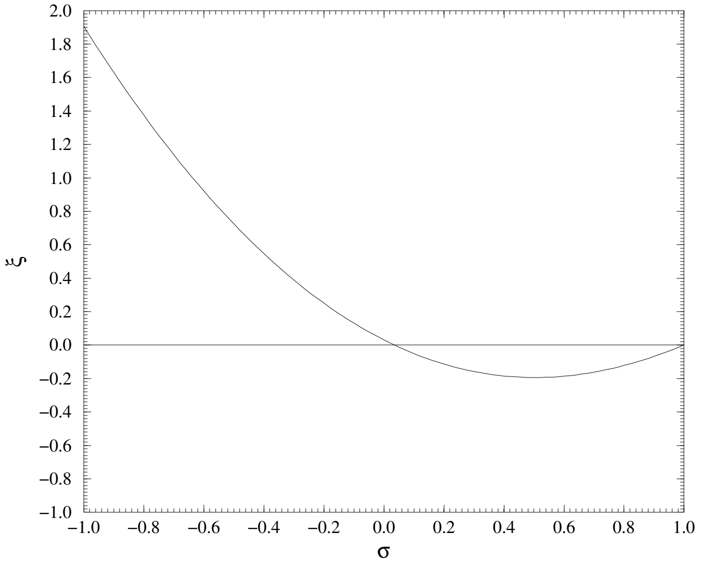

Figure 7: Graph of ξ (the root is small and positive)

We now show to determine the unknown probabilities p_0^0, p_0^1. Let σ denote a zero of ξ in [-1,1] since Q_0, Q_1 are analytic in this interval it follows that a root of the denominator must also be root of the numerator. We thus have that

$$A(\sigma) = B(\sigma) = 0. \tag{55}$$

However it is easy to see that if $\xi(\sigma) = 0$ then $A(\sigma) = 0$ and $B(\sigma) = 0$ imply the same condition on p_0^0, p_0^1. To see this note that the two conditions $A(\sigma) = B(\sigma) = 0$ require

$$\frac{p_0^0 F^{00}(\sigma) + p_0^1 F^{10} - p_0 \sigma F^{10}(\sigma)}{p_1^1 F^{11}(\sigma) + p_0^1 F^{11}(\sigma) - p_0 \sigma F^{11}(\sigma)} = \frac{F^{01}(\sigma)}{F^{00}(\sigma) - \sigma} = \frac{F^{11}(\sigma) - \sigma}{F^{10}(\sigma)}. \tag{56}$$

The equality of the two right hand terms is equivalent to $\xi(\sigma) = 0$, thus there is only one non-trivial condition arising from a zero in $\xi(\sigma)$, which is

$$(p_0^0 F^{00}(\sigma) + p_0^1 F^{10}(\sigma) - p_0 \sigma F^{10}(\sigma))(F^{11}(\sigma) - \sigma) = F^{10}(\sigma)(p_0^1 F^{11}(\sigma) + p_0^0 F^{01}(\sigma) - p_0 \sigma F^{11}(\sigma)) \tag{57}$$

We may also obtain an inhomogeneous linear equation as follows. Observe that the generating function for the number of customers left on departure is $Q(z) = Q_0(z) + Q_1(z)$ so that $Q(1) = 1$ since Q is a probability generating function. We know that ξ has a root at 1 so that

$$\lim z \to 1^- \quad Q(z) = \frac{A(z) + B(z)}{\xi(z)} = 1. \tag{58}$$

Applying L'Hopital's rule we see that

$$A'(1) + B'(1) = \xi'(1). \tag{59}$$

These two equations enable the unknown quantities p_0^0, p_0^1 to be determined and thus so are the generating functions given in equation 45. As explained earlier these can be numerically inverted to yield the tail of departure queue distribution which provides a bound on the tail of the sojourn time distribution. This forms the basis for rate determination.

We note that the left hand sides are well defined for complex $z, \mid z \mid \leq 1$, and are analytic in that region. It follows that any root of the determinant, $\xi(z)$ must also be a root of the numerators $A(z), B(z)$.

Let $d(z)$ be such a root. We now demonstrate that the determinant has a real root $z, z \in (-1, 1)$.

Define the matrix $\mathcal{F}(z)$ to be

$$\mathcal{F}(z) = \begin{pmatrix} F^{00}(z) & F^{10}(z) \\ F^{01}(z) & F^{11}(z) \end{pmatrix} \quad . \tag{46}$$

For $z \in [0, 1]$ this matrix has a Perron eigenvector lying in the interval $(0, 1]$. Note that the Perron eigenvector $\nu_1(1) = 1$ at $z = 1$ because $\mathcal{F}(1)$ is a probability matrix and hence $\xi(1)$ is 0. Observe that $\xi(z)$ is continuous and differentiable in [-1,1]. We will show that $\xi(-1) > 0$. If we then show that $\xi'(1) > 0$ it will follow that ξ has a root in (-1,1) since ξ is continuous. That this latter derivative is positive will follow from the stability criterion for the queue.

Observe that,

$$\xi(-1) = \mid \mathcal{F}(-1) + I \mid, \tag{47}$$

and also that $\mathcal{F}(-1) + I$ is a diagonally dominant (Hadamard) matrix. To see this note that

$$1 - \mid F^{00}(-1) \mid - \mid F^{01}(-1) \mid \geq 1 - F^{00}(1) - F^{01}(1) = 0 \tag{48}$$

with a similar inequality for the second row. The dominance is strict unless F^{00} has only odd powers and F^{01} has only even powers etc. However diagonally dominant matrices have nonnegative determinants and strictly diagonally dominant matrices have positive determinants and so $\xi(-1) \geq 0$ with strict inequality except in the special case mentioned above.

Let us denote $F^{00}(1) = F^{00}$ etc. and use a similar convention for derivatives. Differentiating $\xi(z)$ at 1 yields,

$$\xi'(1) = 2 - F^{00} - F^{11} - F^{00'}(1 - F^{11}) - F^{11'}(1 - F^{00}) - F^{01'}F^{10} - F^{10'}F^{01}. \tag{49}$$

The condition for stability is that the mean number of arrivals during a unit interval be smaller than 1, i.e.

$$q_0 F^{00'} + q_0 F^{01'} + q_1 F^{10'} + q_1 F^{11'} = \gamma < 1 \tag{50}$$

where q_i are the equilibrium probabilities that the source is in state i at the commencement of a service. They satisfy

$$\begin{aligned} q_0 F^{00} + q_1 F^{10} &= q_0 && (51) \\ q_0 F^{01} + q_1 F^{11} &= q_1 \\ q_0 + q_1 &= 1 && (52) \end{aligned}$$

from which we obtain that

$$\begin{aligned} q_0(2 - F^{00} - F^{11}) &= F^{10} = 1 - F^{11} && (53) \\ q_1(2 - F^{00} - F^{11}) &= F^{01} = 1 - F^{00}. \end{aligned}$$

Substituting we thus obtain that,

$$\xi'(1) = (2 - F^{00} - F^{11})(1 - \gamma) > 0 \tag{54}$$

It follows that the graph of $\xi(z)$ in [-1,1] must be as shown in figure.

