Ensemble Polling Strategies for Increased Paging Capacity in Mobile Communication Networks

Christopher Rose & Roy Yates

Department of Electrical and Computer Engineering
Rutgers University
PO Box 909
Piscataway NJ 08855

ABSTRACT

The process of finding a unit in a mobile communications system is called paging and requires the use of limited network resources. Although it is understood how to minimize the use of network resources and satisfy delay constraints when paging a *single* unit, optimal policies for paging multiple units are difficult to derive. Here we adapt single unit polling methods to obtain simple *ensemble* polling schemes for use with multiple units which can greatly increase the rate at which page requests can be processed while maintaining acceptable average delay.

1 Introduction

Before information may be routed to a unit in a communication network, its location must be established by the system. Since unit mobility usually implies some degree of location uncertainty, the unit must be *polled* in various locations until found. This process is called *paging*. Since polling requires use of network signaling resources, we seek methods which minimize the amount of signaling required while still maintaining acceptable average delays between page initiation and unit localization.

Since the number of locations searched is often a surrogate for signaling cost, previous work considered the joint minimization of the number of locations polled and the mean delay in finding a *single* unit given a probability distribution on location [1, 2]. However, a multiuser system may need to service many paging requests simultaneously using only limited signaling bandwidth. Thus arises the problem of poll service discipline for a *group* of paging requests.

Classically, paging of a single unit is accomplished by polling all locations[1] within some group simultaneously; i.e. *blanket polling*. Alternately, a two-step procedure where a group of preferred locations is polled, followed by blanket polling if the first polling event is unsuccessful [3] could also be applied. Both these schemes, however, are needlessly wasteful if information about unit location is available: for example, a location probability distribution [1, 2].

In this work we consider simple planned ensemble polling strategies for the reduction of paging delay based on single unit optimal paging methods [1, 2]. We find that even in the absence of specific information about unit location, these methods permit larger page request rates than can be handled by blanket polling. Similar results have been obtained independently in [4] with an emphasis on the computational complexity of the solution as opposed the emphasis here on policies applicable to general unit location probability distributions.

Our approach is based on the concept of an *effective paging load* which depends not only upon the page request rate, but the average unit mobility and the paging discipline employed. As might be anticipated, lower mobility, or judicious grouping of locations polled requires fewer places to be searched and thereby results in an overall lower effective paging load for a given page request rate. Since polling queue delay dominates for blanket polling at higher page request rates, any reduction in the effective paging load can significantly reduce the queueing delay which in turn allows larger page request rates to be accommodated by the system.

[1] I.e., cells in a wireless system or nodes in a fixed-network for mobile computing.

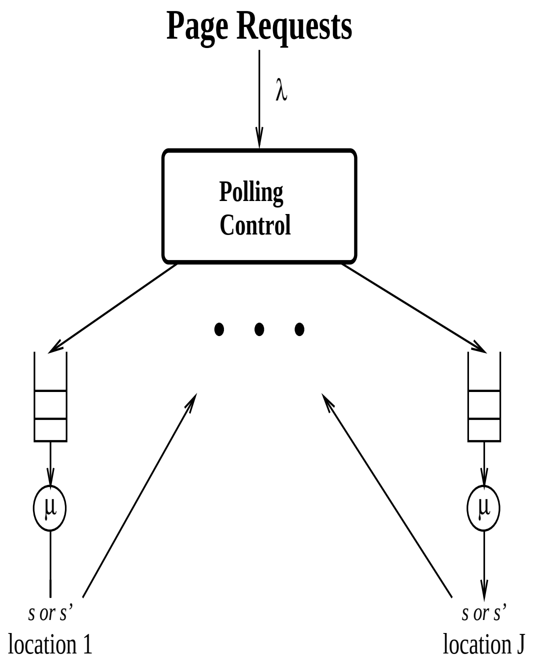

Figure 1: Block diagram of polling system. Page requests arrive according to a Poisson process with rate λ_p, each having an associated probability distribution p_j. Polling Control issues polling requests to appropriate locations. These requests are queued at the locations and serviced FCFS at average rate μ. The result (success: s, failure: s') of completed polls are fed back to the controller for appropriate further action.

2 Problem Statement

We assume page requests (incoming calls) arrive to the system according to a Poisson process with rate λ_p. Each unit m to be located has some known location probability distribution[2] $p_j^{(m)}$ over locations $j = 1, ...J$ and that individual polling events at given locations are serviced at a rate μ independent of service at other locations. A J-location system can then execute J simultaneous independent polling requests at any given time. Finally, assume that success/failure feedback (s/s') is available to the scheduling process. A block diagram of the procedure is supplied in Figure 1. Our goal is to find polling sequence policies which minimize delay.

The state of the paging system is determined by M, the number of units to be found and their associated location probability distributions $\{p_j^{(m)}(k)\}$, $m = 1, 2...M$ at the start of each paging step k. Since the next state, $[M, \{p_j^{(m)}(k+1)\}]$, depends only on the arrivals and the polling strategy, the process is Markovian. Thus, one can

[2]This distribution can be conditional on whatever is known: the mobility process, the time and place of last sighting, the time of day, the registration procedure favored by the mobile unit, etc. [5, 6].

envision designing polling strategies based on the current state which minimize the expected system paging delay. Unfortunately, the number of possible states is infinite and optimal policy selection methods such as Markovian decision process (MDP) theory [7] cannot be applied. Therefore, we turn to simple but effective heuristic methods inspired by results for optimal paging of single units [1, 2] to maintain acceptable average delay.

3 Analysis and Results

A queueing analysis of simple poll scheduling procedures is provided. In general, the results are approximate and based on independence assumptions about the network of polling queues and the arrival streams at each location. The accuracy of these approximations is later tested through Monte Carlo simulation.

