Comparison of Call Gapping and Percent Blocking for Overload Control in Distributed Switching Systems and Telecommunications Networks

Arthur W. Berger, *Member, IEEE*

***Abstract*— Two overload control techniques are compared, percent blocking and call gapping, which have been used in distributed switching systems and in telecommunications networks. The comparison is based on nine criteria, seven of which concern robustness. The results are useful in the design of practical, effective overload controls.**

I. Introduction

OVERLOAD controls are investigated for a star topology network with throttles at the peripheral nodes (PN) and with a bottleneck at the central node (CN). Customers initiate jobs that pass through the PN and CN, but may be abandoned by the customer if the response time is high; in overload, and without throttling, the throughput of nonabandoned jobs declines substantially. The contribution of this paper is to compare the qualitative differences between two throttle schemes: percent blocking and call gapping. A "percent-blocking throttle" blocks and rejects an arrival with a given probability. A "call-gapping throttle" closes for a deterministic time interval, the gap size; after this interval, the next job to arrive passes through and the throttle again closes for the deterministic time interval. These throttle schemes have been used in network management of public switched telephone networks, and in distributed switching systems, [1]–[5].

The comparison of the throttle schemes is based on nine criteria, seven of which concern robustness. Robustness has practical consequences when one tries to design a throttle in a particular, real application. Typically, the "state" of the CN can not be transmitted continuously in time to the throttles at the PN's but rather is transmitted periodically and may change significantly between update intervals. Moreover, the value transmitted is typically from a small set, e.g., NORMAL, MINOR, MAJOR, CRITICAL. Given the typical granularity in the update intervals and in the value of the transmitted signal, a more robust throttle performs better over a larger range of contingencies. For example, when the throttle turns on and jobs are blocked, then customers may reattempt, resulting in an increase in the total arrival rate of jobs. A robust throttle could appropriately handle this increase, at least for the time period until the next signal from the monitor. To examine the robustness of the throttle schemes, we consider fixed, nonadaptive control settings and examine the resulting performance when exogenous parameter values change. Thus, the robustness discussed herein allows an easier design of practical, good controls.

In a related paper [6], we compare two patently dissimilar overload control schemes: 1) a dynamic, percent blocking throttle at the peripheral nodes and a standard first-in–first-out (FIFO) discipline at the central node, versus 2) no throttling at the peripheral nodes and a last-in–first-out (LIFO) discipline at the central node. The LIFO discipline with variations has been shown to be effective as an overload control strategy in stored program control switching systems [7]. In [6], a combination of the two schemes is suggested where the dynamic throttle regulates the arrivals to a level optimal for the LIFO discipline. In the present paper, the comparison is between two relatively similar overload control schemes (either of which could be used in conjunction with a LIFO discipline at the central node); nevertheless, the two schemes possess significantly different qualitative features.

II. Model

A block diagram of the model is given in Fig. 1. We assume that the arrival process to the ith PN, $i = 1, \cdots, n$ is Poisson with parameter λ_i and let λ denote the sum of the arrival rates, $\sum_{i=1}^{n} \lambda_i$. The arrival streams are independent of one another. If the call, or job, is blocked by the throttle at the PN, then it is rejected and lost from the system. If the call is not blocked at the throttle, then it departs the PN and moves to the CN. The superposition of the n departure processes from the PN's is the arrival process to the CN. Arrivals to the CN are queued and served according to a first-in–first-out discipline. The waiting space is assumed to be infinite, and we allow the possibility that the total arrival rate to the CN, λ', is greater than the service rate (causing an unstable queue). We assume that the CN cannot test prior to serving the call whether the customer has abandoned it. If it turns out that the call has been abandoned, then we say that the call is "bad", otherwise it is "good". We assume the service time distribution is exponential with parameter μ and is the same for both good and bad calls. Calls that are served at the CN may have become bad if their sojourn time in the system is large. The probability a call is good is assumed to decay exponentially as a function of sojourn time.[1] For example,

$$\text{Prob(call is good} \mid \text{sojourn time} = t) = e^{-\beta\mu t} \quad (1)$$

where β is the rate that calls turn bad, normalized by μ, [8]. We are interested in the case where the CN is the bottleneck of the system, thus we assume any time spent at the PN is negligible and is not modeled. This includes the case where the call returns to

Paper approved by the Editor for Communication Switching of the IEEE Communications Society. Manuscript received June 23, 1989; revised December 14, 1989.

