Toward Understanding of Metastability in Cellular CDMA Networks: Emergence and Implications for Performance.

Daniel Genin
NIST, Gaithersburg, MD
Email: dgenin@nist.gov

Vladimir Marbukh
Senior Member, IEEE,
NIST, Gaithersburg, MD
Email: marbukh@nist.gov

***Abstract*—We investigate the metastable behavior in a model of a cellular CDMA network with multiple service classes. While the Markov model provides an accurate "microscopic" model of the network behavior, the dimension of this model grows exponentially with the number of cells precluding solution of the corresponding Kolmogorov equations. Dimension of the mean-field approximation model grows only linearly with the number of cells making this approximation computationally tractable. Through numerical analysis we show that the equilibrium manifold of the mean-field model develops "folds" under increasing network load which give rise to multiple stable equilibria. These multiple equilibria can be interpreted as describing network's metastable states. By comparing simulation data with numerical computations we show that mean-field approximation can be used to predict equilibrium states of a realistic network. We construct a sample phase diagram showing distribution of metastable regions in the user load plane, derive a formula for the likelihood of successful service completion for metastable states and discuss performance characteristics of the network.**

***Index Terms*—cellular network, performance, mean-field approximation, metastability.**

I. Introduction

Following [2] we consider a Markov model of a cellular CDMA network with multiple user service classes. The model is applicable not only to cellular networks but to any network with multiple service classes, and migrating users. Ubiquity of such networks makes their optimal design and control a topic of great interest.

In [2] it was proved that for certain parameter values the Markov model exhibits metastability as the number of cells in the network approaches infinity. Simulations performed in [1] confirmed that metastability can appear in realistic models of CDMA networks with two service classes and as few as 49 cells. Understanding which network parameters contribute strongly to metastability and how metastable states depend on these parameters may help in operation and design of complex service networks such as CDMA cellular networks.

Metastability is strong heterogeneity in the distribution of mass of the stationary probability distribution. If the stationary distribution is concentrated on a few relatively small disjoint subsets of the phase space a realization of the Markov process will be found most of the time in one of the states contained within the heavily weighted subsets, with the rest of the phase space visited only occasionally. This means that the Markov process can be approximated by another Markov process with a smaller phase space, whose states correspond to the above subsets and are the metastable states of the original Markov process. In practice, this means that the time series of the system with metastable states exhibits long periods of apparently steady state behavior with rapid transitions in between. Furthermore, due to asymmetry in the probability mass distribution between the metastable states, the time spent in the most likely state is exponentially long compared to the time spent in other metastable states.

We build on the results of [1] by using the mean-field approximation (already developed in [2]) and computer simulations to begin piecing together the structure of the model's phase space with regard to metastability and the implications of metastability for the network's performance. By using numerical computations we construct a generic phase space diagram for the mean-field approximation model. The advantage of the phase diagram is that it allows to see at a glance how parameter regions corresponding to metastable regimes are distributed in the parameter space and are related to the parameters of the network. It also provides a kind of a "topographic" map which the network operator can use to steer the network in to the desirable state or away from undesirable states. Mean-field approximation also permits computation of the metastable states with considerable precision, which makes estimation of the quality of service possible for metastable states. We derive a general formula for quality of service and give a simple expression for it in the special case of the homogeneous network traffic. All this suggests that with sufficient understanding of metastability and an appropriate admission control policy, capable of stabilizing metastable states, metastable states may be practical in network operation.

The paper is organized as follows. Section II describes the performance model and analysis. We briefly introduce the "microscopic" Markov model and the corresponding mean-field approximation broadly following [2], and derive an expression for quality of service. Section III presents results on the structure of the phase diagram and data from simulations testing the predictions of the mean-field approximation. Section IV is devoted to performance analysis in terms of quality of service. We conclude with the summary of our findings and directions for future research in Section V.