3.1 Blanket Polling

In many systems, each page request generates simultaneous polling requests at all locations. If a unit resides at location i, then the paging response time for unit i is the response time of the paging queue at location i. The paging response time at a location $j \neq i$ is irrelevant since the unit will not be found there. Since at each location, the polling load is λ_p and the service rate is μ, the response times at all queues are identically distributed. The *normalized localization delay* $\bar{D}$ is the average time between request arrival and unit localization normalized by the average service time $1/\mu$. When a unit is in queue i, the normalized location delay is simply the system time at paging queue i. By the Pollaczek-Khinchin formula [8, 9],

$$\bar{D}(\rho) = 1 + \rho \frac{C_s^2 + 1}{2(1-\rho)} \tag{1}$$

where $\rho = \lambda_p/\mu$ and $C_s^2 = \sigma_s^2 \mu^2$ is the variance of the service time normalized by $1/\mu^2$, the mean service time squared. Exponential service provides $C_s = 1$ whereas deterministic service provides $C_s = 0$. We will hereafter assume exponential polling service at each location so that $C_s = 1$ and

$$\bar{D}(\rho) = \frac{1}{1-\rho} \tag{2}$$

Notice that if λ_p equals or exceeds μ, then the page request queueing system is unstable and the delay is infinite. Thus, as λ_p is increased, a service provider is forced to increase the rate of polling service μ and this may be unacceptable.

3.2 Sequential paging

From a heuristic standpoint we first note that global polling is wasteful when the unit location probability distribution is concentrated over a small subset of the J locations. In fact, from the standpoint of number of locations searched, it is usually almost a factor of two more wasteful to employ a global polling discipline [1]. We therefore first consider a heuristic approach to polling schedules based on previous analytic work and compare it to blanket polling

Specifically, in [1], it was shown that the expected number of paging requests is minimized if the locations in which a unit can be found are searched sequentially in decreasing order of probability. That is, if user m has a location distribution $\{p_j^{(m)}\}$ satisfying $p_{i_1}^{(m)} \geq p_{i_2}^{(m)} \geq \cdots$, then expected number of paging requests is minimized by sequentially searching locations $i_1, i_2, \ldots$. The number of locations searched L_m satisfies $P\{L_m = j\} = p_{i_j}^{(m)}$. The mean number of locations searched is the mean of the ordered distribution $p_{i_j}^{(m)}$ which can be expressed as

$$\bar{L}_m = \sum_{j=1}^{J} j p_{i_j}^{(m)}$$

In addition, [1] verifies that over all possible ordered distributions, $\bar{L}_m$ is maximized when all J locations are equally likely. In this case, $\bar{L}_m = (J+1)/2$. For now we assume that $\bar{L}_m = \overline{L}$ for all units. As stated previously, we will treat the location queues as if they were independent and assume the queue arrival streams are Poisson.

Thus, the overall arrival rate of polling requests to the system is $\rho\overline{L}$. However, unlike blanket polling, each polling request is processed at one of J locations. The *effective paging load* per polling queue is then $\rho\alpha$ where $\alpha = \bar{L}/J$. Therefore, following equation (1), the average normalized delay between a page request arrival and unit localization for this sequential polling algorithm is

$$\bar{D}_1 = \bar{L}\bar{D}(\rho\alpha) = \frac{\alpha J}{1 - \rho\alpha} \tag{3}$$

where the subscript 1 in $\bar{D}_1$ denotes that we are searching 1 location at a time. Since $\bar{L} \leq (J+1)/2$, we note that $\alpha \leq \frac{J+1}{2J}$ for any unit location probability distribution. Thus, sequential polling can always sustain larger page request rates than blanket polling.

In Figure 2 delay is plotted as a function of load ρ for blanket polling and sequential

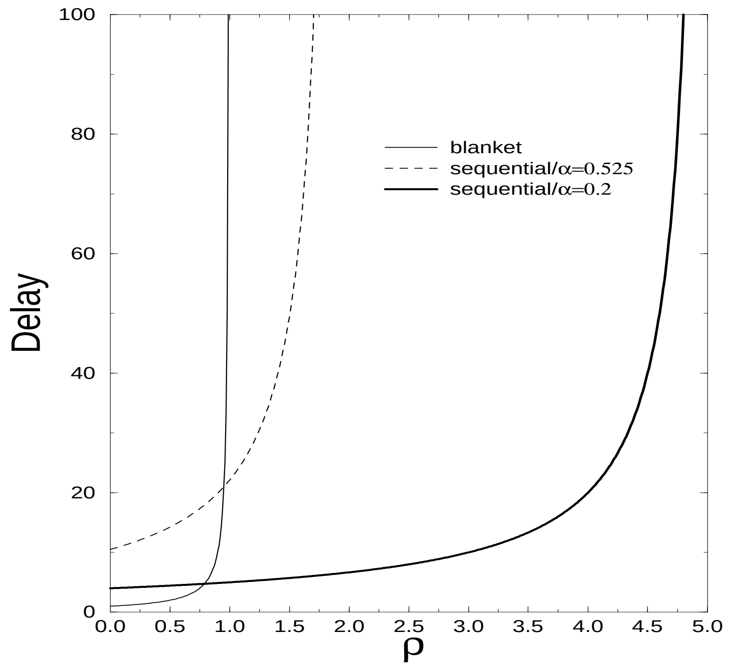

Figure 2: Delay as a function of offered load ρ for sequential and blanket polling. Twenty locations ($J = 20$), uniform ($\overline{L} = 10.5$, $\alpha = 0.525$), and nonuniform ($\alpha = 0.2$) location distributions.

polling for $J = 20$ locations, $\alpha = (J+1)/2J = 0.525$ (uniform distribution[3]) and $\alpha = 0.2$. The figure suggests that we can define a *crossover* load ρ^* as the load above which sequential paging delivers better delay performance than blanket polling. By setting equation (1) equal to equation (3), we have

$$\rho^* = \frac{\alpha J - 1}{\alpha(J-1)} \tag{4}$$

We plot ρ^* as a function of α in Figure 3. We see that crossover load decreases with decreasing location uncertainty (α). This implies that decreasing location uncertainty renders sequential polling effective over an increasingly wider range of offered page request load, ρ. One might therefore use a plot such as that in Figure 2 to describe a polling policy based on ρ and α; i.e., if offered load is below ρ^* use blanket polling, otherwise use sequential polling.