The author is with AT&T Bell Laboratories, Holmdel, NJ 07733.

IEEE Log Number 9143107.

[1] If calls turned bad due to protocol timeouts, then a step function would be more appropriate. Also, if prior to serving the call, the CN could test whether it has been abandoned, then the probability a call is good is more appropriately modeled as a function of waiting time in the queue.

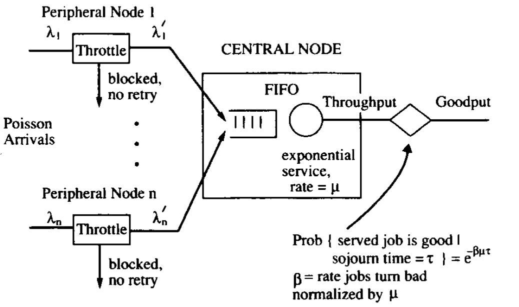


Fig. 1. Block diagram of the model.

a PN after being served at the CN and before exiting the system. Thus the modeled sojourn time in the system is the waiting and service time at the CN.

A. Calculation of Goodput

As defined in [8], "goodput" is the throughput times the probability of being good. It provides both a convenient scalar criterion for the tradeoff of throughput and delay, and also a reasonable abstraction for customer abandonments due to dial tone delay or post-dialing delay in telecommunication networks and switching systems, [8]. Let Γ = the normalized goodput, which is goodput divided by μ, i.e.,

$$\begin{aligned}\Gamma &= \frac{1}{\mu} \cdot \text{goodput} \\ &= \frac{1}{\mu} \cdot \text{throughput} \cdot P(\text{good}) \\ &= \frac{1}{\mu} \cdot \min(\mu, \lambda') \cdot P(\text{good})\end{aligned}$$

where $\lambda' = \sum_{i=1}^{n} \lambda_i'$ = arrival rate to the CN. The utilization of the server at the CN, denoted ρ, equals $(1/\mu)\ \min(\mu, \lambda')$; thus, $\Gamma = \rho \cdot P(\text{good})$. Conditioning on the sojourn time yields:

$$\Gamma = \rho \cdot \int_0^\infty \text{Prob}(\text{good} \mid \text{sojourn time} = t) \cdot dS(t)$$

where $S(t)$ = probability distribution function of the sojourn time. Substituting in (1) yields

$$\Gamma = \rho \cdot \int_0^\infty e^{-\beta\mu t} \cdot dS(t).$$

The last expression is simply ρ times the Laplace–Stieltjes transform of the sojourn time distribution evaluated at $\beta\mu$. Thus, finding the goodput for calls that exponentially turn bad amounts to finding the Laplace–Stieltjes transform of the sojourn time distribution.

B. Throttle Schemes

For percent blocking, the control variable is the probability of blocking, denoted b; for call gapping, the control variable is the gap size, denoted g. We adopt the viewpoint that the controls are set at the PN based on information from the CN that is sent periodically and simultaneously to all the PN's. Here, we are interested in behavior within periods of constant control value, and we make the simplifying assumption that the departure processes from the throttles are in statistical equilibrium.

III. Results

Table I summarizes the analytic results from the model. The results are straightforward to obtain; they either follow trivially from the definitions or exploit standard results from renewal theory and from $M/M/1$ and $G/M/1$ queues, [9], [10]. For the interested reader, Appendix A outlines the derivations. Table II summarizes the comparison of the two throttle schemes, and the following subsections provide details.

