Proceedings of the 1990 Winter Simulation Conference
Osman Balci, Randall P. Sadowski, Richard E. Nance (eds.)

BIAS PROPERTIES OF INFINITESIMAL PERTURBATION ANALYSIS FOR MULTI-SERVER QUEUES

Michael C. Fu

College of Business and Management
University of Maryland
College Park, Maryland 20742

Jian-Qiang Hu

Division of Applied Sciences
Harvard University
Cambridge, Massachusetts 02138

ABSTRACT

It is well-known that infinitesimal perturbation analysis gives biased estimates for multi-server queues when the service time distributions of the servers are not equal. In such cases, it is natural to ask how serious is the bias. In this paper, we analytically calculate the bias for infinitesimal perturbation analysis gradient estimators of mean steady-state system time (with respect to parameters in the interarrival or service time distributions) in a two-server Markovian queue where reversibility holds.

1. INTRODUCTION

The technique of infinitesimal perturbation analysis (IPA) was introduced by Ho and Cao [1983] for the sensitivity analysis of throughput in queueing networks. Suri and Zazanis [1988] applied the technique to the GI/G/1 queue for calculating sensitivities of system time of a customer w.r.t. a parameter of the arrival or service distribution, and proved strong consistency of the IPA estimators for the M/G/1 queue. Since then, numerous consistency proofs for the G/G/1 have followed. Fu and Hu [1990] studied the extension to the multi-server case, the G/G/m queue, proving strong consistency for the M/M/m queue. Unfortunately, the IPA algorithm fails (i.e., the estimator is neither unbiased nor consistent) when the servers are not identical. A heuristic argument as to why it fails is given in [Fu and Hu 1990]. Oftentimes when IPA fails, other perturbation analysis techniques such as those introduced in [Gong and Ho 1987; Ho and Li 1988] can be applied. However, the estimators derived using these techniques do not always have as desirable statistical or computational properties as IPA, so it is natural to investigate the actual magnitude of the IPA estimator error. This paper is a first step in that direction. In this paper, we study an analytically tractable, non-identical, two-server Markovian queueing system and calculate the steady-state bias of the IPA estimator, employing the technique of time-reversibility of Markov chains, which was used to prove consistency of the M/M/m queue in [Fu and Hu 1990]. Since it is believed that the statistical properties of IPA are insensitive to the actual underlying distributions, the calculation of the bias for this analytically tractable system should provide insight for the general case, e.g., on the order of the error with respect to server non-identicality. In section 2, we describe the IPA algorithm; in section 3, we present and discuss the bias result, with the details of the calculation provided in the appendix; and in section 4, we make some conclusions.

2. THE IPA ALGORITHM

We consider a multi-server first-come, first-served (FCFS) queueing system with a general renewal arrival process and general service time distributions. Steady-state system time, denoted by T, is our performance measure of interest, and we wish to estimate $dET/d\theta$ and $dET/d\alpha$, where θ is a parameter of the service time distribution and α is a parameter of the interarrival time distribution. Consider an arbitrary busy period of the system, and let $X_i(\theta)$ represent the service time of the ith (to arrive) customer in the busy period and $A_i(\alpha)$ represent the interarrival time between the (i-1)th and ith customers in the busy period. To describe the IPA estimator, we introduce the important concept of a server's **local** busy period: *the length between two adjacent idle times of the (same) server.* Thus, a single global busy period (busy period of the system) may contain any number (including 0) of local busy periods of a particular server.

Using this idea of local busy periods, and defining the set of customers preceding i in the same local busy period $L(i) = \{j < i : \text{j in the same local busy period as i}\}$, we have (see [Fu and Hu 1990] for details)

$$\frac{dT_i}{d\theta} = \sum_{j \in L(i)} \frac{dX_j}{d\theta} + \frac{dX_i}{d\theta}, \tag{1}$$

where T_i is the system time of the ith customer (in the busy period). This expression can also be written in recursive form:

$$\frac{dT_i}{d\theta} = \begin{cases} \frac{dX_i}{d\theta} & \text{if i initiates local busy period} \\ \frac{dT_{\hat{i}}}{d\theta} + \frac{dX_i}{d\theta} & \text{otherwise} \end{cases} \tag{2}$$

where $\hat{i} = \max_{j \in L(i)} j$, i.e., customer $\hat{i}$ is the index of the customer preceding i in the local busy period. Intuitively, the change in system time of customer i is the sum of the change in system time of the customer just preceding him in the same local busy period (if any such customer exists) plus the change in customer i's own service time. Glasserman [1990] has shown that this estimator (for the derivative of system time of the ith customer, not for steady state) is unbiased for the GI/G/m queue.