[3]It should be noted that the order of locations searched should always be purposely randomized for a uniform distribution since following a particular order would clearly lead to load imbalances: with later-polled locations having extremely low load and first-polled locations having much higher load.

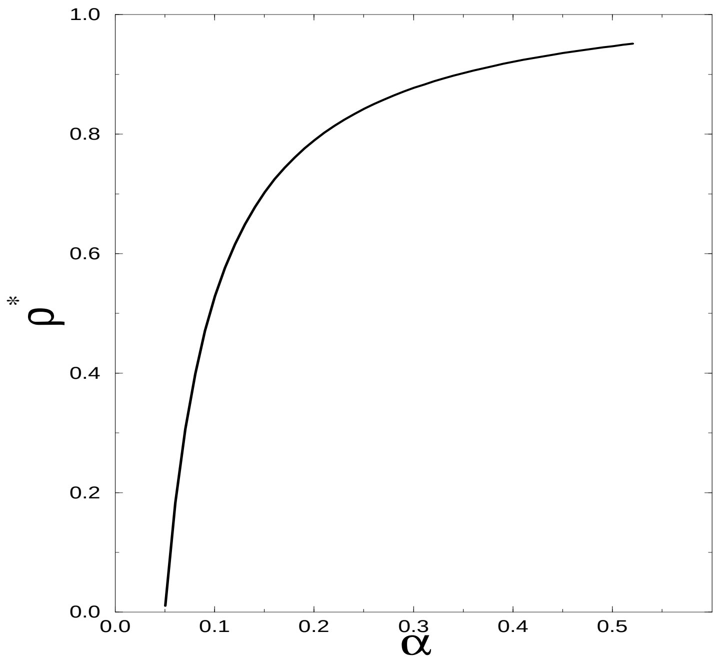

Figure 3: Crossover load ρ^* above which sequential polling gives lower delay than blanket polling as a function of α, the average fraction of locations which must be searched before the unit is found. Twenty locations ($J = 20$).

3.3 Sequential Group Paging

The poorer delay performance of sequential paging at lower loads is a consequence of the sequential nature of the search algorithm. It was shown in [1] that sequential search maximizes the expected delay. However, it was also shown that by constructing groups of locations to be paged simultaneously, the mean number of locations searched could be held reasonably close to the optimal obtained by sequential search while the delay performance improved dramatically.

Thus, we consider a variation on the sequential paging procedure and instead of searching individual locations in decreasing order of probability, we search groups of k locations where each group $\mathcal{G}_i$ contains the k most likely locations not already searched.[4] Note that the last group will have fewer than k elements if J/k is non-integer. The total number of groups is $G = \lceil J/k \rceil$.

First we define q_i as the probability the unit will be found in location group i; i.e.,

[4]Such a uniform grouping with k elements each is in general suboptimal [1]. However, uniform groupings simplify the analysis here.

$q_i = \sum_{j \in \mathcal{G}_i} p_j$. The mean number of groups searched is then,

$$\bar{g} = \sum_{i=1}^{G} i q_i \tag{5}$$

If a unit is in location group $i < G$, then ki locations will be searched while if the unit is in location group G then all J locations are searched. Hence, the mean number of locations searched is

$$\bar{L} = J q_G + k \sum_{i=1}^{G-1} i q_i = (J - kG) q_G + k\bar{g} \tag{6}$$

As before, the offered load seen by each polling queue is $\rho\alpha$ where $\alpha = \bar{L}/J$. Thus, the mean normalized polling delay for each queue is given by $\bar{D}(\rho\alpha)$

Now notice that if a unit is in location j of group $\mathcal{G}_i$, the mean response time after polling group $\mathcal{G}_i$ is simply $\bar{D}(\rho\alpha)$, the mean response time of polling queue j. That is, the responses of the other locations in group $\mathcal{G}_i$ are irrelevant given the unit resides in location j. Now consider that before group $\mathcal{G}_i$ may be polled, negative responses must be received from *all* members of groups $j < i$. We define $\hat{D}_k$ as the normalized mean delay to determine a polling failure in a group of k locations. We may then write the total normalized average unit localization delay as

$$\bar{D}_k = (\bar{g} - 1)\hat{D}_k + \bar{D}(\rho\alpha) \tag{7}$$

We now calculate $\hat{D}_k$.

If we assume that the individual queues behave as if they were M/M/1 and independent, the known CDF for the queueing delay at each location, $F_X(x)$, may be used to calculate $\hat{D}_k$. Specifically, define $Z = \max(X_1, X_2, \ldots X_k)$ as the random variable corresponding to the time required for all k polls to complete service. The CDF of Z is $F_Z(z) = F_X^k(z)$. Thus,

$$\bar{Z} = \int_0^\infty (1 - F_X^k(z)) dz \tag{8}$$

Continuing with the M/M/1 assumption, the stationary probability of m paging requests is $\pi_m = (1 - \alpha\rho)(\alpha\rho)^m$ where $\alpha = \bar{L}/J$. If a polling request arrives and finds m customers in the queue, then the probability density on system time is Erlang of order $m + 1$:

$$f_{X|M}(x|m) = \frac{\mu^{m+1} x^m \mathrm{e}^{-\mu x}}{m!} \tag{9}$$

Using π_m to compute the marginal $f_X(x)$ results in

$$f_X(x) = \mu(1-\alpha\rho)\mathrm{e}^{-\mu(1-\alpha\rho)x} \tag{10}$$

The CDF is therefore

$$F_X(x) = 1 - \mathrm{e}^{-\mu(1-\alpha\rho)x} \tag{11}$$

Making the variable substitution $u = F_X(z)$ in equation 8 yields

$$\begin{aligned} \bar{Z} &= \frac{1}{\mu(1-\alpha\rho)} \int_0^1 \left(\frac{1-u^k}{1-u}\right) du \\ &= \frac{1}{\mu(1-\alpha\rho)} \int_0^1 \left(1 + u + \cdots + u^{k-1}\right) du \\ &= \frac{1}{\mu(1-\alpha\rho)} \sum_{\ell=1}^{k} \frac{1}{\ell} \end{aligned} \tag{12}$$

