum_{k=1}^{N} \rho_k (r_k - \overline{x}_k + T_k), & j = 1, \\ \overline{x}_j + \sum_{k=1}^{j-1} \rho_{j-k} T_k + \sum_{k=1}^{N-j+1} \rho_{k+j-1} T_k, & j > 1. \end{cases} , \tag{3}$$

where $r_k \equiv \overline{x_k^2}/2\overline{x_k}$. For the fixed cycling policy ($q_i = 0$, $i < N$, $q_N = 1$), with $\overline{x}_i = \overline{x}$, and $\overline{x_i^2} = \overline{x^2}$, (**??**) reduces to

$$T_j = \begin{cases} x + N\rho(r - \overline{x}) + \rho \sum_{k=1}^{N} T_k, & j = 1, \\ x + \rho \sum_{k=1}^{j-1} T_k + \rho \sum_{k=1}^{N-j+1} T_k, & j > 1, \end{cases} \tag{4}$$

where $\rho \equiv \lambda \overline{x}$ and $r \equiv \overline{x^2}/2\overline{x}$. We can reduce these to a closed form expression by observing that $T_{n-j+2} = T_j$, $2 \leq j \leq N$, and therefore that $T_{j+1} = T_j$, $2 \leq j \leq N$. Also, $T_F(\rho, N) = \sum_{k=1}^{N} T_k$,

giving the following expression for $T_F(\rho, N)$:

$$T_F(\rho, N) = N\overline{x} + N\overline{x}\left(\frac{\rho(1+C_x^2)(1+N\rho)+2\rho(N-1)}{2(1+\rho)(1-N\rho)}\right). \tag{5}$$

2 Results

In this section we compare Bernoulli and Fixed cycling, as well as other feedback policies. In particular, we show that C_x^2, the squared coefficient of variation of the service time distribution, strongly impacts the relative performance of the various policies.

To simplify the notation, we hereafter write $T_F(\rho, N)$ and $T_B(\rho, q)$ as T_F and T_B respectively. In figure 1 we plot T_F and T_B as functions of C_x^2, for different utilizations. In calculating T_B, we set $q = 1 - 1/N$ so that the expected number of visits by a job to the queue is identical in both policies. If the service time on each visit is exponentially distributed, the two systems have identical mean response times because both are product form networks [Baskett et al. 1975] and the average response time only depends on the average total service requirement. In fact, this conclusion extends from the exponential case to all service time distributions with $C_x^2 = 1$ because the mean response time only depends on the first and second moments of service time, as (**??**) shows. For other service time distributions, the mean response time increases linearly with C_x^2 for both policies, as is expected from the properties of the M/G/1 queue,

Perhaps less obvious is the effect of C_x^2 on the *relative* values of T_F and T_B, for $C_x^2 \neq 1$. For $C_x^2 < 1$ ($C_x^2 > 1$), for all utilizations $0 < N\rho < 1$, $T_F < T_B$ ($T_F > T_B$). To show this we form the ratio

$$R(\rho, C_x^2) \equiv \frac{T_F(\rho, N)}{T_B(\rho, 1-\frac{1}{N})} = \frac{\rho(1+C_x^2)(1+N\rho)+2(1-N\rho^2)}{(1+\rho)(2-\rho+\rho C_x^2)} \tag{6}$$

which is the ratio of two linear expressions in C_x^2. Since $(ax+b)/(cx+d)$ (for positive a, b, c, d) is an increasing function of x if, and only if, $a/c > b/d$, it follows that T_F/T_B is an increasing function of C_x^2. Figure 2 shows how $R(\rho, C_x^2)$ varies with C_x^2 for different utilizations.

The behavior of the relative response times as C_x^2 is varied may be explained as follows. When C_x^2 is high, each time a job cycles to the tail of the queue there is a significant probability that some of the jobs ahead of it have large service times. In the Fixed cycling policy all jobs undergo exactly N cycles, whereas, in the Bernoulli feedback policy, the number of cycles has a geometric distribution and a large fraction of the jobs undergo only a few cycles before leaving the system. Although the average number of cycles is the same as the Fixed cycling policy, the jobs with fewer cycles have a very low response time and this decreases the *average* response time for the Bernoulli policy in comparison to that of Fixed cycling.