The explicit estimator for $dET/d\theta$ is then given by summing over all customers and can be implemented by the following algorithm (given for simplicity for a single parameter, but of course easily extendable to a vector of parameters):

IPA Algorithm for a Service Parameter of the GI/G/m Queue.
Initialize:
DTSUM=0
DXSUM(S)=0 for S=1,...,m, where m=# servers
At the end of service of customer j at server S:
DXSUM(S)=DXSUM(S)+$dX_j/d\theta$
DTSUM=DTSUM+DXSUM(S)
If no one waiting in queue, then server S becomes idle and DXSUM(S)=0
At the end of N customers served:
$(dT/d\theta)_M$=DTSUM/N

Analogously, for the arrival time parameter, we have

$$\frac{dT_i}{d\alpha} = -\sum_{j=i^*+1}^{i} \frac{dA_j}{d\alpha}, \tag{3}$$

and corresponding to Equation (1), we have:

$$\frac{dT_i}{d\alpha} = \begin{cases} 0 & \text{if i initiates local busy period} \\ \frac{dT_{\hat{i}}}{d\alpha} - \sum_{j=\hat{i}+1}^{i} \frac{dA_j}{d\alpha} & \text{otherwise} \end{cases}, \tag{4}$$

where i^* is the index of the customer who initiates the local busy period of customer i and $\hat{i}$ is the index of the customer who precedes i in the local busy period. Intuitively, the change in system time of customer i is the sum of the change in system time of the customer just preceding him in the same local busy period (if any such customer exists) plus the total change in customer i's own arrival time relative to the customer just preceding him in the local busy period.

An algorithm for $dET/d\alpha$ is given by

IPA Algorithm for an Arrival Parameter of the GI/G/m Queue.
Initialize:
DTSUM=0
DASUM(S)=0 for S=1,...,m, where m=# servers
At the entrance of service for customer j at server $\tilde{S}$:
DASUM(S)=DASUM(S)-$dA_j/d\alpha$ for all busy servers S
If customer j initiates a local busy period at $\tilde{S}$, then DASUM($\tilde{S}$)=0.
DTSUM=DTSUM+DASUM($\tilde{S}$)
At the end of N customers served:
$(dT/d\alpha)_M$=DTSUM/N

3. A TWO-SERVER MARKOVIAN QUEUE

For the M/M/m queue, Fu and Hu [1990] proved strong consistency for the IPA estimators by direct comparison with the analytic expressions. The technique used in the proof was time-reversibility of Markov chains. In this section, we utilize this technique to calculate the analytic values of the IPA estimators for a two-server case where the servers are unequal, and compare it with the analytic expressions for the actual derivatives. Many of the details are left to the appendix.

The system under consideration is a single queue, two-server system with unlimited capacity. The arrival process is Poisson with rate $\lambda = 1/\alpha$, and the service times are exponentially distributed with means θ_1 and θ_2. If both servers are available, then either server is chosen with equal probability. (This is the necessary and sufficient condition for reversibility to hold in the two-server case; see, e.g., [Wolff 1989, p.311].) Thus, the system can be represented as a continuous-time Markov chain, and so the stationary probabilities for number in system can be found by solving the flow balance equations, yielding the following:

$$p_n = \frac{\lambda}{\mu^*}\rho^{n-1}p_0, \tag{5}$$

$$\text{with } p_0 = \frac{1-\rho}{1+k\rho} \text{ and } k = \frac{1}{2}\left(\frac{\mu_1}{\mu_2} + \frac{\mu_2}{\mu_1}\right),$$

where $\mu^* = 2\mu_1\mu_2/(\mu_1+\mu_2)$, $\rho = \lambda/\hat{\mu}$, and $\hat{\mu} = \mu_1 + \mu_2$. The expected number in system is given by

$$E[N] = \frac{\lambda/\mu^*}{(1-\rho)(1+k\rho)}, \tag{6}$$

and applying Little's Law, the expected system time is

$$E[T] = \frac{1/\mu^*}{(1-\rho)(1+k\rho)}. \tag{7}$$

Differentiating with respect to α, we get

$$\frac{dE[T]}{d\alpha} = \frac{-\rho\lambda}{\mu^*}\frac{1-k+2k\rho}{(1-\rho)^2(1+k\rho)^2}. \tag{8}$$