Since $\hat{D}_k$ is the normalized expected time, $\hat{D}_k = \mu\bar{Z}$. From equation (7), $\bar{D}_k$ becomes

$$\bar{D}_k = \frac{1}{(1-\alpha\rho)} \left[(\bar{g}-1)\sum_{\ell=1}^{k}\frac{1}{\ell} + 1\right] \tag{13}$$

Notice that for $k = 1$ we have $\bar{g} = \bar{L}$ and the result of equation (13) reduces to that in equation (3).

Analytic insight into the behavior of $\bar{D}_k$ with k can be obtained by assuming a uniform location probability distribution. We then have

$$q_i = \begin{cases} k/J & i = 1, 2, \ldots G-1 \\ 1 - k(G-1)/J & i = G \end{cases} \tag{14}$$

so that

$$\bar{g} = \sum_{i=1}^{G-1} \frac{k}{J} i + G(1 - k(G-1)/J) = G\left(1 - \frac{k(G-1)}{2J}\right) \tag{15}$$

and

$$\overline{L} = J - k(G-1) + \frac{k^2 G(G-1)}{2J} \tag{16}$$

so that

$$\alpha = \bar{L}/J = 1 - (G-1)\frac{k}{J}\left(1 - \frac{k}{2J}G\right) \tag{17}$$

Remembering that $G = \lceil J/k \rceil$ allows us to plot $\bar{D}_k$ as a function of ρ for various k in Figure 4. For offered loads where $\lambda_p/\mu \approx< 0.8$, global polling ($k = 20$) provides the lowest delay. However, as the offered load rises, successive reductions of group

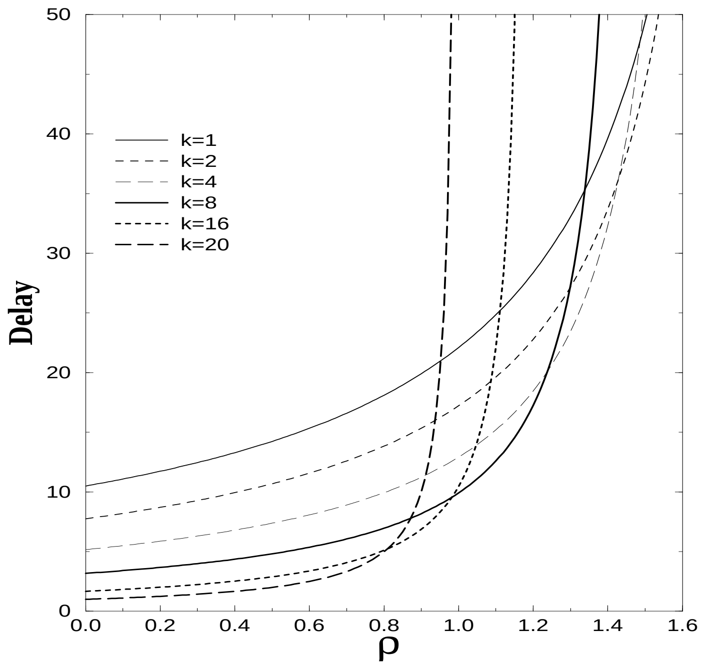

Figure 4: Plot of delay versus ρ for various group sizes k. Twenty locations ($J = 20$) with uniform location distribution ($\alpha = 0.525$).

size k, allow the delay to be kept manageably low. Numerical calculations for nonuniform distributions show that this general behavior becomes more pronounced as α decreases. The family of curves in k compresses vertically and expands horizontally so that smaller values of k result in acceptable delay at smaller ρ than for the uniform case. An example is shown in Figure 5.

4 Verification Using Monte Carlo Simulation

The analysis of section 3 assumes that each unit has the same location probability distribution except for a random ordering of locations. Here we introduce a slightly more realistic model for deriving the user location probability distributions and compare the results of Monte Carlo simulations to the analysis.

When a page request for unit u is serviced at some time t_0, the system learns the user location, x_0. If the next page request for unit u occurs at time t_1, then the system has available both the last sighting time t_0 and the last known location x_0. In principle, this information can be used to form an estimate of the unit location if something is known about the stochastic process which describes the motion of unit u.