We begin by writing down the forward recurrence equations for these probabilities. Using the theorem of total probabilities we have that,

$$\begin{aligned} p_j^0(n+1) &= \sum_{k=1}^{j+1} p_k^0(n) f_{j+1-k}^{00} + \sum_{k=1}^{j+1} p_k^1(n) f_{j+1-k}^{10} + p_0(n) f_j^{10} \\ p_j^1(n+1) &= \sum_{k=1}^{j+1} p_k^0(n) f_{j+1-k}^{01} + \sum_{k=1}^{j+1} p_k^1(n) f_{j+1-k}^{11} + p_0(n) f_j^{11}. \end{aligned} \tag{38}$$

Let us define the transform of $p_j^i(n), i = 0, 1$ as

$$Q_n^i(z) = \sum_{j=0}^{\infty} p_j^i(n) z^j, \quad i = 0, 1. \tag{39}$$

We may thus take the transform of the left hand side of (38) to obtain that,

$$\begin{aligned} Q_{n+1}^0(z) &= \sum_{j=0}^{\infty} \sum_{k=1}^{j+1} p_k^0(n) f_{j+1-k}^{00} z^j + \sum_{j=0}^{\infty} \sum_{k=1}^{j+1} p_k^1(n) f_{j+1-k}^{10} z^j + \sum_{j=0}^{\infty} p_0(n) f_j^{10} z^j \\ &= z^{-1} \sum_{k=1}^{\infty} z^k p_k^0(n) \sum_{j=k-1}^{\infty} f_{j+1-k}^{00} z^{j+1-k} + z^{-1} \sum_{k=1}^{\infty} p_k^1(n) z^k \sum_{j=k-1}^{\infty} f_{j+1-k}^{11} z^{j+1-k} + p_0(n) \sum_{j=0}^{\infty} f_j^{10} z^j \\ &= z^{-1} F^{00}(z) \left[Q_n^0(z) - p_0^0(n) \right] + z^{-1} F^{10}(z) \left[Q_n^1(z) - p_0^1(n) \right] + p_0(n) F^{10}(z) \end{aligned} \tag{40}$$

Repeating this calculation, we also obtain that

$$Q_{n+1}^1(z) = z^{-1} F^{01}(z) \left[Q_n^0(z) - p_0^0(n) \right] + z^{-1} F^{11}(z) \left[Q_n^1(z) - p_0^1(n) \right] + p_0(n) F^{11}(z) \tag{41}$$

For the moment we assume that a stationary distribution exists, $Q^i(z), i = 0, 1$. In which case the following must be satisfied,

$$Q^0(z) \left(F^{00}(z) - z \right) + Q^1(z) F^{10}(z) = p_0^0 F^{00}(z) + p_0^1 F^{10} - p_0 z F^{10}(z) \tag{42}$$

$$Q^0(z) F^{01}(z) + Q^1(z) \left(F^{11}(z) - z \right) = p_0^0 F^{01}(z) + p_0^1 F^{11}(z) - p_0 z F^{11}(z) \tag{43}$$

The above equations are linear in Q and can be solved using Cramer's rule,

$$\begin{aligned} Q^0(z) &= \frac{A(z)}{\xi(z)} \\ Q^1(z) &= \frac{B(z)}{\xi(z)} \end{aligned} \tag{44}$$

where ξ, A, B are

$$\begin{aligned} \xi(z) &= \begin{vmatrix} F^{00}(z) - z & F^{10}(z) \\ F^{01}(z) & F^{11}(z) - z \end{vmatrix} \\ A(z) &= \begin{vmatrix} p_0^0 F^{00}(z) + p_0^1 F^{10} - p_0 z F^{10}(z) & F^{10}(z) \\ p_0^1 F^{11}(z) + p_0^0 F^{01}(z) - p_0 z F^{11}(z) & F^{11}(z) - z \end{vmatrix} \\ B(z) &= \begin{vmatrix} F^{00}(z) - z & p_0^0 F^{00}(z) + p_0^1 F^{10} - p_0 z F^{10}(z) \\ F^{01}(z) & p_0^1 F^{11}(z) + p_0^0 F^{01}(z) - p_0 z F^{11}(z) \end{vmatrix}, \end{aligned} \tag{45}$$

If the source is on the entire period we then obtain,

$$g^1_{11}(y)dy = e^{-\alpha}\delta(y-1)dy. \tag{33}$$

Summing over r we obtain that

$$g_{11}(y)dy = \alpha\beta e^{(\beta-\alpha)y}e^{-\beta}\sqrt{\frac{y}{\alpha\beta(1-y)}}I_1(2\sqrt{\alpha\beta y(1-y)})dy + e^{-\alpha}\delta(y-1)dy. \tag{34}$$

The final case is when the source is in the off state at the start and end of service. Let us suppose there are $r \geq 1$ active periods.

$$\begin{aligned} g^r_{00}(y)dy &= \frac{\alpha^r y^{r-1}}{(r-1)!}e^{-\alpha y} \times \int_{v=0}^{1-y} \frac{\beta^r v^{r-1}}{(r-1)!}e^{-\beta v}e^{-\beta(1-y-v)}dvdy \\ &= \alpha^r\beta^r \frac{y^{r-1}(1-y)^r}{(r-1)!r!}e^{(\beta-\alpha)y-\beta}dy \end{aligned} \tag{35}$$

If there is no active period we obtain

$$g^0_{00}(y)dy = e^{-\beta}\delta(y)dy. \tag{36}$$

Summing over r we obtain,

$$g_{00}(y)dy = \alpha\beta e^{(\beta-\alpha)y}e^{-\beta}\sqrt{\frac{(1-y)}{\alpha\beta y}}I_1(2\sqrt{\alpha\beta y(1-y)})dy + e^{-\beta}\delta(y)dy. \tag{37}$$

From these 4 densities we may obtain the generating functions for the number of arrivals which take place during a service period,

$$\begin{aligned} F_{01}(z) &= \int_0^1 g_{01}(y)e^{-y\lambda(1-z)}dy \\ &= \int_0^1 \beta e^{(\beta-\alpha)y}e^{-\beta}I_0(2\sqrt{\alpha\beta y(1-y)})e^{-y\lambda(1-z)}dy, \\ F_{10}(z) &= \int_0^1 g_{10}(y)e^{-y\lambda(1-z)}dy \\ &= \int_0^1 \alpha e^{-\beta}e^{(\beta-\alpha)y}I_0(2\sqrt{\alpha\beta y(1-y)})e^{-y\lambda(1-z)}dy, \\ F_{11}(z) &= \int_0^1 g_{11}(y)e^{-y\lambda(1-z)}dy, \\ &= \int_0^1 \alpha\beta e^{(\beta-\alpha)y}e^{-\beta}\sqrt{\frac{2y}{\alpha\beta(1-y)}}I_1(2\sqrt{\alpha\beta y(1-y)})e^{-y\lambda(1-z)}dy + e^{-\alpha-\lambda(1-z)}, \\ F_{00}(z) &= \int_0^1 g_{00}(y)e^{-y\lambda(1-z)}dy \\ &= \alpha\beta\int_0^1 e^{(\beta-\alpha)y}e^{-\beta}\sqrt{\frac{2(1-y)}{\alpha\beta y}}I_1(2\sqrt{\alpha\beta y(1-y)})e^{-y\lambda(1-z)}dy + e^{-\beta} \end{aligned}$$

C.2 Stationary Probabilities for the Number in Queue on Departure

Let $p^0_j(n)$ be the probability that the nth customer departs to leave j customers in queue and the source in the off state. Also let $p^1_j(n)$ be similarly defined. We are interested in the limiting stationary values of these probabilities.