A. Robustness to the Rate that Calls Turn Bad, β

As β increases, calls turn bad more quickly and goodput declines. In the limit as $\beta \rightarrow \infty$, the goodput falls to zero. However, typical values for β are small. For the case of abandonments of call attempts due to dial-tone delay, the Prob(call is good | delay = 5 s) is typically in the range of 0.1–0.2, [11], and for abandonments due to post-dialing delay, it is typically in the range of 0.5–0.9. For a service rate of 1 call/s ($\mu = 1$), β ranges from 0.021 to 0.461 as Prob(call is good | delay = 5 s) varies from 0.9 to 0.1 (1). (Likewise, if $\mu = 100$, then β ranges from 0.00021 to 0.00461.) Moreover, for this range of β, both throttles are robust, i.e., there exists a fixed control setting that maintains the goodput reasonably close to the goodput that would have been attained with the optimal control setting. To illustrate this point, consider a throttle designed for abandonments due to post-dialing delay, and suppose the control is set optimally for a β corresponding to Prob(call is good | delay = 5 s) = 0.7. Holding this control fixed and letting β vary, Table III shows that the resulting decrease in goodput is relatively small, even for Prob(good | delay = 5 s) = 0.4. For completeness, Table III includes probabilities below 0.4, though these parameter values are more appropriate for a throttle designed to control abandonments due to dial-tone delay.

For the remainder of the paper, β is set to the bench mark value of 0.071 corresponding to Prob(good | delay = 5 s) = 0.7 when μ = 1.[2]

B. Sen.itivity to Control Settings and Maximum Goodput

When plotting goodput versus the control settings of the two throttles, we need to account for the control variables having different ranges: $b \in [0,1]$ and $g \in [0,\infty]$. We obtain a natural correspondence by associating a given value of b with a given value of g if they both yield the same utilization of the central processor. Fig. 2 compares the goodput for varying control settings, where the arrival rate is twice the service rate. (Note that the control value increases towards the left.) Scaled according to utilization, the plots have roughly the same shape and roughly the same degree of sensitivity to control settings; note in particular that *underthrottling* causes a sharp drop off in goodput.

Fig. 2 also shows that the goodput obtained from call gapping is higher than that from percent blocking, for equal levels of CN utilization. This occurs because the output process from the call-gapping throttle is smoother (interdeparture times have coefficient of variation < 1), and this leads to shorter queueing delays. For one active PN, the maximum goodput from call gapping is 11% higher than that from percent throttling; for other values of β, the maximum goodput from call gapping can be over 20% greater. In tightly designed systems, a 10–20% increase in goodput can be significant. However, as the number of

[2] The choice of parameter values for μ and β was made for illustrative purposes. The choice does effect the value of normalized goodput shown below in Figs. 2–4; however, the choice is not important to the qualitative comparisons demonstrated by the figures nor to the conclusions in Table II.

TABLE I
ANALYTIC RESULTS

	PERCENT BLOCKING	CALL GAPPING
Departure process from throttle i	Poisson. $\lambda_i' = \lambda_i(1-b).$	Renewal. Interdeparture times have a delayed exponential distribution $= 1 - e^{-\lambda_i(t-g)}, t \geq g. \lambda_i' = \frac{\lambda_i}{1+\lambda_i g}.$
Arrival process to CN (i.e., superposition of departure processes from throttles)	Poisson $\lambda' = \sum_{i=1}^{n} \lambda_i(1-b)$ $= \lambda(1-b).$	Superposition of renewal processes. $\lambda' = \sum_{i=1}^{n} \frac{\lambda_i}{1+\lambda_i g}.$ For balanced loads ($\lambda_i = \lambda/n$ for all i): $\lambda' = \frac{n\lambda}{n+\lambda g}.$
Criterion for Stability at CN	$b > \max\left(0, 1 - \frac{\mu}{\lambda}\right).$	Given implicitly by: $\sum_{i=1}^{n} \frac{\lambda_i}{1+\lambda_i g} < \mu.$ For balanced loads: $g > \max\left(0, n\left(\frac{1}{\mu} - \frac{1}{\lambda}\right)\right).$
Utilization of Server at CN	$\rho = \min\left(1, \frac{\lambda(1-b)}{\mu}\right)$	$\rho = \min\left(1, \frac{1}{\mu}\sum_{i=1}^{n} \frac{\lambda_i}{1+\lambda_i g}\right)$
Control setting that yields a specified utilization (Suppose $\lambda > \mu$, and a utilization of ρ^o is desired.)	$b = 1 - \frac{\mu\rho^o}{\lambda}.$	Given implicitly by: $\frac{1}{\mu}\sum_{i=1}^{n} \frac{\lambda_i}{1+\lambda_i g} = \rho^o$ For balanced loads: $g = n\left(\frac{1}{\rho^o\mu} - \frac{1}{\lambda}\right).$
Normalized Goodput	$\Gamma = \frac{\rho(1-\rho)}{1-\rho+\beta},$ with ρ given above.	$\Gamma = \frac{\rho(1-\sigma)}{1-\sigma+\beta}$, for n = 1, with ρ given above & σ given implicitly by: $\sigma = \frac{\mu}{\mu+\lambda}\sigma^2 + \frac{\lambda}{\mu+\lambda}e^{-\mu g(1-\sigma)}$
Optimal Control Setting	$b^{\text{opt}} = 1 - \frac{\mu\rho^{\text{opt}}}{\lambda}$ where $\rho^{\text{opt}} = \min\left(\frac{\lambda}{\mu}, 1 + \beta - \sqrt{\beta(1+\beta)}\right)$	Determined by numerical iteration, for n = 1.
Maximum Goodput	$\Gamma^{\text{opt}} = \frac{\rho^{\text{opt}}(1-\rho^{\text{opt}})}{1-\rho^{\text{opt}}+\beta}.$ For $b^{\text{opt}} > 0$, Γ^{opt} simplifies to $2\rho^{\text{opt}} - 1$.	Determined by numerical iteration, for n = 1.