Assuming that each unit moves according to similar stochastic processes, if the

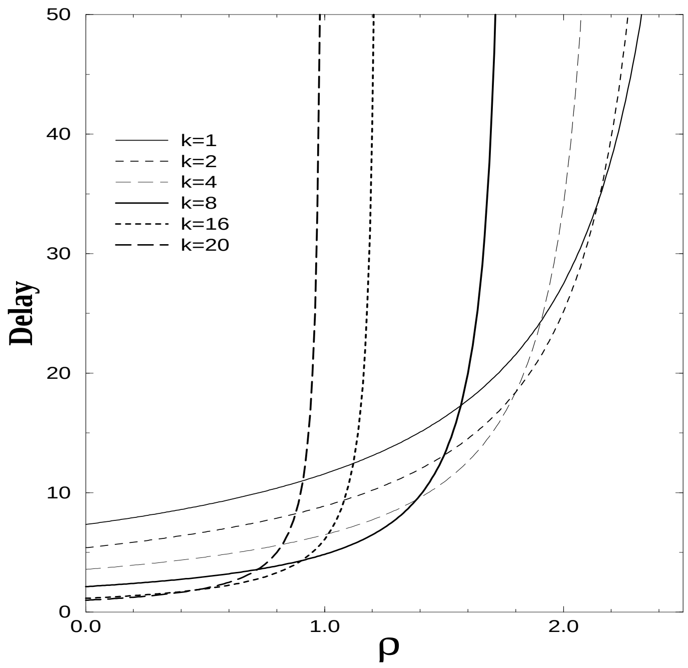

Figure 5: Plot of delay versus ρ for various group sizes k. Twenty locations ($J = 20$) with a linearly decreasing location distribution, $p_j = \frac{2}{J}(1 - \frac{j}{J+1})$. $E[j] = (J+2)/3$ so that $\alpha = 0.3\bar{6}\bar{6}$ for $J = 20$.

times between last sighting and page request is some constant $t = T$ for all units, then the ordered values of location probability for each unit are identical except that the physical ordering is non-random. More precisely, the stochastic motion process can introduce the notion of location contiguity in that likely (or unlikely) locations might be clustered together. This may lead to correlations among the polling queues. As a further complication, we might also assume that the time between last sighting and page request is a random variable t. Then each unit can have a potentially different location distribution each time it is paged [5, 6].

Despite these complications, we shall see that the analytical results seem robust with respect to the manner in which the location distributions are derived. Through use of average distribution properties (*mobility indices*), the delay characteristics for a variety of situations not explicitly covered by the analysis can be adequately predicted.

4.1 The Mobility Model

We use a simple mobility model. Locations are arranged in a J-element "racetrack" of radius 1 as depicted in Figure 6. We will assume that the units move according to a continuous diffusion process described in Appendix A. In this diffusion model, we verify in Appendix A, that a unit that starts at angular position θ_0 at time $t = 0$

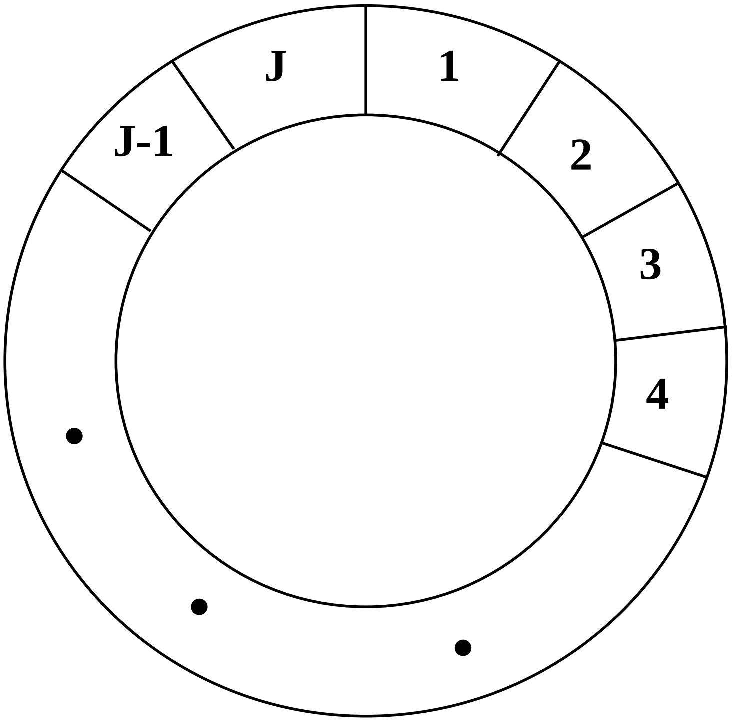

Figure 6: Racetrack model of unit locations. Units may move between adjacent sites according to some motion process.

is at location j at time t with probability

$$p_j(t) = \sum_{k=-\infty}^{\infty} \left[\Phi\left(\frac{2\pi(j/J+k)-\theta_0}{\sqrt{\Gamma t}}\right) - \Phi\left(\frac{2\pi((j-1)/J+k)-\theta_0}{\sqrt{\Gamma t}}\right) \right] \tag{18}$$

where Γ is the diffusion coefficient and $\Phi(x)$ is the normal CDF

$$\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx \tag{19}$$

We will also assume that the initial position on the track, θ_0, is random and takes on only J discrete values evenly spaced between 0 and 2π; i.e., $\theta_0 = 2\pi m/J$, $m = 1, 2 \ldots J-1$.

4.2 Results Using Exponential Service

We first consider the case of uniform location probability by assuming Γt is large; that is, the time between the last sighting and the page request is large enough that the resulting distribution is effectively uniform. This most closely represents the assumptions of the analysis in section 3. We verify Figure 4 for various group sizes k with results shown in Figure 7. For both $k = 1$ and $k = 20$, the agreement is almost exact as might be predicted from queueing network theory. For other k, where

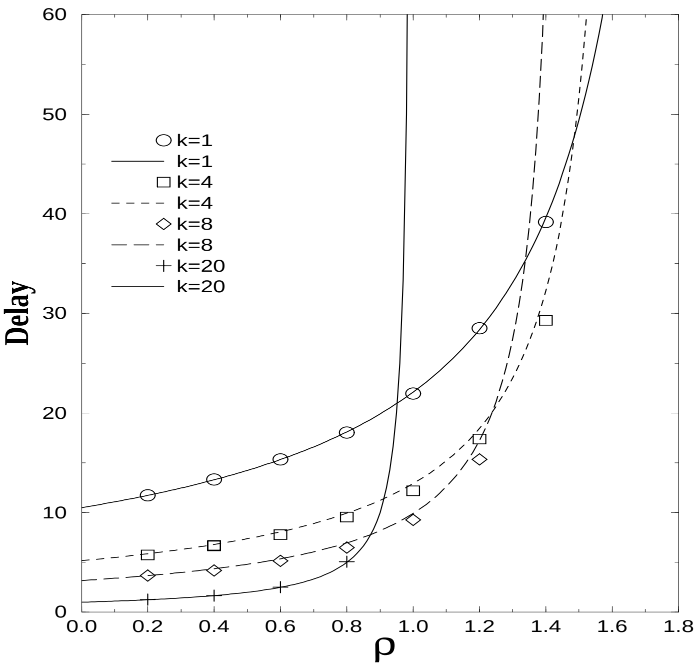

Figure 7: Monte Carlo simulation points superimposed on a plot of delay versus ρ for group sizes $k = 1, 4, 8, \&20$. Twenty locations ($J = 20$) and uniform location distribution.

the correlations between arrival and departure processes are more complicated, the agreement is better at lower loads.