The sum of r on periods is gamma with r degrees of freedom with scale parameter α. There are $r-1$ corresponding off periods with cumulative time V and a final rth off period with duration Z such that $Y+V<1$ and $Y+V+Z\geq 1$. We wish to derive the (improper) density of Y. The probability that $y\leq Y\leq y+dy, y<1$ is

$$\begin{aligned} g_{10}^r(y)dy &= \frac{\alpha^r y^{r-1}}{(r-1)!}e^{-\alpha y}\times\int_{v=0}^{1-y}\frac{\beta^{r-1}v^{r-2}}{(r-2)!}e^{-\beta y}e^{-\beta(1-y-v)}dvdy \\ &= \alpha^r\beta^{r-1}\frac{y^{r-1}(1-y)^{r-1}}{(r-1)!^2}e^{(\beta-\alpha)y-\beta}dy \end{aligned} \tag{24}$$

this holds for $r\geq 2$ active periods. If there is just one, we obtain

$$g_{10}^1(y)dy = \alpha e^{-\alpha y}e^{-\beta(1-y)}. \tag{25}$$

If we sum over r we obtain the overall density as

$$g_{10}(y) = \alpha e^{-\beta}e^{(\beta-\alpha)y}I_0(2\sqrt{\alpha\beta y(1-y)}). \tag{26}$$

We now examine the case where the source is in the off state at the start of service and in the on state at the end with r periods in which the source is off. Thus the cumulative off time has β density with r degrees of freedom. The total active time is therefore $1-V$. Thus we may write the probability that $v\leq V<v+dv, v<1$ as

$$\begin{aligned} g_{01}^r(v)dv &= \frac{\beta^r v^{r-1}}{(r-1)!}e^{-\beta v}\times\int_{w=0}^{1-v}\frac{\alpha^{r-1}w^{r-2}}{(r-2)!}e^{-\alpha w}e^{-\alpha(1-v-w)}dwdv \\ &= \beta^r\alpha^{r-1}\frac{v^{r-1}(1-v)^{r-1}}{(r-1)!^2}e^{(\alpha-\beta)v-\alpha}dv \end{aligned} \tag{27}$$

this holds for $r\geq 2$ active periods. We may now substitute $y=1-v$ to obtain the corresponding density for Y as

$$g_{01}^r(y)dy = \beta^r\alpha^{r-1}\frac{y^{r-1}(1-y)^{r-1}}{(r-1)!^2}e^{(\beta-\alpha)y-\beta}dy \tag{28}$$

If there is just one, we obtain

$$g_{01}^1 = \beta e^{(\beta-\alpha)y-\beta}. \tag{29}$$

Summing over r we obtain that

$$g_{01}(y)dy = \beta e^{(\beta-\alpha)y}e^{-\beta}I_0(2\sqrt{\alpha\beta y(1-y)}) \tag{30}$$

We now examine the case where the source is on at both the start and the end of service. Let us suppose there are $r\geq 1$ off periods and the same number of on periods with an additional final on period.

$$\begin{aligned} g_{11}^r(v)dv &= \frac{\beta^r v^{r-1}}{(r-1)!}e^{-\beta v}\times\int_{w=0}^{1-v}\frac{\alpha^r w^{r-1}}{(r-1)!}e^{-\alpha w}e^{-\alpha(1-v-w)}dwdv \\ &= \beta^r\alpha^r\frac{v^r(1-v)^r}{(r-1)!r!}e^{(\alpha-\beta)v-\alpha}dv \end{aligned} \tag{31}$$

As before we substitute $y=1-v$ to obtain,

$$g_{11}^r(y)dy = \beta^r\alpha^r\frac{(1-y)^{r-1}y^r}{(r-1)!r!}e^{(\beta-\alpha)y-\beta}dy. \tag{32}$$

B.1 Iterative Improvement

We now discuss ways to iteratively improve the solution ϕ_i^* we using the following steps. Hereafter, we use the refined Chernoff approximation for all probabilities.

For later convenience, we define the following functions denoting the ratios of the achieved probability exponent to the desired one.

$$G_i(\phi) = \frac{1}{\delta_i}\left(s_i^*\phi_i\left(\frac{P}{C_iR_i} - 1\right) - \sum_{j\neq i} F_j(\phi_j s_i^*) + \frac{1}{2}\ln[2\pi \sum_{j\neq i}(s_i^*\phi_j)^2 F_j''(\phi_j s_i^*)]\right). \tag{23}$$

where, as before, s_i^* satisfy equation 22 with the ϕ^* replaced by ϕ.

Algorithm ITER

- Begin with $\phi^{(0)} = \phi^*$ computed as described above.
- Determine the index k corresponding to the best served user, i.e,

$$k = \arg\max_i G_i(\phi^{(n)}).$$

- Update ϕ as $\phi^{(n+1)} = \phi^{(n)} + t^*\mathbf{e}_k$, where $\mathbf{e}_k$ is the basis vector with 1 in the kth entry and 0 in all others, and t^* is defined as

$$t^* = \arg\max_t \min_i G_i(\phi^{(n)} + t\mathbf{e}_k).$$

- Stop when all the G_i are equal. Feasibility is established if this value exceeds unity.

It can be shown that the G_i are increasing functions of ϕ_i and decreasing functions of ϕ_j, $j \neq i$. As a result, the above algorithm is guaranteed to increase the value of the smallest G_i at each stage, and involves a series of one-dimensional optimizations that are guaranteed to converge when all the G_i are equal.

C Queueing Analysis for Determining the Rate Allocation

C.1 Arrivals During a Service

We will be constructing the Markov chain of customers left behind in queue following each departure. This process is not Markov since the total number of arrivals depends on whether the state of the source is on or off at the commencement of service. If the on/off state following each departure is included then the process is Markov. We will thus be interested in the number of arrivals during a service in which the state of the source at the beginning and the end of the service period are given.

We will do one case in detail, the derivation of the other three cases being similar. Suppose that the state of the source is on at the start of service and that the state at the end is off. Suppose further that there are $r \geq 1$ active periods. We are interested in the probability density of Y the total amount of on time and that the sources are in the given states at the start and end of the period.

$$A = \max_{\phi} \min_{i} \max_{s_i} \frac{s_i \phi_i \left(\frac{P}{C_i R_i} - 1 \right) + F_i(\phi_i s_i) - \delta_i}{\sum_j F_j(\phi_j s_i)} \tag{16}$$

which then guarantees feasibility if $A \geq 1$, with weights set to the solution values ϕ_i^*. Note that this is only a sufficient condition, and not necessary, on account of the Chernoff bound which is typically slack in the absence of sufficient aggregation.

However, equation 16 poses a difficult problem for any realistic real-time applications, and is likely to require sophisticated optimization techniques. In some previous work, we examined a simpler problem by eliminating the Chernoff parameters by setting them to $s_i = 1$ for all i. The resulting solution is further conservative, since we have eliminated the step of maximizing over s_i. However, this yields the following more tractable problem.

$$B = \max_{\phi} \min_{i} \frac{\phi_i \left(\frac{P}{C_i R_i} - 1 \right) + F_i(\phi_i) - \delta_i}{\sum_j F_j(\phi_j)} \tag{17}$$

It can be shown that the functions

$$G_i = \frac{\phi_i \left(\frac{P}{C_i R_i} - 1 \right) + F_i(\phi_i) - \delta_i}{\sum_j F_j(\phi_j)} \tag{18}$$

are increasing functions of ϕ_i and decreasing functions of $\phi_j \ : \ j \neq i$. This implies that the solution that maximizes B satisfies $G_i = \Delta$ (a user independent constant) for all i. To obtain the optimum, we then maximize over this single parameter Δ. In summary, problem 16 is equivalent to

$$B = \max_{\Delta} \frac{\Delta}{\sum_j F_j[\phi_j(\Delta)]} \tag{19}$$

where the functions $\phi_j(\Delta)$ are obtained by inverting

$$\phi_j \left(\frac{P}{C_j R_j} - 1 \right) + F_j(\phi_j) - \delta_i = \Delta. \tag{20}$$

For a small number of users, one may efficiently solve for B using bisection over Δ, while solving for the $\phi_j(\Delta)$ by bisection at each stage. As before, if the solution B^* exceeds unity, we obtain a sufficient condition for feasibility with the corresponding ϕ_i^*. However, this condition is usually conservative due to the looseness of the regular Chernoff bound. We hence recommend, at this stage, to compute the rate violation probabilities using the refined Chernoff bound to perform admission control, i.e, verify the validity of the following condition for each user i.