TABLE II
SUMMARY OF COMPARISON OF CALL GAPPING AND PERCENT BLOCKING

	Advantage of		
Criteria	Call Gapping	Percent Blocking	Comments
Maximum goodput	✓		The smoother departure process from the call-gapping throttle yields shorter delays and higher goodput. Section III-B.
Robustness to changes in arrival rate, λ	✓		For static control setting, call gapping is more robust to changes in arrival rate, λ. Section III-C.
Robustness to changes in number of active sources, n		✓	Goodput from percent blocking is invariant to n; not so for call gapping. Section III-D.
Robustness to unbalanced loads		✓	Goodput from percent blocking depends on the individual arrivlal rates only via their sum; not so, for call gapping. Section III-F.
Range of stabilizing control values		✓	Call gapping requires more information; it depends on the number of sources, n, in addition to λ and μ. Table I.
Attaining a desired utilization of central processor		✓	Call gapping requires more information; it depends on n and the distribution of individual arrival rates. Table I.
Robustness to the rate calls turn bad, β	✓	✓	Both are robust over likely range for β. Section III-A.
Equity of who is blocked under unbalanced loads	✓	✓	Depending on circumstances and viewpoint, either throttle can be considered more equitable. Section III-E.
Sensitivity to control settings	—	—	Goodput is sensitive to underthrottling for both schemes. When we compare the control variables ($b\varepsilon[0, 1], g\varepsilon[0, \infty]$) according to equal utilization of CN, the sensitivity is the same. Section III-B.

TABLE III
ROBUSTNESS TO THE RATE CALLS TURN BAD, β

Arrival rate is twice service rate, $\lambda/\mu = 2$
Control is tuned for Prob(good | delay = 5 s) = 0.7

Prob(good given delay = 5 s)	Decrease in Γ due to fixed control as opposed to optimal control for:			
	Percent Blocking		Call Gapping	
	$\mu = 1$	$\mu = 100$	$\mu = 1$	$\mu = 100$
0.9	3.7%	0.5%	3.2%	0.4%
0.8	0.7%	0.1%	0.6%	0.1%
0.7	0.0%	0.0%	0.0%	0.0%
0.6	0.5%	0.1%	0.4%	0.1%
0.5	1.7%	0.4%	1.6%	0.3%
0.4	3.5%	0.9%	3.4%	0.7%
0.3	5.9%	1.7%	5.8%	1.4%
0.2	8.8%	2.9%	8.9%	2.4%
0.1	12.8%	5.2%	13.3%	4.4%

active PN's increases, the maximum goodput from call gapping declines. In Fig. 2, for 100 active sources and balanced loading (λ_i is the same at all PN's), then the maximum goodput from call gapping is only 2% greater. (The goodput from 100 active PN's is obtained by simulation. The 95% confidence intervals for the points plotted from the simulation are smaller than the height of the symbol "X" in the plots, and thus the confidence intervals have not been shown.)

