Performance of a Global Congestion Measure for CDMA Networks

Lachlan L. H. Andrew and Stephen V. Hanly
Electrical and Electronic Engineering
University of Melbourne, Parkville, Vic 3052, AUSTRALIA
Ph +61 3 9344 9208 Fax +61 3 9344 9188
{L.Andrew,S.Hanly}@ee.mu.oz.au

Abstract— **A recently proposed global congestion measure for CDMA networks is applied to call admission control, call dropping, and band allocation in multi-carrier CDMA networks. It is shown to perform comparably with existing techniques for call dropping, and better than existing techniques for admission control and band allocation. Difficulties arising in the measurement of this parameter are also investigated.**

I. Introduction

Effective management of any communications system requires an accurate real-time measure of how "congested" the system is. In a simple circuit switched network, the congestion is easily measured by counting the number of available circuits on each link. In packet switched networks, it can be assessed by measuring the queue sizes at all of the nodes. Code division multiple access (CDMA) networks, with their soft capacity, require a new concept of capacity. In [1], a global congestion measure for CDMA networks was proposed, based on the dominant eigenvalue, r, of a matrix determined by the path gains of the mobile stations to the base stations. This measure was shown to have useful mathematical properties, such as being below a threshold if and only if the system is "feasible", and always increasing when a new user is added. However, no empirical evidence was presented that it can be successfully used for network management tasks. In this paper, we demonstrate that using r in strategies for call admission control and for band allocation in multiband CDMA systems can provide better performance than traditional local congestion measures.

The congestion measure r will be defined formally in Section II, and Section III investigates the correlation between outage events and r being high. Sections IV to VI determine how effective r is in determining which calls to drop or accept, and in which band of a multiband system new calls should be placed. Finally Section VII considers how r can be measured in a real system without eigenvalue calculations.

This work was funded in part by the Australian Research Council (ARC).

II. Measuring congestion — r

This paper will examine the congestion of a system working with an optimal power control scheme proposed in [2]. Under this power control algorithm, users iteratively update their transmit powers to the level which would give them exactly their desired signal to interference ratio (SIR) if the interference remained the same as at the previous iteration. That is, $p_{i,n+1} = I_{i,n}/\alpha_i$, where $p_{i,n}$ is the power transmitted by user i at iteration n, $I_{i,n}$ is the effective interference (after despreading) at the base station to which user i is connected at iteration n, and α_i is the SIR requirement of user i.

It was shown in [2] that the rate of convergence of this algorithm is governed by the dominant eigenvalue of the matrix $\mathcal{A}$, given by

$$\mathcal{A}_{ij} = \frac{\alpha_i}{W}\frac{\Gamma[j,c_i]}{\Gamma[i,c_i]},$$

where α_i is the SIR required by user i, W is the processing gain, c_i is the base station to which user i is connected, and $\Gamma[j,c_i]$ is the path gain from user j to base station c_i. If r, the Perron-Frobenius eigenvalue of $\mathcal{A}$, is less than 1, the algorithm converges. If $r > 1$, then no allocation of powers to users can achieve all of the users' SIR requirements. It was conjectured in [3] that r may be a useful measure of congestion. For a given network state, it can easily be measured in a decentralised manner by observing the rate of convergence of the power control algorithm. It also provides a measure of the global congestion, rather than the specific congestion of a single cell.

0-7803-5435-4/99/$10.00 © 1999 IEEE

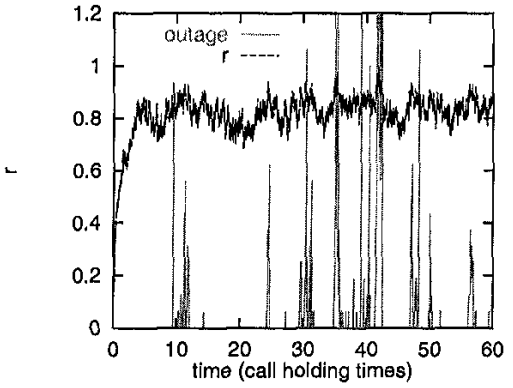

Fig. 1. r and system-wide outages (divided by 8) vs time.