$$s_i^* \phi_i^* \left(\frac{P}{C_i R_i} - 1 \right) - \sum_{j \neq i} F_j(\phi_j^* s_i^*) + \frac{1}{2} \ln[2\pi \sum_{j \neq i} (s_i^* \phi_j^*)^2 F_j''(\phi_j^* s_i^*)] \geq \delta_i \tag{21}$$

where s_i^* satisfies

$$\sum_{j \neq i} \phi_j^* F_j'(\phi_j^* s_i^*) = \phi_i^* \left(\frac{P}{C_i R_i} - 1 \right). \tag{22}$$

(a) (b)

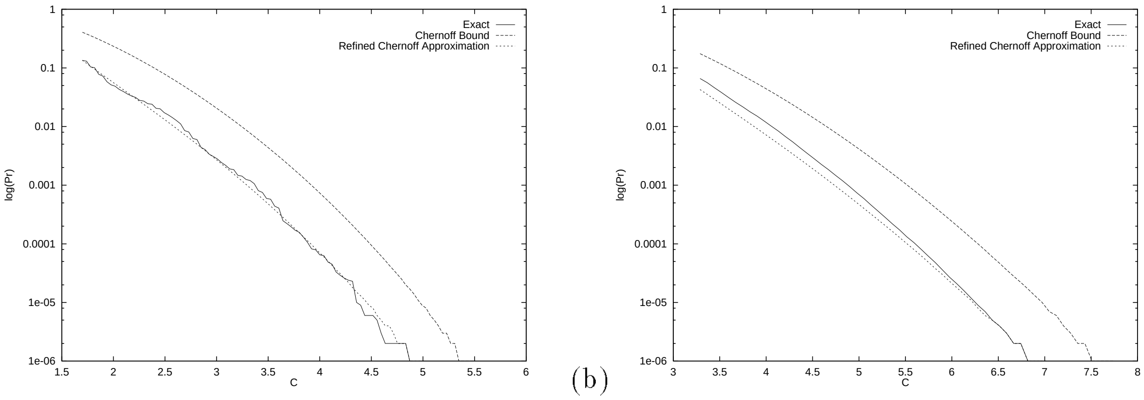

Figure 6: $Pr[\sum_{i=1}^{n} w_i x_i > C]$ as a function of C for binary $x_i \in \{0,1\}$ and fixed $w_i \in [0,1]$. $x_i = 1$ with probability 0.1 for all i. (a) n=10. (b) n=20.

This expression may be refined, under the assumptions that the probabilities of interest are small and x is an aggregate of a sufficiently large number of random variables, to obtained the Refined Chernoff approximation

$$-\ln Pr[x > C] \geq Cs^* - F(s^*) + \ln[\sqrt{2\pi} s^* \sigma(s^*)] \tag{12}$$

where $\sigma^2(s^*) = \left.\frac{d^2 F(s)}{ds^2}\right|_{s=s^*} \equiv F''(s^*)$ and

$$s^* = \arg\max_s Cs - F(s),$$

i.e. s^* satisfies $C = F'(s^*)$. For illustration, figure 6 shows a comparison of exact tail probabilities with these approximations for the cases of weighted combinations of binary variables. Notice that the refined Chernoff approximation provides an accurate estimate even with as few as 10 variables, while the Chernoff bound differs only by a constant factor.

We now apply these approximations to compute rate violation probabilities and set the weights. Recall equation 7 that expresses the conditions to be satisfied by the GPS weights ϕ_i.

$$Pr\left[R_i < R_i\right] = Pr\left[\frac{\phi_i}{\phi_i + \sum_{j \neq i} \phi_j \xi_j} \frac{P}{C_i} < R_i\right] \leq e^{-\delta_i}. \tag{13}$$

We use the Chernoff bound on equation 13 to write the following sufficient condition after some algebra.

$$s_i \phi_i \left(\frac{P}{C_i R_i} - 1\right) - \sum_{j \neq i} F_j(\phi_j s_i) \geq \delta_i \tag{14}$$

where $s_i \geq 0$ are the so-called Chernoff parameters, and $F_j(s) = \ln E[e^{s\xi_j}] = \ln[1 - p_j + p_j e^s]$ is the log-moment-generating-functions of ξ_j, which is 1 with probability p_j. Equation 14 can be rewritten as

$$\frac{s_i \phi_i \left(\frac{P}{C_i R_i} - 1\right) + F_i(\phi_i s_i) - \delta_i}{\sum_j F_j(\phi_j s_i)} \geq 1 \tag{15}$$

for each user i. This requires us to solve the following optimization problem to obtain

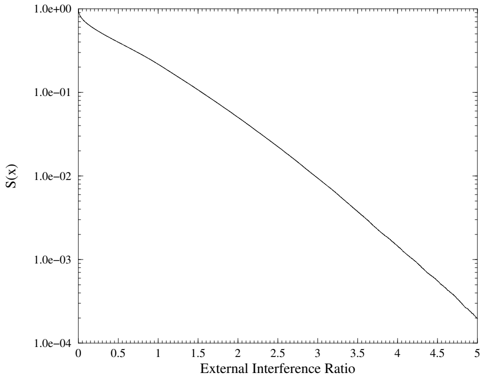

Figure 5: The survivor function for external cell interference, $S(x)$.

A Calculation of c_i values

We suppose that all base-stations transmit at maximum power at all times. When this maximum power is 1, let I_{ext} be a random variable equal to the relative out-of-cell interference experienced by a user, distributed over all positions in the cell and over log-normal fading. The relative out-of-cell interference is defined to be the out-of-cell interference divided by the received power from the in-cell base-station. The *survivor function* $S(x) := \Pr(I_{ext} \geq x)$ is given in Figure 5. For details on how this function may be calculated, see Viterbi [6]. We in fact assume that each base-station has maximum transmit power 2, but only half of this power is dedicated to data. For simplicity, we also take the in-cell interference to be equal to the in-cell received power. Let the target E_b/I_0 be 7dB, and let the bandwidth of the system be 4MHz. Then, using the formula

$$\frac{E_b}{I_0} = \frac{\text{Bandwidth}}{\text{Transmit Rate}} \text{Signal-Noise Ratio},$$

we have

$$c_i = 2 \times \frac{10^{0.7}}{4 \times 10^6}(1 + I_{ext}).$$

B Setting the weights using the Chernoff Bound

The famous Chernoff bound(see for example [3]) for the tail probability of a random variable x is expressed as

$$-\ln Pr[x > C] \geq Cs - F(s) \tag{10}$$

for $s \geq 0$, where $F(s)$ is the log-moment-generating-function of the random variable x, defined as

$$F(s) = \ln \mathcal{E}[e^{sx}]$$

with $\mathcal{E}[]$ denoting expected value. In particular the tightest value of this bound is obtained as

$$-\ln Pr[x > C] \geq \max_s \; Cs - F(s). \tag{11}$$

10) this probability may be computed exactly. These calculatations form the basis for admission control at the time when a new user applies for admission.

As our numerical results showed the choice of admission criterion has a large impact on capacity and therefore how economic any 3rd generation CDMA data service will be. The number of users that could be supported in a cell could be increased by a factor of between 2 or 3 depending on the extent access to users with high effective weights (typically near the edge of a cell) was granted.

The various other aspects of the scheme have yet to be examined in detail. In particular we have yet to describe how the effective weights for each mobile are to be computed. These are critical in placing constraints on the probability of overload as a result of changes in propagation and interference as well as through chance fluctuations of power control. An additional area for investigation is the question of how resources are to be allocated at a network level (as opposed to the cell level description given here).

References

[1] J. Abate and W. Whitt (1992) "The Fourier Series Method For Inverting Transforms of Probability Distributions" *Queueing Systems 10 pp 5-88*

[2] D. M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar and P. Whiting. (November 1998) CDMA Data Forward Link Rate Scheduling. Bell Labs Internal Report. April'99.