With respect to the service time distribution, there are numerous choices for the parameter, e.g., θ_1 and θ_2 each separately. We have chosen to take $\theta_1 = \theta$ and $\theta_2 = c\theta$, so $k = (c+1/c)/2$, in which case we get

$$\frac{dE[T]}{d\theta} = \frac{c+1}{2}\frac{[1+k\rho^2]}{(1-\rho)^2(1+k\rho)^2}. \tag{9}$$

To calculate the expectation of the IPA estimators, we consider a customer, shown in Figure 1 and denoted by C_*, who arrives to the system in steady-state at time A_* and is the ith customer in the busy period. As in the previous section, we let $X_j(\theta)$ represent the service time of the jth customer served in the same busy period as C_*, and $A_j(\alpha)$ represent the interarrival time between the (j-1)th and jth customers in the busy period. Denote C_*'s system time by T, waiting time by W, and service time by X_* (so $T = W + X_*$).

For exponential interarrival times with mean α, we have $dA_j/d\alpha = A_j/\alpha$, and for exponential service times with mean θ or $c\theta$, we have $dX_j/d\theta = X_j/\theta$, so Equation (3) becomes

$$\frac{dT}{d\alpha} = \frac{-1}{\alpha}\sum_{j=i^*+1}^{i} A_j, \tag{10}$$

where recall that i^* is the index of the customer who initiates the local busy period of customer i, and Equation (1) becomes

$$\frac{dT}{d\theta} = \frac{1}{\theta}\sum_{j\in L(i)} X_j + \frac{X_*}{\theta}, \tag{11}$$

where recall that $L(i)$ is the set of indices of customers who precede customer i in the local busy period.

These two equations can be rewritten as

$$\frac{dT}{d\alpha} = \frac{-S}{\alpha}, \tag{12}$$

and

$$\frac{dT}{d\theta} = \frac{1}{\theta}(S+T), \tag{13}$$

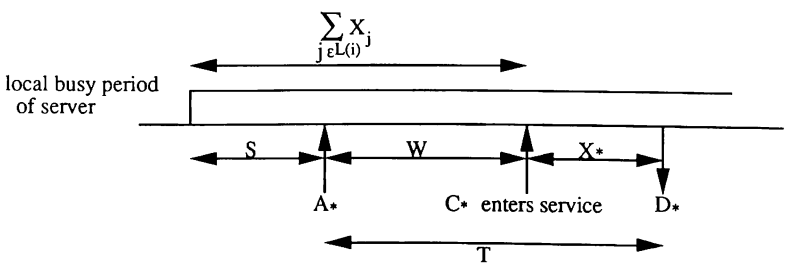

Figure 1. Customer C_* in Steady-State System

where S is the length of C_*'s local busy period upon C_*'s arrival (see Figure 1). Thus, the expectations of the IPA estimators are given by

$$E\frac{dT}{d\alpha} = \frac{-1}{\alpha}E[S], \tag{14}$$

and

$$E\frac{dT}{d\theta} = \frac{1}{\theta}(E[S] + E[T]). \tag{15}$$

To complete the calculations, we need E[S], for which we use the reversibility of the system - the details of which are provided in the appendix - to get

$$E[S] = \frac{2\rho^2/\mu^*}{(1-\rho)^2(1+\rho)(1+k\rho)}. \tag{16}$$

Substituting, we have

$$E\frac{dT}{d\alpha} = \frac{-2\rho^2\lambda/\mu^*}{(1-\rho)^2(1+\rho)(1+k\rho)}, \tag{17}$$

and

$$E\frac{dT}{d\theta} = \frac{c+1}{2}\frac{\rho^2+1}{(1-\rho)^2(1+\rho)(1+k\rho)}, \tag{18}$$

which match Equations (8) and (9) when $c = 1(\Rightarrow k = 1)$.

The bias terms then are just the differences between Equations (8) and (17), and between Equations (9) and (18), which we denote by $b_\alpha(k)$ and $b_\theta(k)$, respectively:

$$b_\alpha(k) = \frac{\rho^2(k^2-1)}{(1-\rho^2)(1+k\rho)^2}, \tag{19}$$

$$b_\theta(k) = \frac{-1}{\lambda\theta}\frac{\rho^2(k^2-1)}{(1-\rho^2)(1+k\rho)^2}. \tag{20}$$

where it can be checked that $b_\alpha(1) = 0$ and $b_\theta(1) = 0$, i.e., IPA is unbiased for the equal-server case. Interestingly enough, the two bias terms differ only by a multiplicative term not dependent on the non-identicality of the servers.