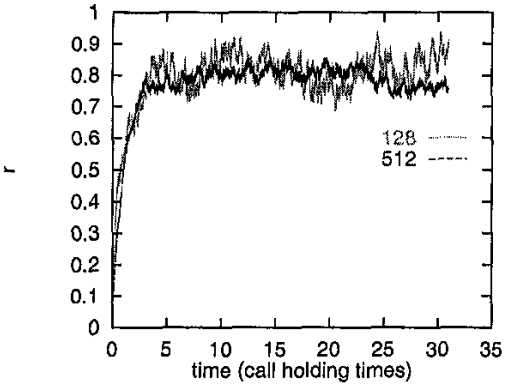

Fig. 2. r vs time for small (processing gain 128) and large (512) system.

III. VARIATION OF r

A new measure of congestion should be correlated with existing measures. Figure 1 shows the changes in r over time, and also periods of high outage. Clearly outage is higher when r is high. However r can be measured by any base station, whereas outage can only be measured at the connected base stations. Detecting congestion before users go into outage is important for good network management. In particular, techniques such as handoff and admission control, which work to prevent outage, should be triggered before outage occurs to maintain sufficient quality of service. Figure 1 indicates that r can be useful in predicting outage events.

Figure 2, showing r for systems with spreading gain 128 and 512, shows that fluctuations in r decrease for "larger" systems.

IV. DROPPING

When a CDMA system is overloaded, some or all calls will be unable to attain their required signal to interference ratio (SIR) and go into outage. In order to ensure sufficient quality of service for the majority of users, some users must be dropped. The simple scheme of dropping users when they have been in outage for longer than a threshold time has proved to be quite effective. However it may not be optimal. If the aim of dropping users is to drop the fewest users required to make the system feasible, this can be achieved by dropping the users which have the greatest impact on r, which need not be those currently in outage. To determine the effectiveness of r as a congestion measure, three dropping strategies were tested by simulation. In all cases, a 4×4 toroidal hexagonal grid of base stations was used, and there was no admission control. Users did not move, but mobility was simulated by recalculating the shadowing occasionally. When outage occurs (i.e., the SIR drops below 6dB), the first strategy simply drops all of the users currently in outage. The second strategy drops a randomly selected user which is currently in outage. The third drops the user which makes the largest contribution to r. Figure 3 clearly shows that the most aggressive strategy, which drops all users currently in outage, performs substantially worse than the either scheme which drops one user at a time, but that there is little difference between the other two schemes. This shows that r correctly identifies users which are contributing to the poor performance of the system. These results also highlight the importance of a conservative dropping algorithm. When a strong interferer connects to a neighbouring cell, several users may go into outage. Typically all of these are dropped simultaneously (say after a constant timeout). However if one user is dropped and the power control is allowed to stabilise, the reduction in transmit power of all of the other users may be sufficient for the other users to return from outage. This could be achieved in practice by using a random timeout for dropping.

The drawback of this strategy is that it is difficult to assess the impact of a single user on r. The most obvious method is tentatively dropping the user, and measuring the change in rate of convergence of the power control algorithm. Not only is this very time-consuming, but each user will briefly go into outage while its impact is being measured, even if only a single user is initially in outage. However, these results support r as a measure of network congestion, which can in principle be used for network management. Investigation is continuing into whether users which contribute highly to r can be characterised by more measurable parameters (like relative position or signal strength) which may be able to be used for practical dropping algorithms.

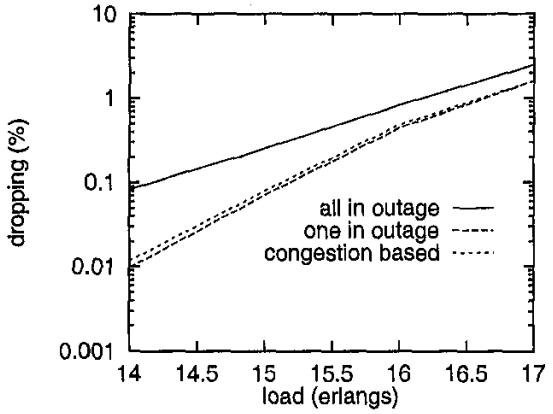

Fig. 3. Dropping probability vs. arrival rate for outage-based and congestion-based dropping.