[3] J. A. Bucklew. *Large Deviation Techniques in Decision, Simulation and Estimation.* Wiley-Interscience, New York, 1990.

[4] Kleinrock L. (1975) Queueing Systems,Volume 1, *Wiley.*

[5] Kelly F.P. (1979) Reversibility and Stochastic Networks *Wiley.*

[6] Viterbi A.J. (1995) CDMA. Principles of Spread Spectrum Communication *Addison-Wesley.*

[7] Lucatoni D. (1992) The BMAP/G/1 Queue: A Tutorial *Performance 1993.*

[8] A. Elwalid and D. Mitra. Design of Generalized Processor Sharing Schedulers which Statistically Multiplex Heterogeneous QoS Classes. *Proc. of IEEE INFOCOM'99* , March 1999, pp. 1220-1230.

[9] A. Elwalid, D. Mitra, and R. H. Wentworth. A New Approach for Allocating Buffers and Bandwidth to Heterogenous, Regulated Traffic in an ATM Node. *IEEE Journal on Selected Areas in Communications*, 13(6):1115–1127, Aug. 1995.

[10] S. Keshav. *An Engineering Approach to Computer Networking.* Addison Wesley, Reading, MA, 1997.

[11] G. Kesidis. *ATM Network Performance.* Kluwer, Boston, MA, 1996.

[12] D. Stiliadis and A. Varma. Efficient Fair Queueing Algorithms for Packet Switched Networks. *IEEE/ACM Trans. Networking*, 6:175–185, Apr. 1998.

[13] D. Stiliadis and A. Varma. Rate Proportional Servers: A Design Methodology for Fair Queueing Algorithms. *IEEE/ACM Trans. Networking*, 6:164–174, Apr. 1998.

requirements as described before, i.e. $\phi_i = c_i R_i$, as we consider $\delta_i = \delta = 0.01$ uniformly for all users. In this context, it is worthwhile to mention that our attempt to refine the weights using the procedure given in Appendix B produced negligible benefits in all our examples, thus reinforcing the fact that the simple choice $\phi_i = c_i R_i$ is very effective.

Figure 3 shows the results for capacity supposing that mobiles are taken uniformly from the cell area.

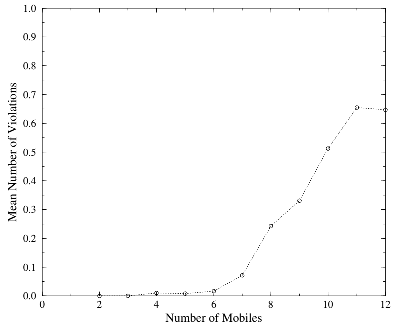

Figure 3: Capacity when selecting the users uniformly from the cell

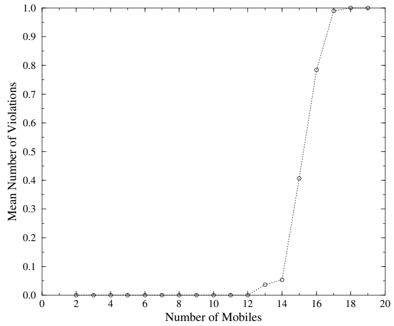

Figure 4: Capacity when selecting the N users with lowest c_i

As was to be expected the results show that accepting users with relatively high c_i leads to a reduction in capacity. A much higher number of users can be supported therefore if we offer a restricted service to users in "worst case" positions. For example if we restrict rate according to the effective weights then we are effectively allocating a certain fraction of forward link power whenever there is data to send to the mobile.

5 Conclusions

Rate processor sharing provides an effective means to allocate resources on the forward link of a CDMA cell. In particular it enables admission control to be performed between users with heterogenous QOS requirements. This is achieved by transforming these requirements into a minimum rate which is supplied with high probability. This probability can be computed accurately and efficiently by using a refined Chernoff approximation. In the case of a small number of users (¡

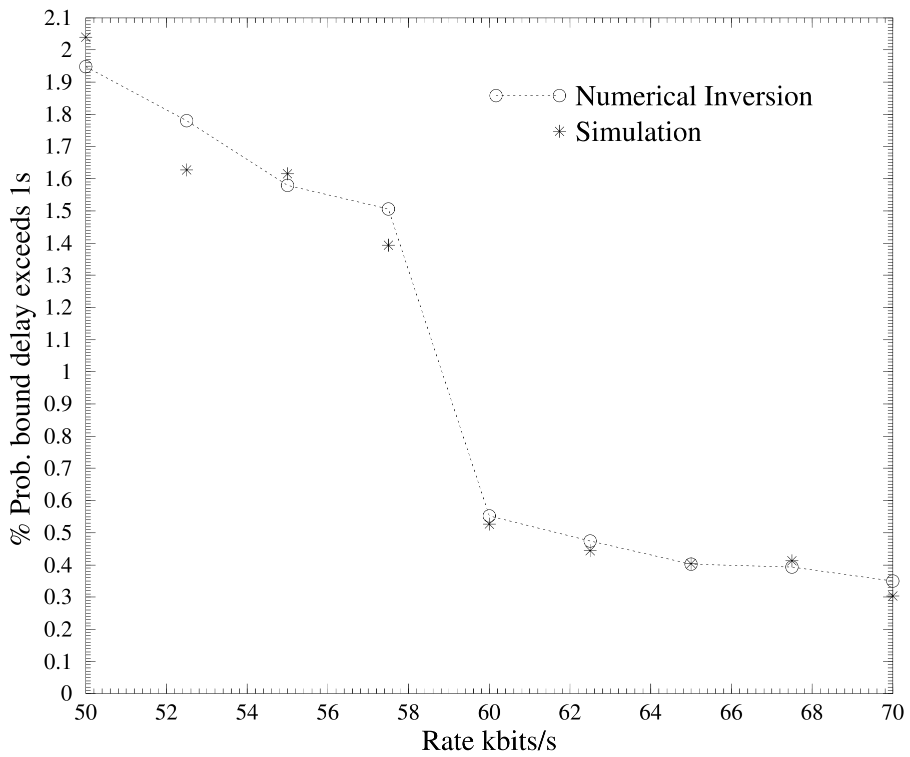

Figure 2: Probability bound for delay

to the true capacity. However, we expect RPS to be efficient and hence provide a good indication of the true capacity of the system. For a more detailed discussion of capacity and for the following definitions we refer the reader to [2].

For simplicity we assume that the data traffic model for all users is the same. Let ε and f be some fixed (small) numbers in $[0, 1]$; the meaning of $1 - \varepsilon$ is the cell "Grade-of-Service".
1. The maximum value of N such that such that the *average* fraction of users for which QoS is achieved is at least $1 - \varepsilon$.
2. The maximum number $(1 - f)M$ out of M randomly distributed users in a cell such that the *average* fraction of users achieving QoS is at least $1 - \varepsilon$, (when the fM "worst" users are *not* served at all).
Note that definition 2 is a generalization of 1 with f being non-zero. Note also that the above definitions do not assume that the QoS requirements of all users are the same. They can be easily extended to the case when the traffic models are different for different users, by replacing the maximum number of users by the maximum throughput achieved. In the case of common mean data rate the two differ by a factor of R.

The ability to meet QoS requirements of a set of users clearly depends not only on the number of users, but also on the distribution of the users within the cell, since users with higher c_i values require more resources to meet their QoS. Thus it is clear that definition 2 will yield higher values of capacity (for any given scheduling algorithm) than 1. This is because the distribution of $N = (1 - f)M$ users with the smallest c_i values chosen out of M randomly placed users is never worse than the distribution of N users placed randomly in the cell. Roughly speaking, the factor "f" reflects the fraction of mobiles that are blocked during admission control.