We investigate the two extremes now: when the servers are nearly identical and when the servers are far from identical. For this purpose, we define δ such that $c = 1 + \delta$.

Case 1: $\delta << 1$: Then, Equation (19) becomes

$$b_\alpha(\delta) \approx \frac{\rho^2\delta^2}{(1-\rho^2)(1+\rho)^2} \propto \delta^2, \tag{21}$$

and Equation (20) becomes

$$b_\theta(\delta) \approx \frac{-1}{\lambda\theta}\frac{\rho^2\delta^2}{(1-\rho^2)(1+\rho)^2} \propto \delta^2. \tag{22}$$

Thus, the biases are proportional to δ^2. Note that the biases go to 0 as ρ goes to 0, and to infinity as ρ goes to 1. Since the bias is caused by finite jumps in the sample performance function arising from event order changes in the sample path due to infinitesimal perturbations in the parameters (analytically, discontinuities in the sample performance function) - in this case causing switching between unequal servers (see [Fu and Hu 1990] for details) - and more occurrences of such phenomenon will occur at higher traffic intensities, this result makes intuitive sense.

Case 2: $c >> 1$ (and $k\rho >> 1$): Then, Equation (19) becomes

$$b_\alpha(c) \approx \frac{1}{1-\rho^2}, \tag{23}$$

and Equation (20) becomes

$$b_\theta(c) \approx \frac{-1}{\lambda\theta}\frac{1}{1-\rho^2}, \tag{24}$$

The biases are bounded by a constant for c large enough, when the system begins to act like a single-server system, except when empty. Again, both biases go to 0 as ρ goes to 0, and to infinity as ρ goes to 1.

4. CONCLUSIONS

Because it is believed that the applicability of IPA is in general distribution-independent, we would conjecture that the order of biases with respect to non-identicality - proportional to δ^2 - may be applicable to more general distributions, as well. This would mean that IPA is relatively insensitive to small deviations from the equal server case. Simulation studies are presently underway to verify this conjectures. Finally, it should be noted again that in cases where IPA is inappropriate, very often other forms of perturbation analysis can be applied, see e.g., [Gong and Ho 1987; Ho and Li 1988].

APPENDIX: CALCULATION OF E[S]

Recall that $\lambda = 1/\alpha$ is the arrival rate, $\mu_1 = 1/\theta$ and $\mu_2 = 1/(c\theta)$ are the service rates, $\rho = \lambda/(\mu_1 + \mu_2)$, and we consider a customer, C_*, who arrives to the system in steady-state at time A_*, shown in Figure 1.

To calculate E[S], we condition on the random variable N, where N = the number of customers in the system upon C_*'s arrival:

$$E[S] = \sum_{n=0}^{\infty} E[S|N=n]p_n = \sum_{n=0}^{\infty} p_n E_n[S] \tag{25}$$

where p_n = probability that C_* finds n customers in the system upon arrival (since Poisson arrivals see time averages) and E_n denotes the conditional (given $N = n$) expectation.

We note that for $N < 2$, C_* has no wait and initiates the local busy period, so Equation (25) becomes

$$E[S] = \sum_{n=2}^{\infty} p_n E_n[S]. \tag{26}$$

We need to calculate the term $E_n[S]$ for $n \geq 2$. For this calculation, we use the reversibility of the system. First, we rewrite the desired term by conditioning on the local busy period that C_* enters (see Figure 2):

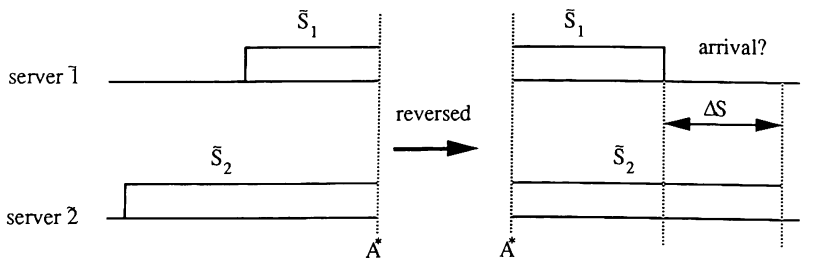

Figure 2. Time-Reversed Sample Path

$$E_n[S] = E_n[\tilde{S}_1]P[C_* \text{ served by } \tilde{1}] + E_n[\tilde{S}_2]P[C_* \text{ served by } \tilde{2}] \quad (27)$$