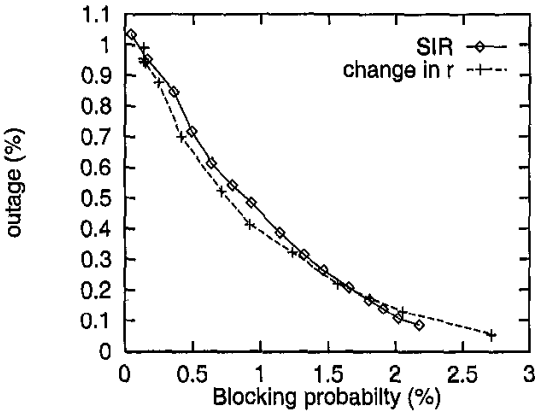

Fig. 4. Outage vs. blocking for r-based admission control, and conventional SIR-based admission control.

V. Admission Control

When a system is becoming congested, it is better to block incoming calls than to allow new users to connect to the system and cause existing users to receive inadequate service. This is commonly done by simply limiting the number of users who may connect to each base station, which entirely ignores the (often substantial) impact of other-cell interference. A more sophisticated approach considers the interference from neighbouring cells and blocks new calls when the total received interference at the target base station is excessive These methods are compared in [4]. A further enhancement would be to consider the interference caused to neighbouring cells. For example, new calls could also be blocked when the interference at neighbouring base stations would become excessive if the new call were accepted. However this requires communication between the base stations, and it is not clear how many tiers of neighbours should be consulted. An alternative approach would be to use r to determine whether or not a call should be admitted.

Because r is a global measure of congestion, admission should not be based simply on the current value of r. This would cause congestion in one part of the network to prohibit access in a very distant, lightly loaded part. Instead, admission should be based on the impact of the new user, that is, the change in r. This can be measured in terms of the "spare capacity", defined as $1 - r$. A maximum limit can be place on the allowable proportional decrease in spare capacity, $(r_{new} - r_{old})/(1 - r_{old})$. It is not clear that this is the optimal criterion, but it is nonetheless effective.

The effectiveness of an admission control scheme is best measured by considering how well it can trade blocking for outage, the aim being to minimise the blocking for a given outage rate. Figure 4 shows the tradeoff possible with the proposed scheme for a load of 17 Erlangs. The acceptance criterion was that $(r_{new} - r_{old})/(1 - r_{old}) < \theta$ for $\theta = 1, 0.5, 0.25, 0.167, 0.125, 0.108, 0.095, 0.083, 0.072, 0.062$. There is a lower limit to the blocking performance of this approach, because it will never allow a call into the system if that makes the system infeasible, but analogous schemes could be developed to consider the case allowing negative spare capacity.

For comparison Figure 4 also shows the results of blocking based on the interference measured at the target base station. The proposed algorithm based on the reduction in spare capacity outperforms the algorithm based solely on the SIR of the target base station over a wide range of operating conditions, most notably for lower blocking probabilities. Note however that blocking based on local measurements can allow the entire system to become infeasible occasionally, which can reduce the blocking arbitrarily much, at the expense of increased outage. While the improvement in performance offered by the proposed approach may not in practice justify the extra computational complexity, these results show that the proposed scheme can be used as a benchmark against which other less optimal admission algorithms can be compared.

VI. Band allocation

In some CDMA systems, multiple non-interfering bands are used to increase capacity [5,6]. The performance of such systems depends on how new users are allocated to bands. This is typically done by allocating new users to the band with the fewest current users in the cell of interest ("least load") [6]. An ob-

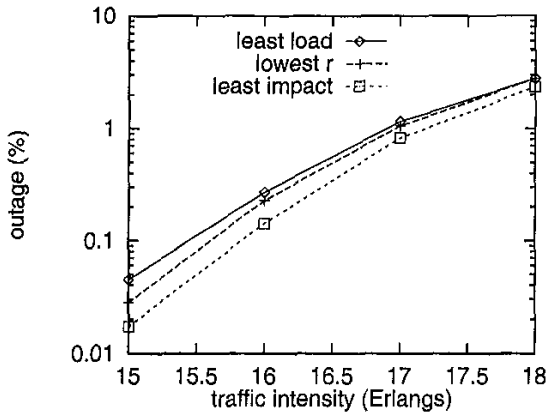

Fig. 5. Outage probabilities using least load, lowest r, and least impact strategies.

vious extension is to allocate users to the least congested band (with the smallest r). A better strategy is to allocate users to the band where they contribute least to the congestion. The user can be added to the band in which it causes the smallest proportional decrease in spare capacity. Figure 5 shows the total outage probability when calls were allocated using the least load, lowest r, and decrease in spare capacity strategies. Clearly the latter approach is superior.