For all our numerical results, we have set the user weights to be proportional to their power

Number of Customers n	Analysis	Simulation
0	0.491459	0.491737
1	0.222269	0.222510
2	0.092868	0.093123
3	0.036515	0.036679
4	0.011497	0.013996
5	0.005233	0.005275
6	0.001963	0.001983
7	0.000738	0.000741
8	0.000277	0.000280
9	0.000104	0.000105
10	0.000039	0.000040

Table 1: Survivor Distribution for the Number of Customers Left Behind in Queue n

4 Numerical Results

In this example we consider groups of mobiles with independent statistically identical sources and a common service requirement. The source is an on-off bursty source with on and off periods being mutually independent, exponentially distributed with mean on time 93 ms. and mean off time 907 ms. . The rate of generation of packets when a source is on is 9 pkts/s so that,

$$MeanPacketArrivalRate = 0.093 \times 9 = 0.837 pkts/s. \tag{9}$$

The packet sizes are taken to be constant with length 10kbits. The common QoS requirement is that packets be transmitted with a delay of greater than 1s with a probability of 0.01 or better.

As described before, we may treat each source *in isolation* from others and seek to determine the bit rate R sufficient to meet the delay requirement mentioned above. The queueing analysis in Appendix C, which exploits the fact that the arrival and departure distributions coincide in equilibrium (see for example [4]), allows us to bound the tail probability for delay beyond 1 second for any given rate. Table 1 gives the survivor distribution for the number of customers left behind in queue with results obtained from inversion of the generating function (column 2) and from simulation (column 3) in which a total of 10^7 events were used to produce the results. They were obtained using the above parameters and a service rate of $R = 60$ kbits/s. The centre column was obtained from the generating function using inversion techniques outlined in [1]. The third column shows the results which were obtained from simulation which show close agreement.

Figure 2 shows a graph of the probability of delay exceeding 1s for bit rates in the range of 50-75 kbits/s, using departure results such as those described above. It is well known that this distribution coincides with the number an arrival finds in queue. We may thus bound the tail probability of transmission delay by assuming that the transmission of the first packet in queue at the time of an arrival has only just been initiated. The large drop in probability around 60 kbits/s corresponds to the fact that the bound requires that there be no more than 4 packets in queue for rates $R < 60$ kbits/s whereas a delay of 1s can always be achieved for rates $R \geq 60$ kbits. As the graphs shows the minimum rate which is required to meet the delay requirement is ≈ 60 kbits/s.

4.1 Capacity Calculations

We now proceed to the discussion and estimation of capacity for our CDMA cell. Clearly, the realized capacity is a function of the algorithm, in our case RPS, and only provides a lower bound

For the choice $\phi_i = c_i R_i$ we have $\mathcal{R}_i = \mathcal{R} = Pr\left[\sum_j x_j c_j R_j > P\right]$ for all i. If we also assume that no one user dominates the power, i.e. $\frac{P}{c_i R_i} >> 1$ for all i, and $a_i >> e^{-\delta}$, e.g. a_i is of order 10^{-1} while $e^{-\delta}$ is of order 10^{-3} or less, it follows that the $\mathcal{P}_i$, when feasible, are roughly "equal" in a logarithmic sense. This means that they differ from each other by a factor of atmost $\frac{1}{\min_i a_i}$, or their logarithms differ by atmost $\log(\min_i a_i)$, which we take to be much smaller than $\log(\mathcal{P}_i)$. In this case, the optimal choice satisfies $\phi_i = c_i R_i + \epsilon_i$ where the ϵ_i are incremental. This is due to the fact that the tails of all our distributions are nearly exponential (when the number of users is appreciably larger than 1) and hence the $\mathcal{P}_i$, which are off by no more than an order of magnitude when a_i are of order 10^{-1}, are highly sensitive functions of ϕ_j of the form $e^{-(\cdot)\phi_j}$. Also, when $\frac{P}{c_i R_i} >> 1$ for all i, any choice of ϕ_i differing substantially from $c_i R_i$ cannot equalize the $\mathcal{P}_i$ since the $\mathcal{R}_i$ would all differ substantially in this case. It hence follows that the solution in the neighborhood of $c_i R_i$ is the unique optimum.

Thus, we see that when the users have identical rate violation probabilities, the users weights simply take the form of their minimum power requirement $\phi_i = P_i = c_i R_i$. One may then achieve heterogeneity in service by setting the R_i aggressively or conservatively. However, if different δ_i are desirable for some applications, the appendix (section B) shows how to use the Chernoff approximation on equation 7 to compute the ϕ_i in this case. The iterative improvement described there may also be used to refine the choice $\phi_i = c_i R_i$ when the δ_i are all equal. However, these calculations are more computationally involved and less attractive for implementation.

3.3 Connection Admission Control (CAC) for RPS

We now discuss the criteria that must be used in conjunction with the above parameter settings to decide whether the QoS of an arriving user can be met. In the general case, given the weights ϕ_i, we may use the refined Chernoff approximation described in Appendix B (or an exact computation if the number of users is much smaller than 10) to compute the rate violation probabilities $\mathcal{P}_i$ defined above, and admit an incoming user if all the $\mathcal{P}_i$ are acceptable.

In the special case when the a_i and δ_i are all equal, it can be shown that the most constrained user would be the one with the largest cost, say user k with cost c_k. In this case, we only need to check one condition, i.e. that $\mathcal{P}_k$ is acceptable. With the simple choice of weights $\phi_i = c_i R_i$, a_i different but all of the same order of magnitude (say 10^{-1}), and $\delta_i = \delta$ much larger than $-\ln a_i$, we can use $\frac{\mathcal{R}}{\min_i a_i}$ as a conservative estimate of the worst $\mathcal{P}_i$. In this case, the admission criterion just simplifies to

$$-\ln \mathcal{R} \leq \gamma$$

where $\gamma = \delta - \ln[\min_i a_i]$. If any of these assumptions are valid, we see that the admission control can be boiled down to checking a single condition using the refined Chernoff approximation.

In the general case, when a large number of heterogeneous users are present, RPS requires us to check global feasibility. If this is inconvenient, one may consider the alternative approach of classifying users based on their QoS requirements, traffic characteristics and cost of transmission c_i. Such classification would allow the pre-computation and storage of an "admissible region" in this multi-dimensional user space using RPS, which may then be used to make fast decisions in real-time. One may simplify this further to avoid storing such information by approximating this region by a linear one which associates an "effective power" with each user. These approaches, which are standard in conventional networking, are likely to be more useful as the number and types of data users becomes large.

3.2 Weight Determination

We now describe our method to set the GPS weights ϕ_i. We rewrite the GPS equation 3 as

$$P_i = \frac{\phi_i}{\sum_j x_j \phi_j} P. \tag{6}$$

where x_j is an instantaneous binary activity indicator random variable of user j that determines an instantaneous power assignment for all active users. We assume the users to be independent and non-colluding, i.e. the x_j are i.i.d. To impose the QoS criterion described before, we must pick ϕ_i so as to satisfy

$$Pr\left[\frac{\phi_i}{\sum_j x_j \phi_j}\frac{P}{c_i} \leq R_i\right] \leq e^{-\delta_i} \tag{7}$$

or equivalently

$$Pr\left[\sum_{j\neq i} x_j \phi_j > (\frac{P}{c_i R_i} - 1)\phi_i\right] \leq e^{-\delta_i} \tag{8}$$

We observe that when the $\delta_j = \delta$ are all equal and much larger than unity, and the activity factors $a_i >> e^{-\delta_i}$, the choice $\phi_i = c_i R_i$ is bound to be a reasonably good one. This conclusion, borne out also by our simulation results later, is a very convenient one for practical implementations of the GPS scheme. To see this, consider the following properties of the optimal choice of weights in equation 8.

Balancing of QoS Whenever a feasible weight vector ϕ exists, there is a choice ϕ^* that equalizes all the probabilities

$$\mathcal{P}_i \equiv Pr\left[\sum_{j\neq i} x_j \phi_j > (\frac{P}{c_i R_i} - 1)\phi_i\right]$$

and achieves feasibility.