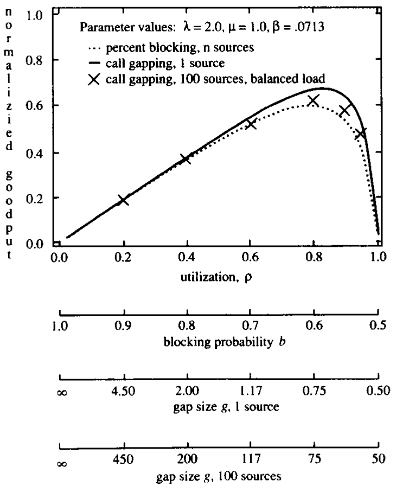


Fig. 2. Normalized goodput versus control settings, given fixed total arrival rate.

C. Robustness to the Total Arrival Rate, λ

Since call gapping can guarantee a minimal interdeparture time from the throttle, it is more robust to varying arrival rates than is percent blocking. However, although call gapping is better, it is not ideal, because if the gap size is set to guarantee stability for $\lambda \longrightarrow \infty$, then for nominal overloads, the control is overthrottling. Likewise, if the control is set for a nominal overload, then as $\lambda \longrightarrow \infty$, arrivals increasingly come just after the throttle opens and underthrottling occurs. Fig. 3 shows the case where control levels are optimally set for an arrival rate twice the service rate. Note that call gapping is more robust: it maintains high goodput over a broader range of arrival rates. (Simulation was used for the case of 100 active sources.)

The greater robustness from call gapping is particularly important when customers may reattempt. When the throttle is active, blocked customers may retry, and the arrival rate can noticeably increase prior to the next control update from the monitor. Hence, it is important that the given control setting maintain high goodput when faced with this increasing offered load. If we make the simplifying approximation that the combined load of first attempts and retries is Poisson, then the impact of reattempts is seen by letting λ grow greater than 2 in Fig. 3.

D. Robustness to the Number of Sources, n

Sometimes the number of active sources n is static and known, and robustness to variations in n is not of concern. However, the number of potential sources may be static while the number of active sources may vary. Percent blocking is more robust than call gapping to variations in the number of active sources; in fact, percent blocking is invariant to n, for given total arrival rate. Fig. 4 shows the sensitivity of call gapping for variable n where the control is optimally set for $n = 100$ and balanced loading. For call gapping, both simulation and approximate analytic results are shown. The analytic results make the approximation that the superposition of the renewal streams is Poisson. Under this approximation, the goodput from call gapping can be expressed in closed form as: $\Gamma = \rho(1-\rho)/(1-\rho+\beta)$ where $\rho = \min[1,(1/\mu)\cdot(n\lambda)/(n+\lambda g)]$.

E. Fairness Under Unbalanced Loads

In unbalanced loads, the arrival rates, λ_i $i = 1,\cdots,n$, are not all identical. Of typical interest are cases where some of the λ_i are significantly higher than others. Regarding the equity of who gets blocked and rejected under unbalanced loads, either control scheme could be preferred depending on circumstances. Call gapping more strongly controls the heavily loaded PN's, which is an advantage when load from one PN ought not to interfere with load from another. Percent blocking regulates all calls with the same probability, which is an advantage when there is no prior right to capacity, and preference ought not to be given to any PN, including ones that are lightly loaded.

F. Robustness to Unbalanced Loads

Consider a given total arrival rate λ and a variety of possible values for the arrival rate from each source, $\lambda_i, i = 1,\cdots,n$. For percent blocking, the goodput is invariant, since it only depends on the sum of the λ_i's. For call gapping, however, the goodput is not invariant, and thus we would like to choose the gap size to be robust over different apportionments of the λ_i's, i.e., over different possible values for the λ_i's given $\sum_{i=1}^{n} \lambda_i$ is constant. Suppose a g is found that is optimal for a given apportionment of the arrival rates, and suppose the apportionment were to change. The g would then be either underthrottling or overthrottling. Underthrottling has the more severe consequences, as the system may become unstable. Thus, it would be prudent to set the control for an apportionment such that any deviation would lead to overthrottling. This is attained if the g is set assuming a balanced load. One can show that any other apportionment of load yields a lower total arrival rate to the CN (a lower λ') and thus a lower ρ. Thus, the deviation from optimality would be in the direction of overthrottling, as desired.