We then consider the case where Γt is chosen so that the unit location probability distribution is nonuniform ($\alpha = 0.25$ when $k = 1$) but identical for each page request, except for the random rotation θ_0. This corresponds to a fixed time between last sighting and page request for all units. The result is shown in Figure 8.

Finally, we consider the case where the parameter t is itself a random variable. This corresponds to a more realistic scenario where a unit is paged some random time after its last contact with the system so that its last position is known. For simplicity we approximate t as an exponential random variable with parameter λ_u. Such an approximation is reasonable if we also assume that $E[t] = 1/\lambda_u$ is much greater than the sum of the mean page request sojourn time in the queueing system and the mean time spent in conversation.[5] In this way, each page request for any given unit arrives with a potentially different location probability distribution since Γt is a random variable. We define $\overline{L}(t)$ as the mean of the corresponding parametrized distribution. Thus, $\overline{L}(t)$ is a random variable.

[5]Tens of minutes for intercall arrival intervals for a single unit is typical. Comparatively, a tolerable page service time should under a few seconds and the mean conversation time is generally on the order of 90 seconds for voice telephony.

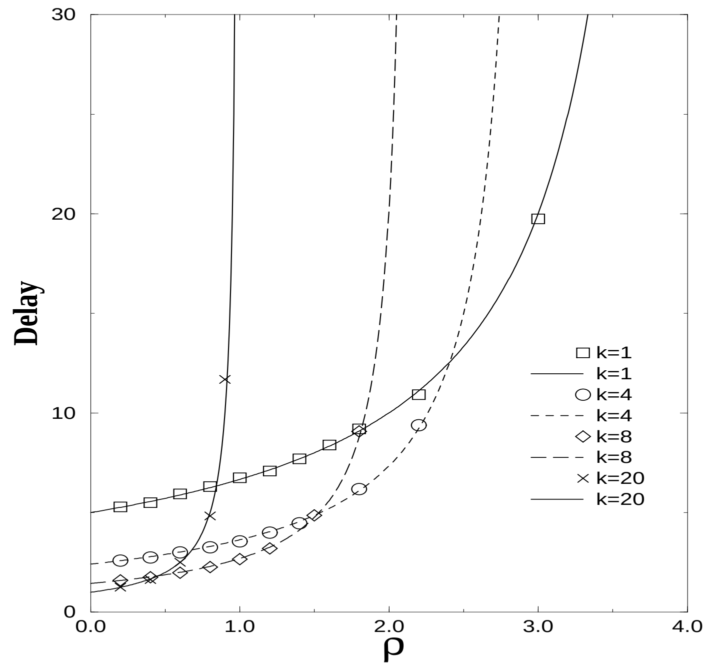

Figure 8: Monte Carlo simulation points superimposed on a plot of delay versus ρ for group sizes $k = 1, 4, 8, \&20$. Twenty locations ($J = 20$) and nonuniform location distribution with Γ in equation (18) chosen such that $\alpha = 0.25$ when $k = 1$.

We define $\mathcal{M} = \frac{\Gamma}{\lambda_u}$ as the mobility index. Each value of $\mathcal{M}$ corresponds to a unique value of $\alpha = E_t[\overline{L}(t)]/J$ via equation (18). Thus, α may be considered a surrogate for mobility index. Values of α approaching $\frac{J+1}{2J}$ imply high mobility index whereas values of α approaching $1/J$ imply low mobility index. It should be noted that mobility index is a measure not of the absolute increase of unit location uncertainty with time, but rather a measure of the expected increase in location uncertainty between times of unit contact with the system [10].

The result of simulations using $\alpha \approx 0.25$ (when $k = 1$) are provided in Figure 9 and show good agreement with the analytic approximations.

4.3 A Brief Experiment with Deterministic Service

Extension of the analytic approximation used in the previous sections to deterministic service is difficult for at least two reasons:

- At higher loads, the arrival streams to the queues will be composed mainly of failed prior requests. Since these prior requests were served deterministically, the overall arrival process to the queues can be extremely regular rather than Poisson. This regularization of the arrival process tends to reduce the average system time (reminiscent of [11]) so that delays calculated based on Poisson