This follows from the fact that that each of the $\mathcal{P}_i$ are monotonically decreasing functions of ϕ_i and monotonically increasing in $\phi_j,\ j \neq i$. Thus, given any feasible vector ϕ, for which we must have $\max_i \mathcal{P}_i \leq e^{-\delta}$, one may adjust the components till the $\mathcal{P}_i$ are equalized so as to always lower $\max_i \mathcal{P}_i$. This may be done, for example, by lowering ϕ_k where k is the index corresponding to smallest value of $\mathcal{P}$, thus raising $\mathcal{P}_k$ and lowering all other $\mathcal{P}_i,\ i \neq k$.

A Useful Identity can be obtained by noting that

$$a_i Pr\left[\sum_{j\neq i} x_j \phi_j > (\frac{P}{c_i R_i} - 1)\phi_i\right] + (1-a_i) Pr\left[\sum_{j\neq i} x_j \phi_j > \frac{P}{c_i R_i}\phi_i\right] = Pr\left[\sum_{j} x_j \phi_j > \frac{P}{c_i R_i}\phi_i\right].$$

We write this more concisely as

$$a_i \mathcal{P}_i + (1-a_i)\mathcal{Q}_i = \mathcal{R}_i.$$

Given that $0 < a_i < 1$, this identity implies the orderings $\mathcal{P}_i > \mathcal{R}_i > \mathcal{Q}_i$ and $\mathcal{Q}_i < \mathcal{P}_i < \frac{\mathcal{Q}_i}{a_i}$.

To motivate statistical multiplexing in our wireless context, and the benefits of knowing a_i, see figure 1. The figure plots the number of identical users, all with the same transmission cost C, that can be supported as a function of C with and without statistical multiplexing, and the former is seen to be a factor of 2-4 larger. This was done for a unit rate requirement with activity 0.1 for all users and a total power of unity. Notice that the multiplexing advantage grows rapidly for small C, as a large number of users provide much better averaging in cumulative activity.

3.1 Rate Determination

We now deal with the rate allocations R_i in greater detail. As mentioned before, regulated connections can be assigned rates in a simple manner, while unregulated ones can make rate requests using criteria of the users choice. We believe that transferring the burden of rate specification to the user is a practical and realistic option, as applications vary widely in traffic characteristics and QoS demands. Further, pricing may be used to enforce efficient use and thereby keep the throughput of the system high. Attempts to meet tight delay/loss guarantees in error-prone wireless environments promise to be formidable and could complexify the resource management. With these aspects in mind, we nevertheless provide guidelines to choose rates and activity factors for data users.

As mentioned in the introduction RPS provides a robust means of resource sharing in which users with heterogenous requirements can be supported. This flexibilty mainly arises from translating a user requirement into the provision of a given rate (or equivalently power) with some given high target probability. Unlike other service metrics like delay or loss or buffer occupancy, rate can be controlled in a nearly instantaneous manner and produces immediate effect on the user's state.

In the results to be presented subsequently we will consider users who have a delay requirement, for example, which must be satisfied with a given probability. In this case flow i's QoS requirement is of the form

$$\Pr\{S_i > T_i\} \leq \delta_i \ , \tag{4}$$

where S_i is the steady-state delay (or, equivalently, sojourn time = waiting time + transmission time), and parameters T_i and δ_i are delay thresholds and a prescribed tail probability respectively. In the course of our results we will show that this can be met by providing a target rate R_i with a given high probability $1-\epsilon_i$. Other reasonable QoS criteria could potentially be accommodated in a similar manner.

For the purposes of our numerical results later, we will suppose each mobile is receiving traffic from statistically identical independent Markov Modulated ON-OFF sources with mean off time β^{-1} and mean on time α^{-1}. During the on period packets arrive as a Poisson stream with rate λ. Packets are taken to be of constant length and are supposed to arrive instantaneously. We term the mean arrival rate of packets as the source activity factor, $s_i = s$,

$$s_i = s = \frac{\beta}{\alpha + \beta}. \tag{5}$$

Using these parameters, the queueing model presented in Appendix C describes how one may arrive at a minimum rate R_i that would meet the user's delay with near certainty. The queueing model also provides the queue activity factors a_i defined before.

Clearly, better characterization of the traffic allows better and less conservative choices for R_i and a_i, resulting in higher capacity. For ease of implementation, one may tabulate the rate requirements and activity factors for a pre-designated set of *user classes* expected to be served. Also, as mentioned before, allowing a higher granularity in rates will reduce packing losses and improve efficiency.

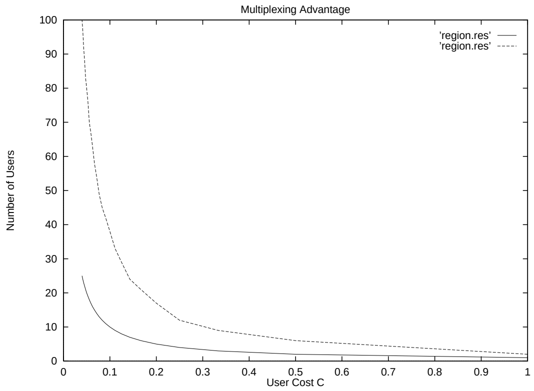

Figure 1: Benefits of statistical multiplexing when all users have a common transmission cost C.

independently as the minimum rates that would meet the users' QoS requirements (loss, delay etc.) with certainty. This allows us to cleanly separate the queueing and QoS related issues from the wireless aspects of the problem.

As in GPS, once such a rate has been determined, we admit the user with a weight ϕ_i that would guarantee this rate when the user's queue is active. These weights, which correspond to minimum power assignments, then control the sharing between users via the customary GPS principle of proportional sharing.

$$P_i = \frac{\phi_i}{\sum_{j\ active} \phi_j} P. \tag{3}$$

Further, to extract multiplexing gain when substantial numbers of bursty users are present, we require a characterization of the activity a_i of each user i, which denotes the fraction of time the user actually uses his assigned power. This factor depends on R_i, since it denotes the fraction of time the user's queue is being served, which is generally larger than the fraction of time the user sends data. If R_i is chosen to be the mean data rate, $a_i = 1$, while choosing R_i to be the peak rate would yield a_i to be just the source activity factor of the user. For intermediate rates, a_i lies inbetween these two extremes. Knowledge of this factor, under the assumption that users are independent and non-colluding, will lead to the appropriate weight assignments and an admission control procedure, as will be described in subsection 3.3.

This approach to statistical multiplexing of traffic using GPS was originally advocated by [8] in the context of wire-line networks, namely IP routers and ATM switches. Our simple linear model for achieved rate as a function of power allows us to cast this wireless packet-scheduling problem in similar terms. However, the techniques used to solve the problem need to be somewhat different here, since the number of data users in each cell is assumed to be typically small, unlike the IP/ATM scenario.

As described in [8], one may formulate this problem for regulated connections with known traffic characteristics and QoS using simple worst-case estimates for R_i and a_i, for example as described for the *Dual-Leaky-Bucket-Regulators* in [9]. Here the authors refer to the lossless *effective-bandwidth* of a connection $e_i^{(0)}$, which we can choose to be our required rate R_i, and the on-off activity factor ω_i for the bandwidth, which we refer to here as a_i. This is an important special case, as this and similar simple forms of traffic regulation have become widely adopted. In the absence of such regulation, the following subsection 3.1 describes a more sophisticated bursty on-off traffic model for data, from which one may infer R_i and a_i using queueing analysis described in appendix section C.

of the user can be set, for the given rate, according to a statistical criterion. We thus see that the c_i is determined as P_i^{Peak}/R_i which in general could depend on rate. In practice the peak power requirement may be chosen according to some percentile of the distribution of the time varying power requirement or to some other related statistical notion. We do not go into this question here but leave it as a topic for further investigation. Under these circumstances, the power allocations P_i must satisfy the following linear constraint,

$$\sum_i P_i = \sum_i c_i R_i \leq P \tag{1}$$

so as not to overload the base station transmitter.