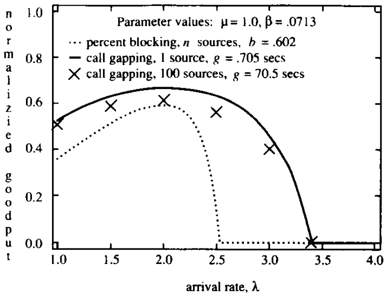


Fig. 3. Normalized goodput versus total arrival rate, given fixed control setting.

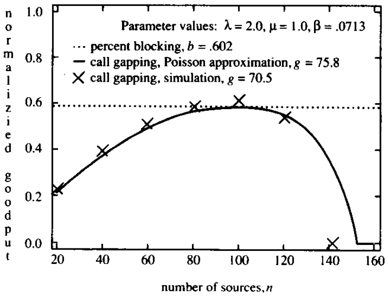


Fig. 4. Normalized goodput versus number of sources, given fixed total arrival rate and fixed control setting.

To make a quantitative comparison of percent blocking and call gapping, consider the scenario where the control is set for balanced loads, held fixed, and then the load changes to become unbalanced. Supposed that there are 100 active PN's, $\lambda_1 = \cdots = \lambda_{10}$, and $\lambda_{11} = \cdots = \lambda_{100}$, and the arrival rates in the first set are greater than those in the second set. Table IV shows that for large imbalances, the goodput from call gapping falls off significantly, while the goodput from percent blocking remains unaffected. (Note that the first row of Table IV corresponds to balanced loads, $\lambda_i = 0.02, i = 1,\cdots,100$.)

IV. Summary

Two overload control throttles, percent blocking and call gapping, have been compared. Each throttle has strengths and weaknesses, summarized in Table I. The key strengths of call gapping are 1) a greater robustness to changes in total arrival rate, and 2) higher goodput. For varying arrival rates, where the control setting is fixed, call gapping maintains reasonable goodput over regions where percent blocking has allowed goodput to fall to zero. Moreover, for optimal control settings, the maximum goodput from call gapping can be 10–20% greater than from percent blocking, due to the smoother departure process from the throttle.

TABLE IV
ROBUSTNESS TO UNBALANCED LOADS

Parameters: $\lambda = 2.,\ \mu = 1.,\ \beta = 0.0713,\ b = 0.602,\ g = 70.5$ s

$\sum_{i=1}^{10} \lambda_i$	$\sum_{i=11}^{100} \lambda_i$	goodput from call gapping	goodput from percent blocking
0.2	1.8	0.607	0.596
0.5	1.5	0.606	0.596
1.0	1.0	0.565	0.596
1.5	0.5	0.430	0.596
2.0	0.0	0.122	0.596

The strengths of percent blocking are 1) robustness to changes in number of active sources and 2) robustness to unbalanced loads. The optimal control setting for percent blocking is a function of the total arrival rate λ and is not a function of the number of active sources nor the individual arrival rates. (Call gapping is a function of these parameters.) Hence, the above robustness is actually an invariance, for given λ.

A possible control design is to use both throttle schemes where, as a function of overload state, an arrival would need to satisfy a call gapping criterion, or a percent blocking criterion, or both. Percent blocking might be used for minor overloads, and call gapping for critical overloads where the firm limit on the departure rate is beneficial.

V. APPENDIX A
DERIVATION OF ANALYTIC RESULTS OF TABLE I

A. Departure Process from Throttle i

For percent blocking, since the departure process is a Bernoulli decomposition of the Poisson arrival process, it is also Poisson. For call gapping, the interdeparture time equals the deterministic gap size g plus the time interval from the epoch the throttle opens to the arrival of the next customer. This latter interval has an exponential distribution since the interarrival times are exponential and the exponential distribution is memoryless. Thus, the interdeparture time distribution is a convolution of a point mass and an exponential, which is a delayed exponential. In particular, probability distribution function (pdf) of interdeparture times:

$$= 1 - e^{-\lambda_i(t-g)}, \quad t \geq g.$$

The mean interdeparture time is $(1/\lambda_i) + g$ and the variance is $1/\lambda^2$. The departure *rate* is $\left(\frac{1}{\lambda_i} + g\right)^{-1} = \frac{\lambda_i}{1+\lambda_i g}$.