Using r directly for band selection is more feasible than for dropping. The network only needs to try the new user in each of the bands, of which there are typically less than ten, rather than disrupt all users in the network, of which there may be thousands.

VII. Measurement issues

This section will investigate the feasibility of measuring r from changes in the received power level. This will be done by considering the dynamics of the system $\mathbf{y}'(t) = (A - I)\mathbf{y}(t)$, which converges to the feasible power allocation (if one exists). Here $\mathbf{y}(t)$ is the vector of powers received at each base station, and A is a matrix whose eigenvalues are the same as those of $\mathcal{A}$ [1]. The total power at a given base station in such a system has the solution $y(t) = K - a_1 e^{(r-1)t} + o\left(e^{(r-1)t}\right)$. Let $y_i = y(t_0 + i\delta)$ for some constants t_0 and δ. For sufficiently large t_0 (i.e., once the lower order terms have died out), and ignoring the effect of measurement errors, r can be determined from

$$\frac{y_1 - y_0}{y_2 - y_1} = e^{(1-r)\delta}$$

In practice there will always be measurement errors in the y_is. If the measured values are $\hat{y}_i = y_i + w_i$ then the relative error in $(\hat{y}_1 - \hat{y}_0)/(\hat{y}_2 - \hat{y}_1)$ will be

$$\frac{1}{y_1 - y_0}\left[\left(1 + e^{(1-r)\delta}\right) w_1 - w_0 - e^{(1-r)\delta} w_2\right]. \quad (1)$$

The main sources of error are the $o\left(e^{(r-1)t}\right)$ term, thermal noise, nonlinearity of the measuring device, fast fading, voice activity and quantisation of the power control. Because measurements are being made almost at equilibrium, the changes in received power will be small, and quantisation of the power control signals will be particularly important.

Two important issues arise here. The sampling instants (governed by t_0 and δ) must be determined, and the accuracy with which the measurements should be made must be determined. Each of these involves a tradeoff. If t_0 is made too small, the terms due to non-dominant eigenvalues will be significant. However, if t_0 is too large, the system will almost have converged before the measurements are taken, and the changes $y_i - y_{i-1}$ will be small compared to the measurement errors. Furthermore, the larger t_0 is, the longer the measurement process takes. If δ is too small then the factor $1/(y_1 - y_0)$ in the relative error becomes large, but if δ is too large the same problems occur as for t_0 too large. Clearly if the accuracy with which the measurements are taken is too low, the resulting estimate of r will be bad. However increasing the accuracy requires increasing the averaging period to reduce the quantisation noise, which places a lower bound on δ. Moreover this may actually decrease the accuracy because the ideal received power will change over the averaging period.

Of these problems, the most fundamental is the impact of the decaying term $o\left(e^{(r-1)t}\right)$ caused by the non-dominant eigenvalues. Even in the absence of measurement errors, this can cause the estimate to be anywhere in the range $(-\infty, +\infty)$ by causing the difference $y_2 - y_1$ (or $y_1 - y_0$) to be approximately zero, and possibly of the wrong sign. This is illustrated in Figure 6, which shows a phase portrait of the system $x = 0.25 + e^{-t} - 1.5e^{-2t}$, $y = 0.25 + e^{-t} + 0.5e^{-2t}$, which is a possible trajectory for the total received powers at each base station of a two-cell system. If measurements are taken at points P_1 and P_2 then either the numerator or denominator of $(\hat{y}_1 - \hat{y}_0)/(\hat{y}_2 - \hat{y}_1)$ could be zero.

However, for purposes such as admission control, the measurements will only be made by the base station, i, to which the new mobile is connected. It is to be expected that the largest change in received power will occur at base station i, and hence the chance of

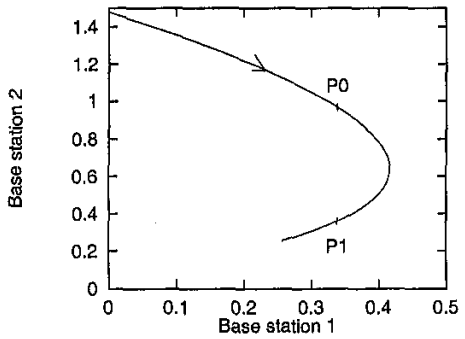

Fig. 6. Possible phase portrait of received powers at base stations of a two-cell system.