Furthermore, whilst it may be possible to obtain the c_i using estimates of propagation and interference this is also a question we do not go into here. Instead in what follows we simply adopt the empirically obtained interference distribution given in [6]. Since our focus is on demonstrating the impact of scheduling and rate adaptation, rather than on the intricacies of power control, this is adequate for our purposes. The details of how the c_i are obtained are given in Appendix A.

We now turn to the argument that suggest the individual user power requirements may be considered as independent even when there is multi-path propagation. We suppose that (a) each cell's total transmit power P^T is constant in time, and (b) the multi-path propagation arises from a fixed number of "rays", which we take for the sake of argument to be 2. Let φ be the proportion of the total signal received in the first ray and $1-\varphi$ be the remaining proportion received in the second ray. Suppose we assume perfect orthogonality between signals within the same ray, and also assume the net signal to noise ratio (SNR) is obtained by adding the individual SNRs from each ray. Then the SNR for user i is given by

$$SNR_i = \frac{P_i G_i \varphi}{I_i + P^T G_i (1-\varphi)} + \frac{P_i G_i (1-\varphi)}{I_i + P^T G_i \varphi}, \tag{2}$$

where P_i is the total signal power transmitted to user i, G_i the propagation gain from the base station to user i, and I_i the net external interference and noise for user i. Note that $P^T G_i$ represents the total power received by user i from its base station. Thus, if the total transmitted power P^T is held constant, we see that each user's received SNR is proportional to the transmitted power P_i. Implicit in this model is the assumption that a user's receiver cannot eliminate "self-interference" from his own signal in the alternative ray. This is a reasonable assumption since most receivers used currently do not have this capability.

Given the above assumptions our model for the CDMA wireless down-link is thus equivalent to a single server queue in which there is net service capacity and the rates each user can obtain depends on a user dependent cost constrained as in (1). In what follows we will assume that these rates can be continuously assigned, indeed that any rate-tuple satisfying (1) is feasible. It then follows that the problem of meeting the users QoS requirements is simply one of determining what instantaneous powers should be assigned. Whilst in principle power adjustments can be done instantaneously, this is clearly far from practical and so we will suppose that all users powers are held fixed within short periods which we refer to as the scheduling intervals.

3 Design Issues

We now proceed to address the issue of how the rates R_i and the parameters of the RPS algorithm are picked for each user i. The simplicity of the RPS approach, like GPS, arises mainly from the isolation it provides to individual users, which we can use to select their guaranteed rates R_i

simple CAC framework of GPS along with its ability to provide user isolated rate specifications. Further RPS is made efficient through the mechanism of statistical multiplexing which allows a greater number of users to be supported by accepting a small probability of link overload arising from user activity (for statistical multiplexing in GPS, see [8]).

In real CDMA systems, users transmissions are power controlled in order to combat the effects of significant and rapid fluctuations in channel conditions and other user interference. This provides a second mechanism for link overload. RPS can take this into account by choosing the transmission costs in such a way that the probability of an overload through these fluctuations is held at an acceptable level. We elaborate on this subsequently.

The RPS approach is, of course, not without it's drawbacks. In the first place it does not adapt rate as the state of the queues change. It can be shown that adapting rate to queue lengths is a necessary characteristic of an optimal algorithm at least for some user criteria. Perhaps more seriously it can be shown that RPS is not always stable when in fact stability can be achieved. These drawbacks, taken together with the problem of parameter setting discussed later, show that this algorithm is not ideal. However it is our view that the flexibility and robustness of RPS nevertheless make it an attractive candidate for resource management of a wireless data link.

The rest of the paper begins by discussing the model for the forward link and the user transmission costs. There follows a description of the user QoS requirements and the method by which RPS meets them. The requirements are not chosen with any particular application in mind but are introduced by way of illustration. Subsequently the question of setting the parameters for RPS is examined in detail. We next present various definitions of capacity along with numerical results to compare them. In the appendices mathematical details are given for the queueing model (see also [7] and references therein) used in obtaining our results together with techniques for refining the parameter settings in RPS. These follow the conclusions.

2 Discussion of Wireless Model

We examine a single CDMA cell with n data users distributed randomly within the cell. The users communicate only with this cell, so that there is no soft handoff. The cell has a total power P with which to perform data transmissions over its forward link. (Note that this power will vary between cells in the network as each cell is allocated a different peak power according to its load and the load of its neighbors. The optimal allocation of these powers is a network-wide problem which we do not go into here.) There is a QoS requirement for each user which in general may take various forms including mean delay, probability of exceeding a given delay threshold etc. The users signals are transmitted using orthogonal codes, we assume, so that they can be taken as non-interfering within the same multi-path ray but not across them.

Our model of the forward link was first presented in [2]. We first suppose that the power a user requires to achieve a given rate R is independent of the power assigned to other users within the cell. Clearly this is the case if there is perfect orthogonality since the within cell interference is removed at the mobile receiver. However as we detail below this is also reasonable when there is imperfect orthogonality arising from multi-path propagation.

But before so doing, let us consider the relationship between power and rate for an individual user. We make the idealized assumption that the peak power a user requires is linearly dependent on the rate. The constant of proportionality for user i, c_i, we term the users effective weight. In practice of course, the power needed to transmit at a given rate is time varying as mentioned in the introduction. To take this variability into account we will suppose that the peak power requirement

Rate Processor Sharing: A Robust Technique for Scheduling Data Transmissions in CDMA Wireless Networks

Krishnan Kumaran Phil Whiting

Bell Laboratories, Lucent Technologies
600-700 Mountain Avenue
Murray Hill, NJ 07974-0636
Tel: +1 908 582 {4182,6912 }
FaX: +1 908 582 3340
{kumaran,pawhiting}@lucent.com

Abstract

We present an algorithm for scheduling data transmissions on the forward link of a CDMA wireless network. Our algorithm, Rate Processor Sharing (RPS), is based on Generalised Processor Sharing (GPS), a widely adopted method for resource sharing in wire-line networks. As such the RPS algorithm, in common with GPS, provides an effective means of performing Call Admission Control (CAC) in the presence of users with heterogenous Quality of Service (QoS) requirements and traffic characteristics. Unlike GPS, however, RPS must take into account user interference and the actions of power control.

Apart from discussing the design and implementation of the algorithm, we also address the issue of capacity. These take into account statistical multiplexing amongst bursty users. Unlike CDMA voice networks, the heterogeneity amongst data users makes it difficult to devise a natural definition of network capacity. Indeed, as we show, capacity varies substantially, depending on the actual definition used.

1 Introduction

We consider the problem of scheduling and rate allocation for mobile data transmissions on the forward link (base-to-mobile) of a CDMA cell. In future such networks it is likely that the mobile users will have heterogenous QoS and traffic characteristics. The problem thus arises as to how to share resources between mobiles in such a way that their QoS requirements can be effectively met whilst operating with a reasonable level of efficiency. This situation is to be contrasted with voice networks where the principle of CDMA allows statistical averaging to take place among a large number of delay intolerant users with a common requirement. For data traffic, which is typically more bursty and delay tolerant, scheduling can be used to manage power usage and as a mechanism to control interference.

In what follows we propose a simple and efficient scheduling scheme which we term Rate Processor Sharing (RPS). Our scheme, based on Generalised Processor Sharing, see for example [11, 10, 13, 12], allocates power to individual users based on their QoS requirements and their cost of transmission in terms of energy-per-bit. In common with GPS, RPS converts user QoS requirements into a minimal rate assignment which is provided with high probability. Combined with knowledge of the user costs this translates into a power allocation. RPS thus inherits the