Ensemble Polling Strategies for Mobile Communications Networks

Christopher Rose and Roy D. Yates
Electrical and Computer Engineering, WINLAB
Rutgers University
Piscataway, N.J. 08855

Abstract

In communications systems, mobile units must be found before information may be routed to them. The process of finding each unit, called paging, requires the use of limited radio and fixed network resources. Thus, the rate at which units can be found on average and the rate at which page requests can be satisfied subject to a delay criterion is intimately tied to the polling discipline employed by the system. Here we consider simple polling schemes which can greatly increase the rate at which page requests can be processed while maintaining acceptable average delay.

1 Introduction

Unit mobility in a communication network usually implies some degree of uncertainty in the location of a unit. Therefore, before information may be routed to a unit, that unit's location must be established by the system. This process is called *paging* and requires that the system *poll* various locations until the unit is found.

Previous work considered the joint minimization of the number of locations polled and the mean delay in finding a unit given the probability distribution on that unit's location [1, 2]. However, in a practical mobile system servicing many units, there may be many paging requests in progress simultaneously and only limited paging bandwidth with which to handle them. In this way arises the problem of poll service discipline for a group of paging requests.

In most current systems, all locations (cells) within the service area are polled for a single unit simultaneously; i.e. *blanket polling*. Others suggest a two step procedure where a group of preferred locations are polled, followed by a blanket page if the first polling event is unsuccessful [3]. This scheme is needlessly wasteful if information about unit location, a probability distribution as in [1, 2] for example, can be obtained. One can easily envision a process whereby polling requests are queued and issued in some sequence with more likely locations polled before less likely locations where possible.

In this work we consider simple planned ensemble polling strategies for the reduction of paging delay which are based on single unit optimal paging methods [1, 2]. We find that these methods permit larger page request rates than can be handled by blanket polling, even when a worst-case uniform unit location distribution is assumed.

Our approach is based on the concept of an *effective paging load* which depends upon the paging request rate, a measure of the average unit mobility and the paging discipline employed. As might be anticipated, lower mobility requires fewer places to be searched and thereby results in an overall lower effective paging load for a given page request rate. This work is presented in greater detail in [4].

2 Problem Statement

We assume page requests (incoming calls) arrive to the system according to a Poisson process with rate λ_p. Also assume each unit m to be located has some known location probability distribution[1] $p_j^{(m)}$ 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 feedback, s, s', is available to the scheduling process immediately after each polling request. A block diagram of the procedure is supplied in Figure 1.

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. Unfortunately, the number of possible states is infinite and optimal methods such as Markovian deci-

[1] This distribution can be conditional on whatever is known: the mobility process, the time of day, or the registration procedure favored by the mobile unit [5, 6, 7, 8].

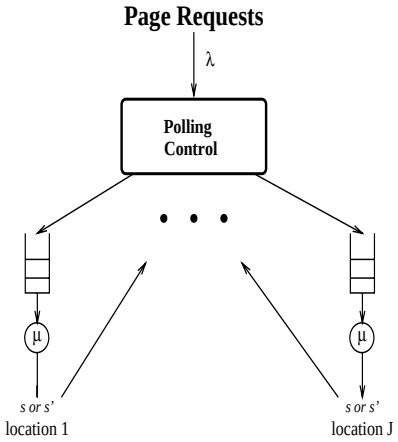

Figure 1: Block diagram of polling system. Page requests arrive according to a Poisson process with rate λ_p, each having an associated probability distribution p_j. Polling Control issues polling requests to appropriate locations. These requests are queued at the locations and serviced FCFS at average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

sion process (MDP) theory [9] cannot be applied. In this work, we turn to heuristic methods inspired by results for optimal paging of single units [1, 2].

3 Analysis and Results

We provide a basic queueing analysis of simple poll scheduling procedures. 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 examined through Monte Carlo simulation in [4] and found to agree well with the analytic results.

3.1 Blanket Polling

In most current systems, each page request generates simultaneous polling requests at all locations. If a unit is at location i, then the response time in paging 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. Assuming exponential service we have [10, 11],

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

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. This translates into increased polling channel bandwidth per cell which 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 hueristic approach to polling schedules based on previous analytic work and compare it to blanket polling

Specifically, in [1], it is verified 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.[2] As stated previously, we will treat the location queues as if they were independent and we 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

[2] This constraint can be relaxed to allow $E[\overline{L}_m] = \overline{L}$ [4].

algorithm is

$$\bar{D}'_1 = \bar{L}\bar{D}'(\rho\alpha) = \frac{\alpha J}{1-\rho\alpha} \qquad (2)$$

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, the sequential polling system always has a larger sustainable page request rates than for blanket polling.

In Figure 2 delay is plotted as a function of load ρ for blanket polling and sequential polling for $J = 20$ locations, for $\alpha = (J+1)/2J = 0.525$ (uniform distribution[3]) and $\alpha = 0.2$.