$$= E_n[\tilde{S}_1] + E_n\Delta S * P[C_* \text{ served by } \tilde{2}], \quad (28)$$

where $\tilde{S}_1$ is the length of the shorter local busy period when C_* arrives and $\tilde{1}$ denotes the server of the shorter local busy period, $\tilde{S}_2$ is the length of the longer local busy period $\tilde{2}$ when C_* arrives and $\tilde{2}$ denotes the server of the longer local busy period, and $\Delta S = \tilde{S}_2 - \tilde{S}_1$, the difference in length between the two local busy periods.

Using the time-reversed sample path shown in Figure 2, the first term is just the time to go from n down to 1 customers for an M/M/1 queue with arrival rate λ and mean service time $\hat{\mu} = \mu_1 + \mu_2$, which is equal to the sum of $(n-1)$ independent, identically distributed busy periods of the M/M/1 queue, i.e.,

$$E_n[\tilde{S}_1] = (n-1)\frac{1/\hat{\mu}}{1-\rho} \text{ for } n \geq 2. \quad (29)$$

To calculate the second term in Equation (28), we condition on whether $\tilde{S}_2$ corresponds to server 1's local busy period or server 2's local busy period:

$$\begin{aligned} E_n\Delta S * P[C_* \text{ served by } \tilde{2}] &= E_n[\Delta S|\tilde{2}=1] * P[C_* \text{ served by } \tilde{2}|\tilde{2}=1]P[\tilde{2}=1] \\ &\quad + E_n[\Delta S|\tilde{2}=2] * P[C_* \text{ served by } \tilde{2}|\tilde{2}=2]P[\tilde{2}=2] \\ &= E_n\Delta S_1 \frac{\mu_1}{\mu_1+\mu_2}\frac{\mu_2}{\mu_1+\mu_2} + E_n\Delta S_2 \frac{\mu_2}{\mu_1+\mu_2}\frac{\mu_1}{\mu_1+\mu_2} \\ &= \frac{\mu_1\mu_2}{(\mu_1+\mu_2)^2}(E_n\Delta S_1 + E_n\Delta S_2), \end{aligned} \quad (30)$$

where $E_n\Delta S_1 = E_n[\Delta S|\tilde{2}=1]$, $E_n\Delta S_2 = E_n[\Delta S|\tilde{2}=2]$, and where we have utilized reversibility and the memoryless property of the exponential distribution to derive the probabilities.

Viewing the time-reversed sample path in Figure 2, we see that the term $E_n\Delta S_1$ is the time from having only one customer in the system – and that one at server 1 – to the time when server 1's local busy period ends. There are two cases to consider, shown in Figure 3, depending on whether or not server 1 finishes service of its current customer before another customer arrives to the system. The probability of the former (finishing) is $\mu_1/(\lambda+\mu_1)$, the probability of the latter (not finishing) is $\lambda/(\lambda+\mu_1)$, while the mean time of the event (either departure or arrival) is $1/(\lambda+\mu_1)$. The former case indicates the end of server 1's local busy period, so $E_n\Delta S_1$ is simply $1/(\lambda+\mu_1)$. The latter case is more complicated. Given the arrival occurs before the departure at server 1, the system now has two customers. The expected time back down to one customer in the system is, as we calculated in the previous paragraph, $(1/\hat{\mu})/(1-\rho)$. The remaining customer is at server 1 with probability $\mu_2/(\mu_1+\mu_2)$ and at server 2 with probability $\mu_1/(\mu_1+\mu_2)$. If the customer is at server 2, then server 1's local busy period has ended, while if the customer is at 1, then we have returned to the situation that we began with, i.e., the remaining time at this point is again $E_n\Delta S_1$ (see Figure 3). Putting this altogether, we have

$$E_n\Delta S_1 = \frac{1}{\lambda+\mu_1} + \frac{\lambda}{\lambda+\mu_1}\left(\frac{1/\hat{\mu}}{1-\rho} + E_n\Delta S_1 \frac{\mu_2}{\mu_1+\mu_2}\right)$$