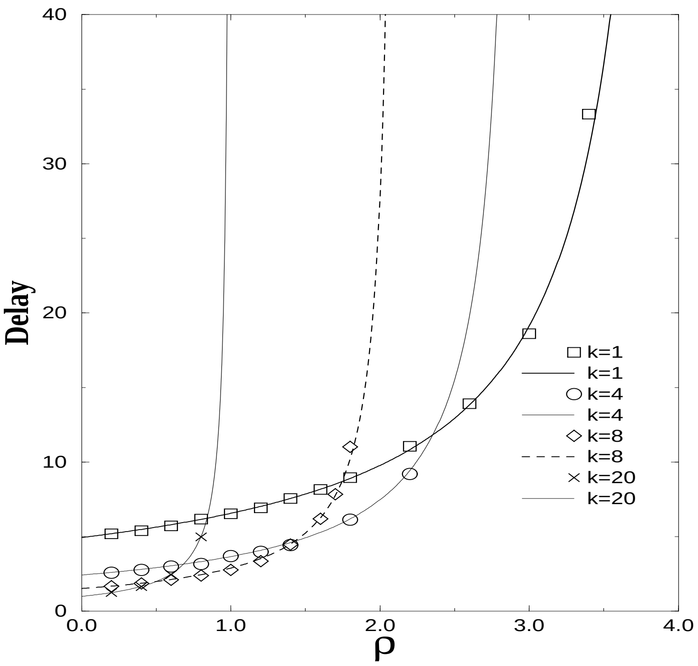

Figure 9: Monte Carlo simulation points superimposed on a plot of delay versus ρ for group sizes $k = 1, 4, 8, \&20$. Twenty locations ($J = 20$) and nonuniform location distribution with mobility index $\mathcal{M}$ chosen such that $E_t[\overline{L}(t)]/J = \alpha \approx 0.25$ (when $k = 1$).

arrivals will be pessimistic.

- The group failure delay calculation of equation (12) was based explicitly upon the delay probability distribution for exponential service and must be redone for deterministic service. The calculation is difficult.

For these reasons we present simulation results for deterministic service and show that the family of delay curves in group size k has a morphology similar to that obtained for exponential service. Specifically, we see in Figure 10 that as offered load increases, low delay may still be obtained by using successively smaller group sizes. The primary difference between the deterministic service curve family and its exponential counterpart is the extremely close nesting of the the curves as k increases.

This close nesting nesting has a simple explanation. At low loads the location polling queues are usually empty. For this reason, the delay incurred by a polling group failure is almost always exactly one service time as compared to $\sum_{\ell=1}^{k} 1/\ell$ service times for exponential service. Thus, for deterministic service, the normalized mean unit localization delay is simply the mean number of groups polled, $\bar{g}$. Horizontal lines are drawn at delays $\bar{g}(k)$ for $k = 1, 2, 8\&20$ in Figure 10. These lines also show close nesting in k and correspond closely to the low load portions of the delay curves

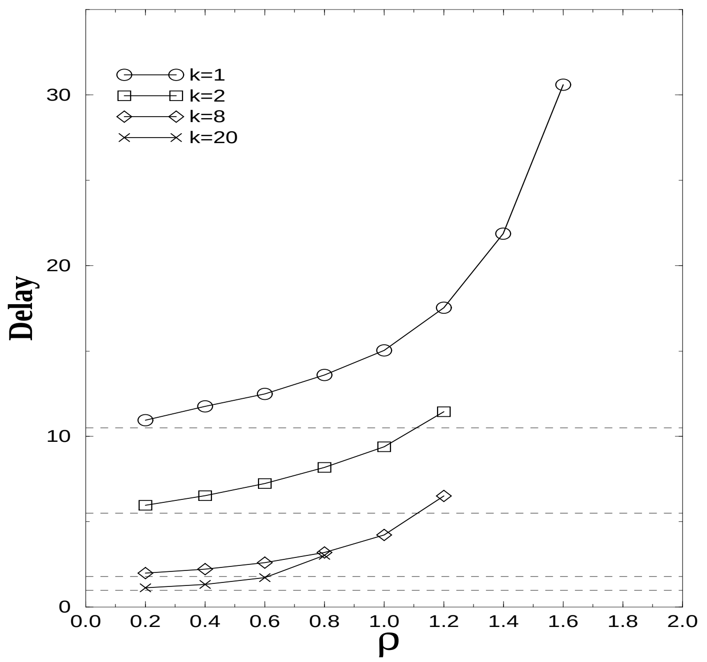

Figure 10: Monte Carlo simulation points superimposed on a plot of delay versus ρ for group sizes $k = 1, 2, 8\&20$. Twenty locations ($J = 20$) with a uniform location distribution and deterministic service. Notice the more rapid decline in delay with group size k as compared to exponential service (see Figure 7). Corresponding zero load delays shown as horizontal dashed lines (see text for explanation).

for $k = 1, 2, 8\&20$ respectively.

5 Summary and Conclusions

We have considered the problem of paging multiple units over J locations as a problem in service discipline selection at a stochastic service facility. Our simple discipline assumes that groups of k locations are polled for each unit in decreasing order of probability. This formulation allows the analytic results to be expressed simply in terms of *average* mobility parameters such as $\alpha = \bar{L}/J$, the mean proportion of locations searched and $\bar{g}$, the mean number of groups searched.

We have found that at low loads, blanket polling ($k = J$) provides the best delay performance. However, as paging rate λ_p increases toward the maximum polling rate μ, the blanket polling delay tends toward infinity. At some point it is advisable to reduce k and indeed sets of discrete values of $\rho = \lambda_p/\mu$ can be found above which group size k should be reduced. The highest sustainable paging rates λ_p are achieved when $k = 1$. It may therefore prove useful to adaptively adjust k based on estimates of λ_p, μ and α similar to methods proposed for mobile routing [12].

The unit mobility also strongly affects the number of locations which must be searched. If few locations are searched on average, then large sustainable page request rates are possible. For example, if 1/4 of the locations need be searched on average (as compared to approximately 0.5 for a uniform location distribution), then the maximum sustainable page request rate is approximately four times that achievable by blanket polling. However, even if little is assumed about unit location (i.e., a uniform distribution), the maximum sustainable rates are nearly a factor of two above those attainable by blanket polling.

In this work we have avoided extending results to other service disciplines besides exponential or characterization of the service time profile for a particular paging system. Our reasons are twofold. First, the analysis of queueing networks with non-exponential service is difficult. Second, an attempt to consider specifics for any of myriad possible scenarios would at best be anecdotal and at worst only muddy the analytic waters.

Instead we simulated deterministic service and found the general morphology of the delay curve family in k to be similar to that for exponential service. Since the statistics of deterministic and exponential service differ appreciably, we venture to guess that results for other service distributions will lie somewhere between the two. Limited experiments with other simple service distributions (such as uniform) have borne out this hypothesis.