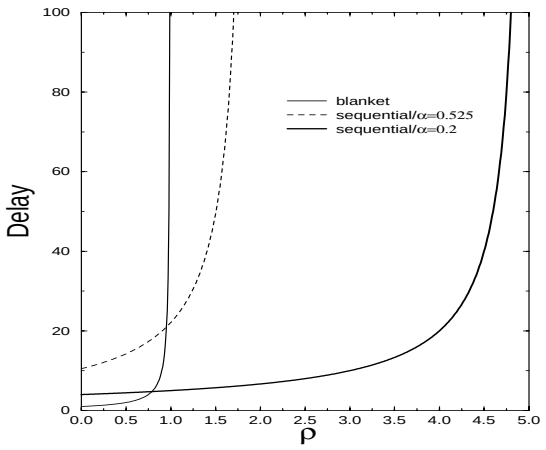

Figure 2: Delay as a function of offered load ρ for sequential and blanket polling. Twenty locations ($J = 20$), uniform ($\overline{L} = 10.5$, $\alpha = 0.525$), and nonuniform ($\alpha = 0.2$) location distributions.

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

$$\bar{g} = \sum_{i=1}^{G} i q_i \qquad (3)$$

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

$$\bar{L} = J q_G + k \sum_{i=1}^{G-1} i q_i = (J - kG) q_G + k\bar{g} \qquad (4)$$

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 when 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

$$\bar{D}'_k = (\bar{g} - 1)\hat{D}'_k + \bar{D}'(\rho\alpha) \qquad (5)$$

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 $\hat{D}'_k$. We obtain[5] for $\hat{D}'_k$,

$$\bar{D}'_k = \frac{1}{(1-\alpha\rho)} \left[(\bar{g} - 1) \sum_{\ell=1}^{k} \frac{1}{\ell} + 1 \right] \qquad (6)$$

Notice that for $k = 1$ we have $\bar{g} = \bar{L}$ and the result of equation (6) reduces to that in equation (2).

[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.

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

[5] Please refer to [4] for details.

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

$$q_i = \begin{cases} k/J & i = 1, 2, ...G-1 \\ 1-k(G-1)/J & i = G \end{cases} \quad (7)$$

so that

$$\bar{g} = \sum_{i=1}^{G-1} \frac{k}{J} i + G(1-k(G-1)/J) = G\left(1 - \frac{k(G-1)}{2J}\right) \quad (8)$$

and

$$\overline{L} = J - k(G-1) + \frac{k^2 G(G-1)}{2J} \quad (9)$$

so that

$$\alpha = \bar{L}/J = 1 - (G-1)\frac{k}{J}\left(1 - \frac{k}{2J}G\right) \quad (10)$$

Remembering that $G = \lceil J/k \rceil$ allows us to plot $\bar{D}'_k$ as a function of ρ for various k in Figure 3. 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 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 4.

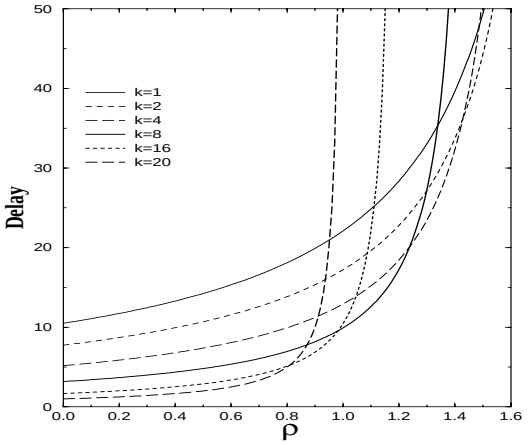

Figure 3: Plot of delay versus ρ for various group sizes k. Twenty locations ($J = 20$) with uniform location distribution ($\alpha = 0.525$).

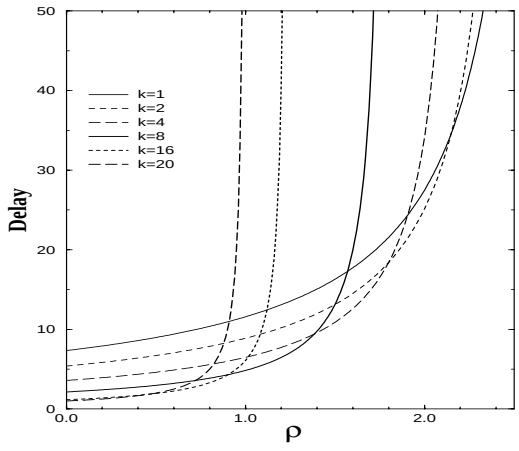

Figure 4: Plot of delay versus ρ for various group sizes k. Twenty locations ($J = 20$) with a linearly decreasing location distribution, $p_j = \frac{2}{J}(1 - \frac{j}{J+1})$. $E[j] = (J+2)/3$ so that $\alpha = 0.3\overline{666}$ for $J = 20$.

4 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 $k = 1$. These results suggest simple adaptive paging control strategies in response to variations in page request load.

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.

There is some need to extend these results to other service disciplines besides exponential, although in general, the analysis of queueing networks with non-exponential service is difficult. However, we venture to guess that results for other service distributions will prove similar. For example, simulation of deterministic service shows the general morphology of the delay curve family in k to be similar to that for exponential service. Limited experiments with other simple service distributions (such as uniform) show similar empirical results.

We have left open the question of *optimal* state-based service disciplines owing to difficulties of complexity suggested in the introduction. 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 complex algorithms to the simple scheme presented here in the near future.

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] C. Rose and R. Yates. Ensemble polling strategies for increased paging capacity in mobile communications networks. Winlab-TR 110, WINLAB, Rutgers University, 1995. (submitted to ACM Wireless Networks 9/95).

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

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

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

[8] U. Madhow, M.L. Honig, and K. Steiglitz. Optimization of Wireless Resources for Personal Communications Mobility Tracking. *IEEE Transactions on Networking*, 3(6):698–707, December 1995.

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

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

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