As figure 2 shows, with moderate values of C_x^2, the difference between the response times for Bernoulli feedback and Fixed cycling is small, but not negligible. As $C_x^2 \to \infty$, the ratio $R(\rho, C_x^2)$ remains finite:

$$\lim_{C_x^2 \to \infty} R(\rho, C_x^2) = \frac{1+N\rho}{1+\rho}. \tag{7}$$

This value increases with ρ for $N > 1$ and approaches $2N/(N+1)$ as $\rho \to 1/N$. It is less than 2 for all finite N. Finally, $R(\rho, C_x^2)$ is minimized when $C_x^2 \to 0$, $\rho \to 1/N$ and $N = 2$, and the minimum value is $8/9$. Thus, $R(\rho, C_x^2)$ satisfies $8/9 \leq R(\rho, C_x^2) \leq 2.0$.

We next determine the expected number of cycles a job must make with Bernoulli feedback to have a response time identical to the Fixed cycling policy with N cycles. In other words, find N^*, depending on N, such that $T_B(\rho, 1-\frac{1}{N^*}) = T_F(\rho, N)$. Setting $T_F = T_B$ with $q = q^* \equiv 1-\frac{1}{N^*}$

yields:

$$
\begin{aligned}
T_F = T_B &= \frac{\overline{x}}{1-q^*} + \frac{\overline{x}}{1-q^*}\left(\frac{\rho(1+{C_x}^2)+\frac{2q^*\rho}{1-q^*}}{2(1-\frac{\rho}{1-q^*})}\right) \\
q^* &= \frac{2(1-\rho)(T_F-\overline{x})-\rho\overline{x}(1+C_x^2)}{2T_F} \\
\frac{N^*}{N} &= \frac{1}{N}\left(\frac{2T_F}{2\rho T_F + 2\overline{x} - \rho\overline{x}(1-C_x^2)}\right),
\end{aligned}
$$

where T_F is given by (**??**). Again, we have a ratio of two linear expressions in C_x^2 and $\frac{N^*}{N}$ is an increasing function of C_x^2.

Figure 3 shows how $\frac{N^*}{N}$ varies with $N\rho$ for different values of C_x^2 (note the expanded scale on the Y-axis). For all values of C_x^2, $N^* = N$ when $N\rho$ is 0 or 1. This is to be expected since at $N\rho = 0$ there is no queueing, and at $N\rho = 1$ a very small increase in q^* would cause a very large increase in response time. For any intermediate value of ρ, $\frac{N^*}{N}$ increases with C_x^2, asymptotically approaching $\frac{N(1+N\rho)}{(N\rho)^2-N\rho^2+1+\rho}$ as $C_x^2 \to \infty$. Differentiating w.r.t. ρ shows that this achieves its maximum at $\rho = (\sqrt{2}-1)/N$. The maximum possible value of $\frac{N^*}{N}$ (attained at $C_x^2 \to \infty$, $N \to \infty$) is, thus, 1.2071.

For $C_x^2 < 1$, $R(\rho, C_x^2) < 1$, and hence $\frac{N^*}{N} < 1$; in fact $\frac{N^*}{N}$ decreases as $C_x^2 \to 0$. Even at $C_x^2 = 0$, however, $\frac{N^*}{N}$ remains very close to 1. This is because, in practical terms, the two systems have almost identical average response times when $C_x^2 < 1$.

To study how these results generalize to other distributions of N_v, the number of visits per job to the queue, we considered two example distributions, the Uniform and the Split. These

are defined by

$$\text{Uniform(N): } P[N_v = k] = \begin{cases} 1/N, & 1 \le k \le N, \\ 0, & k > N, \end{cases} \qquad \text{Split(p,N): } P[N_v = k] = \begin{cases} p, & k = 1, \\ 1-p, & k = N, \\ 0, & \text{otherwise.} \end{cases}$$

The mean response time for both cases can be obtained using (**??**).

We compared the four feedback distributions, keeping the average number of visits the same in all cases. Some of the results are listed in Table 1. In almost all cases, increasing the variance of the number of visits increases (decreases) the expected response time when $C_x^2 < 1$ ($C_x^2 > 1$). It is tempting to conclude that C_x^2 influences the expected response time in exactly this way, regardless of the higher moments of N_v. As seen from the table, however, for two distributions of N_v that have equal mean *and variance*, the response times are not equal, indicating that the response times also depend on the higher moments of N_v. In fact, the table shows a pair of distributions of N_v, viz. Split(0.258,133) and Uniform(197), for which Split has the higher variance of N_v *and* the higher response time.

The mean and variance of N_v nevertheless appear to have a dominant role in determining the response times. If this is correct, and the higher moments can essentially be ignored, then our conclusion that the Bernoulli policy could be used to bound or approximate other distributions of N_v is strengthened. We showed that the difference between T_B and T_F is reasonably small for moderate values of C_x^2; the difference would be smaller for any other distribution of N_v which had a variance between the Geometric and the Fixed distributions. For distributions with higher variance than the Geometric, the accuracy will depend both on the variance of N_v and on C_x^2.