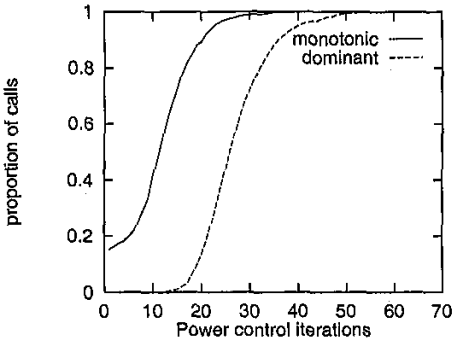

Fig. 7. Proportions of calls such that $p(t)$ is monotonic and after which the dominant term is clearly dominant after time t_0.

the received power at this base station "overshooting" is low. (That is, it is expected that, if the received power vector is resolved into the eigenvectors, component i of each of these resolved vectors will have the same sign.) This would mean that y_2 and y_1 (and y_1 and y_0) are clearly distinct, and the above problem would not occur.

To test this intuition, a simulation was performed with no admission control, and the impact of the call arrivals was measured. Propagation conditions were assumed to be constant and the load was 17 Erlangs. For every arriving call, the change in the total (equilibrium) received power at each base station, $\Delta\mathbf{y}$, was recorded, along with the matrix A describing the system with the new call and the number of the base station, i, to which the new call arrived. Then $\Delta\mathbf{y}$ was decomposed as $\sum a_j x_j$ where $\{x_j\}$ are the eigenvectors of A and $\{a_j\}$ are coefficients. If the eigenvalues, λ_j, of A are distinct, the solution of the linearised system is then $\mathbf{y}(t) = K + \sum a_j e^{(\lambda_j - 1)t} x_j$, with derivative $\mathbf{y}'(t) = -\sum(1 - \lambda_j) a_j e^{(\lambda_j - 1)t} x_j$. Assuming that no soft handoff occurs, a sufficient condition for the ith component of $\mathbf{y}(t)$ to be monotonic for time $t > t_0$ is that $\sum_{j \in B}(1 - \lambda_j) a_j e^{(\lambda_j - 1)t_0} x_{ij} > \sum_{j \notin B} |(1 - \lambda_j) a_j e^{(\lambda_j - 1)t_0} x_{ij}|$, where $B = \{j : a_j x_{ij} > 0 \wedge \lambda_j \in \mathbf{R}\}$ and x_{ij} is the ith component of x_j. However, for the measurement to be meaningful, it is necessary for the dominant term to be an order of magnitude larger than the sum of the other terms. The proportion of call arrivals for which $\mathbf{y}(t)$ is monotonic after time t_0 and the proportion for which the dominant term is an order of magnitude larger than the others after time t_0 are both shown in Figure 7. If the update interval is assumed to be 1.25 ms, as in the IS-95 standard, then acceptable accuracy is generally achieved after 60 ms. This is faster than the timescale at which other factors cause the power level to change, such as shadowing or the admission of new calls in neighbouring cells. (In a large, multicellular system, the arrival of a call in *any* cell has a slight impact on the power control, but new calls in distant cells can generally be ignored. Quantifying their effect is a future research topic.)

VIII. Conclusion

This paper has investigated the effectiveness of a simple global congestion measure for CDMA networks. It has been shown that the measure agrees well with existing measures of congestion, and can be used for admission control or band allocation in a multi-band system. It can in principle also be used to guide dropping of excess calls. The strengths of the global congestion measure are that it can be measured at any point in the network and that it can predict congestion before problems occur.

References

[1] S. V. Hanly, "Congestion measures in DS-CDMA networks," *IEEE Trans. Commun.*, vol. 47, pp. 426–437, Mar. 1999.

[2] S. V. Hanly, *Information Capacity of Radio Networks*. PhD thesis, King's College, Cambridge University, `http://www.ee.mu.oz.au/staff/hanly/publications.html`, 1993.

[3] S. V. Hanly, "An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1332–1340, Sept. 1995.

[4] Y. Ishikawa and N. Umeda, "Capacity design and performance of call admission control in cellular CDMA systems," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 1627–1635, Oct. 1997.

[5] L. L. H. Andrew, "Measurement-based band allocation in multiband CDMA," in *Proc. Infocom '99*, (New York), 1999, pp. 1536–1543.

[6] T. Dean, P. Fleming, and A. Stolyar, "Estimates of multicarrier CDMA system capacity," in *Proc. Winter Sim. Conf.*, (Washington, DC), 1998, pp. 1615–1622.