B. Arrival Process to CN

For percent blocking, since the departure processes from the PN's are Poisson and the superposition of Poisson processes is Poisson, then so is the arrival process to the CN. For call gapping, the arrival process is the superposition of independent non-Poisson renewal processes. The arrival rate to the CN is the sum of individual departure rates from the PN's.

C. Criterion for Stability at CN

The CN is stable if the arrival rate is less than the service rate. For percent gapping, we have $\lambda(1-b) < \mu$. Solving for b and noting that $b \geq 0$, yields $b > \max(0, 1 - (\mu/\lambda))$. For call gapping, we have: $\sum_{i=1}^{n} \lambda_i/(1 + \lambda_i g) < \mu$.

D. Utilization of Server at CN

The utilization of the server ρ is the minimum of 1 and the arrival rate divided by the service rate; the expressions in Table I follow immediately.

E. Control Setting That Yields A Specified Utilization

Suppose $\lambda > \mu$, then any value of $\rho \in [0, 1]$ is attainable and suppose the value of ρ^o is desired. Substitute ρ^o for the utilization in Table I and solve for the control variable.

F. Normalized Goodput

From Section II-A, we have that the normalized goodput Γ equals

$$\Gamma = \rho \cdot \int_o^\infty e^{-\beta\mu\tau} \cdot dS(\tau)$$

where $S(\tau)$ = pdf of the sojourn time. Thus, Γ = $\rho \cdot$ Laplace–Stieltjes transform of sojourn time distribution evaluated at $\beta\mu$.

For percent blocking, since the arrival process to the CN is Poisson and the service times are exponential, then the system is $M/M/1$, and the sojourn time density function is: $\mu(1-\rho)e^{-\mu(1-\rho)t}$. The Laplace transform of the density, evaluated at $\beta\mu$, is $\mu(1-\rho)/(\mu(1-\rho) + \beta\mu) = (1-\rho)/(1-\rho+\beta)$. Thus, for percent blocking

$$\Gamma = \frac{\rho \cdot (1-\rho)}{1-\rho+\beta} \tag{A.1}$$

For call gapping with one active source, the arrival process to the CN is renewal with delayed exponential interarrival times, and exponential service times (a $GI/M/1$ system). The pdf of the waiting time is $1 - \sigma e^{-\mu(1-\sigma)t}$, $t \geq 0$ where σ is given implicitly by: $\sigma = A^*(\mu - \mu\sigma)$, and $A^*(\mu - \mu\sigma)$ is the Laplace transform of the interarrival time density evaluated at $\mu - \mu\sigma$, [10]. For one active source, $(n = 1), \lambda_i = \lambda_1 = \lambda$, and since interarrival times to the CN have density: $\lambda_i(t-g)e^{-\lambda_i(t-g)}$ $t \geq g$, then

$$A^*(\mu - \mu\sigma) = \frac{\lambda}{\lambda + \mu - \mu\sigma} e^{-g(\mu-\mu\sigma)}.$$

Thus, σ is given by $\sigma = [\lambda/(\lambda + \mu - \mu\sigma)]e^{-g(\mu-\mu\sigma)}$, which can be rearranged to

$$\sigma = \frac{\mu\sigma^2}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} e^{-\mu g(1-\sigma)}. \tag{A.2}$$

The Laplace transform of the sojourn time density is the product of the Laplace transform of the waiting time density and that of the service time density. The waiting time density equals $(1-\sigma)\delta_o(\tau) + \sigma\mu(1-\sigma)e^{-\mu(1-\sigma)\tau}$ where $\delta_o(\tau)$ is the unit delta function. Thus, the Laplace transform of the sojourn time density equals

$$\left[1 - \sigma + \frac{\sigma\mu(1-\sigma)}{s + \mu(1-\sigma)}\right] \cdot \left[\frac{\mu}{s+\mu}\right]$$

which simplifies to

$$\frac{\mu(1-\sigma)}{s + \mu(1-\sigma)}. \tag{A.3}$$

Equation (A.3) evaluated at $\beta\mu$ yields: $\frac{1-\sigma}{1-\sigma+\beta}$. Thus, for call gapping,

$$\Gamma = \frac{\rho \cdot (1-\sigma)}{1-\sigma+\beta} \quad \text{for } n = 1$$

where σ is given by (A.2).