Although ensemble paging is generally necessary in any mobile communications system whether land-line or wireless, an interesting problem arises when the system is wireless. Specifically, most wireless mobile systems issue polling signals over the air and await responses from mobile units. The targeted mobiles respond over a shared multi-access channel. For low loads, a fixed average page service model might be reasonable since contention is low. However, at higher loads with many units responding, contention could induce greater average delays per message. This effect could be modeled as a page service rate μ which *decreases* with offered load. Thus, for a system with multi-access page response channels, one might expect the gains afforded by effective paging load reduction to be even more impressive than in the case of fixed average page request service rate.

We have left open the question of *optimal* state-based service disciplines owing to difficulties of problem complexity suggested earlier. Nonetheless, we are currently exploring state-based algorithms using information theory coupled to the paging theory developed in [1]. We hope to compare the performance of these more complex algorithms to the simple scheme presented here in the near future.

Acknowledgements

This work was supported in part by a grant from the National Science Foundation (NRCI-92-06148). We would like to thank Imrich Chlamtac, the Editor-in-Chief of ACM Wireless Networks and the two anonymous reviewers for their invaluable comments on the manuscript.

A Diffusion on a Ring

In a diffusion process of rate Γ, let $X(t)$ denote the position at time t of a particle that starts at position 0 at time $t = 0$. The position $X(t)$ has the Gaussian distribution [13, 14, 15, 16]

$$f_{X(t)}(x) = \frac{1}{\sqrt{2\pi\Gamma t}} e^{-x^2/2\Gamma t} \tag{20}$$

We map the diffusion process on the real line to a unit circle by defining the angular position $\Theta(t)$ of a particle on the unit circle at time t via

$$\Theta(t) = \theta \text{ iff } X(t) = 2\pi k + \theta \text{ for some integer } k$$

Applied to the racetrack model of Figure 6, we say that a unit is at position $j \in \{1, \dots, J\}$ at time t if

$$\frac{2\pi(j-1)}{J} < \Theta(t) \leq \frac{2\pi j}{J}$$

Hence, the probability a unit is at position j is

$$p_j(t) = \sum_{k=-\infty}^{\infty} P\{2\pi((j-1)/J + k) < X(t) \leq 2\pi(j/J + k)\} \tag{21}$$

An initial condition $\Theta(0) = \theta_0$ is equivalent to the initial condition $X(0) = \theta_0$. Since $X(t) + x_0$ describes the position of a diffusion process on the real line started at position x_0, we can include the condition $\Theta(0) = \theta_0$ by mapping $X(t) + \theta_0$ to the unit circle. In this case, a particle started at position θ_0 at time $t = 0$ is at location j on the racetrack at time t with probability

$$\begin{aligned} p_j &= \sum_{k=-\infty}^{\infty} P\{2\pi((j-1)/J + k) < X(t) + \theta_0 \leq 2\pi(j/J + k)\} \\ &= \sum_{k=-\infty}^{\infty} \left[\Phi\left(\frac{2\pi(j/J + k) - \theta_0}{\sqrt{\Gamma t}}\right) - \Phi\left(\frac{2\pi((j-1)/J + k) - \theta_0}{\sqrt{\Gamma t}}\right)\right] \end{aligned}$$

where Γ is the diffusion coefficient and $\Phi(x)$ is the CDF of the normal distribution:

$$\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx \tag{22}$$

References

[1] C. Rose and R. Yates. Minimizing the average cost of paging under delay constraints. *ACM Wireless Networks*, 1(2):211–219, 1995.

[2] S. Madhavapeddy, K. Basu, and A. Roberts. Adaptive paging algorithms for cellular systems. *Fifth WINLAB Workshop on Third Generation Wireless Information Networks*, April 1995. New Brunswick, NJ.

[3] G. Pollini and S. Tabbane. The Intelligent Network Signaling and Switching Cost of an Alternative Location Strategy Using Memory. In *Proc. IEEE Vehicular Technology Conference, VTC'93*, May 1993. Seacaucus, NJ.

[4] David Goodman, P. Krishnan, and Binay Sugla. Minimizing Queueing Delays and Number of Messages in Mobile Phone Location. *ACM-NOMAD*, 1996. to appear.

[5] C. Rose. Minimizing the average cost of paging and registration: A timer-based method. *ACM Wireless Networks*, 1996. In press.

[6] C. Rose. State-based paging/registration: A greedy technique. Winlab-TR 92, Rutgers University, December 1994. (submitted, IEEE Transactions on Vehicular Technology, 12/94).

[7] H.C. Tijms. *Stochastic modeling and analysis : a computational approach.* Chichester; New York: Wiley, 1986.

[8] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory.* Wiley, second edition, 1985.

[9] L. Kleinrock. *Queueing Systems, Vol.1.* Wiley-Interscience, 1975.

[10] C. Rose. Minimization of paging and registration costs through registration deadlines. *IEEE/ICC'95*, pages 735–739, 1995. Seattle.

[11] B. Hajek. The Proof of a Folk Theorem on Queueing Delay with Applications to Routing in Networks. *Journal of the Association for Computing Machinery*, 30(4):834–851, Oct. 1983.

[12] R. Yates, C. Rose, S. Rajagopalan, and B. Badrinath. Analysis of a mobile-assisted adaptive location management strategy. *ACM Wireless Networks*, 1995. In press.

[13] G.F. Simmons. *Differential Equations with Applications and Historical Notes.* McGraw-Hill, 1972.

[14] F.B Hildebrand. *Advanced Calculus for Applications,* 2^{nd}*. ed.* Prentice Hall, Englewood Cliffs, 1976.

[15] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York, third edition, 1991.

[16] S.M. Ross. *Stochastic Processes.* John Wiley & Sons, New York, 1983.