Solving, we get

$$E_n\Delta S_1 = \frac{1/\mu_1}{1-\rho^2}. \quad (31)$$

A completely analogous argument for server 2 yields

$$E_n\Delta S_2 = \frac{1/\mu_2}{1-\rho^2}. \quad (32)$$

Substituting into Equation (30), we get

$$E_n\Delta S * P[C_* \text{ served by } \tilde{2}] = \frac{1/\hat{\mu}}{1-\rho^2}. \quad (33)$$

Substituting into Equation (25), we have

$$\begin{aligned} E[S] &= \frac{\lambda}{\mu^*}p_0\sum_{n=2}^{\infty}\rho^{n-1}\left[(n-1)\frac{1/\hat{\mu}}{1-\rho} + \frac{1/\hat{\mu}}{1-\rho^2}\right] \\ &= \frac{\rho^2/\mu^*}{(1-\rho)^2(1+k\rho)} + \frac{\rho^2/\mu^*}{(1-\rho^2)(1+k\rho)} \\ &= \frac{2\rho^2/\mu^*}{(1-\rho)^2(1+\rho)(1+k\rho)}. \end{aligned}$$

REFERENCES

Fu, M.C. and J.Q. Hu (1990), "Consistency of Infinitesimal Perturbation Analysis for the GI/G/m Queue," to appear in *European Journal of Operational Research.*

Glasserman, P. (1990), "Structural Conditions for Perturbation Analysis of Queueing Systems," to appear in *Journal of Association for Computing Machinery.*

Gong, W.B. and Y.C. Ho (1987), "Smoothed Perturbation Analysis of Discrete-Event Dynamic Systems," *IEEE Transactions on Automatic Control AC-32,* 10, 858-867.

Ho, Y.C. and X.R. Cao (1983), "Optimization and Perturbation Analysis of Queueing Networks," *Journal of Optimization Theory and Applications 40,* 4, 559-582.

Ho, Y.C. and S. Li (1988), "Extensions of Perturbation Analysis of Discrete-Event Dynamic Systems," *IEEE Transactions on Automatic Control AC-33,* 5, 427-438.

Suri, R. and M.A. Zazanis (1988), "Perturbation Analysis Gives Strongly Consistent Sensitivity Estimates for the M/G/1 Queue," *Management Science 34,* 1, 39-64.

Wolff, R.W. (1989), *Stochastic Modeling and the Theory of Queues,* Prentice-Hall, Englewood Cliffs, NJ.

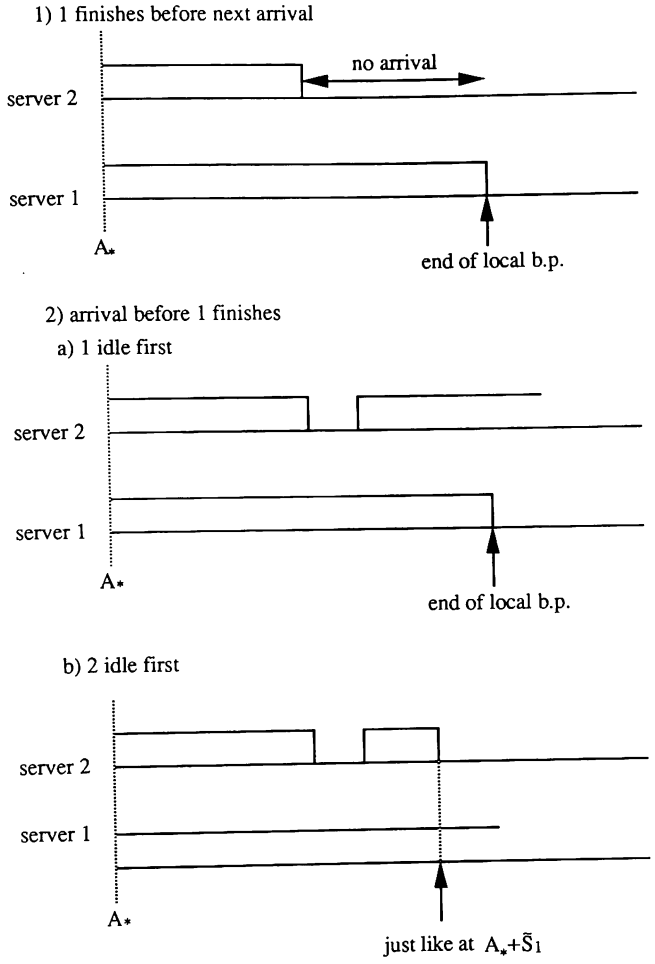

Figure 3. Cases to Consider in Calculating $E_n \Delta S_1$