G. Optimal Control Setting

For percent gapping, the normalized goodput is given by (A.1). Viewing ρ as the control variable in (A.1), we can optimize Γ over ρ, obtaining $\rho^{\text{opt}} = 1+\beta-\sqrt{\beta(1+\beta)}$. One can show that $1+\beta-\sqrt{\beta(1+\beta)}$ is $\epsilon\left[\frac{1}{2},1\right]$ for $\beta\ \epsilon\ [0,\infty)$. However, this optimal value for ρ, which is always $\geq \frac{1}{2}$, may not be attainable if λ is too small, as would be the case when the system is lightly loaded. When $\frac{\lambda}{\mu} < 1 + \beta - \sqrt{\beta(1+\beta)}$, then the blocking should be zero, yielding the highest utilization possible. Thus,

$$\rho^{opt} = \min\left(\frac{\lambda}{\mu}, 1+\beta-\sqrt{\beta(1+\beta)}\right). \tag{A.4}$$

The optimal blocking is the control setting that yields this utilization, namely, $b^{\text{opt}} = 1 - \frac{\mu\rho^{\text{opt}}}{\lambda}$.

H. Maximum Goodput

For percent blocking, maximum goodput is attained when the optimal utilization (A.4) is substituted into the normalized goodput (A.1). For overload where $b^{\text{opt}} > 0$, then $\rho^{\text{opt}} = 1+\beta-\sqrt{\beta(1+\beta)}$, and Γ^{opt} simplifies to $2\rho^{\text{opt}} - 1$.

References

[1] B. T. Doshi and H. Heffes, "Analysis of overload control schemes for a class of distributed switching machines," in *Proc. 10th Int. Teletraffic Congress,* Montreal, Canada, June 8–14, 1983, paper no. 5.2.2.

[2] A. Fukuda, "Input regulation control based on periodical monitoring using call gapping control," *Electron. Commun. Japan,* Part 1, vol. 69, no. 11, pp. 84–93, 1986.

[3] D. G. Haensche, D. A. Kettler, and E. Oberer, "Network management and congestion in the U.S. telecommunications network," *IEEE Trans. Commun.,* vol. 29, pp. 376–385, Apr. 1981.

[4] K. Kawashima, "Queueing models for congestion control in telecommunications systems," in *Oper. Res. '84,* J. P. Brans, Ed. The Netherlands, North-Holland, 1984, pp. 971–982.

[5] A. Kumar, "Adaptive load control of the central processor in a distributed system with a star topology," *IEEE Trans. Comput.,* vol. 38, pp. 1502–1512, Nov. 1989.

[6] A. W. Berger, "Overload control in star networks: Comparison of percent blocking throttle and LIFO queue discipline," submitted for publication.

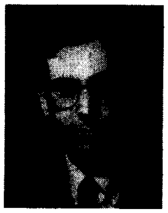
[7] L. J. Forys, "Performance analysis of a new overload strategy," in *Proc. 10th Int. Teletraffic Congress,* Montreal, Canada. June 8–14, 1983, paper no. 5.2.4.

[8] B. T. Doshi and E. H. Lipper, "Comparison of service disciplines in queueing systems with delay dependent behavior," in *Applied Probability-Computer Science: The Interface,* vol. II, R. L. Disney and T. J. Ott, Eds. Cambridge, MA: Birkhauser, 1982, pp. 269–301.

[9] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes.* New York: Academic, 1975.

[10] L. Kleinrock, *Queueing Systems Volume I: Theory.* New York: Wiley, 1975.

[11] L. Burkard, J. J. Phelan, and M. D. Weekly, "Customer behavior and unexpected dial tone delay," in *Proc. 10th Int. Teletraffic Congr.,* Montreal, Canada, June 8–14, 1983, paper no. 2.4.5.



Arthur W. Berger (S'82–M'83) was born in New York City on April 17, 1953. He received the B.S. degree in mathematics from Tufts University in 1974, and the M.S. and Ph.D. degrees in applied mathematics from Harvard University in 1980 and 1983, respectively.

Since 1983 he has been a member of technical staff at AT&T Bell Laboratories, Holmdel, NJ, where he has worked on network planning and on the performance analysis of telecommunication switching systems. His research interests are in the economic dynamics of pricing and the control of queueing systems.

Dr. Berger is a member of the IEEE Communications and the IEEE Control Societies.