Finally, to study how these observations generalize to somewhat more complex queueing systems, we used simulation to study the ratio of response times of Fixed to Bernoulli feedback

with an $M/G/K$ queue, for $K > 1$. In figure 4 we plot this ratio for an $M/G/8$ queue with $N = 5$ and $q = 4/5$, for three service time distributions, all with unit mean: deterministic ($C_x^2 = 0$), exponential ($C_x^2 = 1$) and hyperexponential ($C_x^2 = 5$). We used the regenerative method for output analysis and halted the simulation when the 95% confidence intervals were less than 5% of the simulated mean response times. (Although each simulated response time, when plotted alone, would appear smooth, the curves in figure 4 are jagged because much tighter confidence intervals are needed to obtain smooth ratio curves.) For high utilizations, the figure clearly shows that Fixed cycling has higher response times than Bernoulli, for $C_x^2 = 5$.

3 Conclusions

We analyzed and compared models of an M/G/1 queue with Bernoulli and Fixed feedback policies, and showed that the relative performance of the two policies changes as C_x^2 increases from values less than 1 to values greater than 1. We showed the ratio of the two response times to be bounded between 8/9 and 2, and demonstrated that the two systems show only a small difference in response times for moderate values of C_x^2. It appears that Bernoulli feedback could be used as an approximation for Fixed feedback under these conditions, considerably simplifying the analysis of such feedback systems.

We studied the dependence of the response time on the distribution of N_v, and showed that the first and second moments of N_v dominate in determining the average response times in most cases. We showed that increasing $\overline{N_v^2}$ while keeping $\overline{N_v}$ constant causes the expected response time to decrease (increase) when $C_x^2 > 1$ ($C_x^2 < 1$). We also showed, however, that the higher moments of N_v cannot be ignored, and gave one example of a pair of distributions where the distribution with the higher value of $\overline{N_v^2}$ also had the higher response time (with $C_x^2 > 1$).

Some interesting questions arise from this study. How would these comparisons extend to more complex feedback systems, such as if the M/G/1 queue was replaced by a more complex network of queues? If the central network satisfied the requirements for product form, for instance, the Bernoulli feedback model would have a relatively simple exact analysis, and might be a worthwhile approximation to other feedback policies. Results for higher moments of response time in the M/G/1 case would also be interesting. Also for the M/G/1 case determining the nature, in a stochastic ordering sense, of the relationship between the response times with different distributions of N_v is an interesting problem.

References

Baskett, F., K.M.Chandy, R.R.Muntz and F.G.Palacios. 1975. "Open, closed and mixed networks of queues with different classes of customers," *Journal of the ACM*, Vol.22, no.2, pp.248-260.

Disney, R.L., König, D., and Schmidt, V., 1983. "Stationary Queue Length and Waiting Time Distributions om Somg;e Server Feedback Queues", *Ad. Appl. Probab.*, Vol. 16, pp.437-446.

Disney, R.L. 1981. "A Note on Sojourn Times in M/G/1 Queues with Instantaneous Bernoulli Feedback", *Naval Res. Logist. Quart.*, Vol. 28, pp.679-684.

Disney, R.L. and Kiessler, P.C. 1987. *Traffic Processes in Queueing Networks*, John Hopkins University Press, Baltimore.

Kleinrock, L. 1976. *Queueing Systems, Volume I: Theory*, John Wiley, New York.

Little, J.D.C. 1961. "A Proof of the Queueing Formula $L = \lambda W$", *Operations Research*, pp. 383-387.

Nelson, R. 1990. "A Performance Evaluation of a General Parallel Processing Model," *Performance Evaluation Review*, Vol.18, no.1, pp.13-26.

Nelson, R. 1987. "Expected Response Time for an FCFS Feedback Queue with Multiple Classes," *IBM Research Report RC 13221.*

Takács, L. 1963. "A Single server queue with feedback," *Bell System Technical Journal*, vol.42, pp.505-519.

van den Berg, J.L. 1990. "Sojourn times in Feedback and Processor Sharing Queues," Ph.D. Thesis, Rijksuniversiteit te Utrecht.

van den Berg, J.L., O.J.Boxma and W.P.Groenendijk. 1989. "Sojourn times in the M/G/1 queue with deterministic feedback," *Stochastic Models*, vol.5, pp.115-129.

Wolff, R.W. 1982. "Poisson arrivals see time averages," *Operations Research*, Vol.20, 223-231.