II. Network performance model

In Subsection II-A we briefly introduce the Markov model proposed in [2]. Our model differs somewhat from the original in that the network parameters are permitted to vary across cells and migration occurs only between adjacent cells. In Subsection II-B we derive the corresponding mean-field approximation and the fixed-point equations for the homogeneous network case.

A. Markov Model

Consider a network with a set of cells, representing cell towers, J and S service classes with varying resource demands. The service classes correspond to users using different services e.g. voice and video [1]. Cell $j \in J$ has capacity C_j, while each user of service class $s = 1, \ldots, S$ requires capacity b_s and has an exponentially distributed "lifespan" (call duration) τ_s, with average $\bar{\tau}_s = 1/\mu_s$. Numbers of users at all cells of the network are described by a vector $X = (X_{sj}, s = 1, \ldots, S, j = 1, \ldots, J)$, where X_{sj} is the number of users of service class s at the jth cell. The feasible region for vector X is given by

$$\mathcal{X} = \left\{ X : \sum_{s=1}^{S} b_s X_{sj} \leq C_j, j = 1, \ldots, J \right\} \qquad (1)$$

This assumption describes Frequency Division Multiple Access (FDMA) network. In the case of Code Division Multiple Access (CDMA) assumption (1) can be justified if inter-cell interference is small.

We assume that new users of class $s = 1, \ldots, S$ originate at the cell $j \in J$ according to a Poisson process of rate λ_{sj}. Each user in the network performs a random walk over the set of cells J. The random walk is described by a Markov process $\xi_s(t)$ with a set of states J, continuous time and transition rates between adjacent network cells $i, j \in J$, $i \neq j$ equal to γ_{sij}. For consistency we set $\gamma_{sii} = 0$ and $\gamma_{sij} = 0$, if i and j are not adjacent in the network. If an arriving user takes the cell out of the feasible region he is ejected from the network.

Time evolution of the network state vector X is described by a Markov process with state space $\mathcal{X}$. Under the above assumptions evolution of the vector $X(t) = (X_{sl}(t))$ is a time-homogeneous Markov process with a finite number of states $|\mathcal{X}|$. The resulting process is ergodic provided that arrival, migration and service rates are non-zero and hence possesses a unique invariant probability distribution $P(X) = \lim_{t\to\infty} P(t, X)$, which is uniquely determined by the steady-state Kolmogorov equations:

$$\begin{aligned} P(X) \sum_{j\in J} & \left(\lambda_{sj} + \sum_{s\in S} X_{sj} \left(\mu_s + \Gamma_{sj} \right) \right) \\ &= \sum_{j\in J} \sum_{s\in S} P(X - 1_{sj}) \lambda_{sj} \\ &+ \sum_{j\in J} \sum_{s\in S} P(X + 1_{sj})(X_{sj} + 1) \left(\mu_s + \sum_i \check{\gamma}_{sji}(X) \right) \\ &+ \sum_{j\in J} \sum_i \sum_{s\in S} P(X + 1_{sj} - 1_{si})(X_{sj} + 1) \hat{\gamma}_{sji}(X - 1_{si}) \end{aligned} \qquad (2)$$

supplemented with normalization condition $\sum_{\mathcal{X}} P(X) = 1$.

In (2) the vector $1_{si} = (\delta_{sh}\delta_{ij} : h \in S, j \in J)$, where δ_{ij} is the Kronecker symbol $\delta_{ij} = 1$ if $i = j$ and 0 otherwise and

$$\begin{aligned} \check{\gamma}_{sji}(X) &= \begin{cases} \gamma_{sji} & \text{if } X + 1_{si} - 1_{sj} \notin \mathcal{X} \\ 0 & \text{otherwise} \end{cases} \\ \hat{\gamma}_{sji}(X) &= \begin{cases} \gamma_{sji} & \text{if } X + 1_{si} - 1_{sj} \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases} \\ \Gamma_{sj} &= \sum_i \gamma_{sji} \end{aligned}$$

We point out several obvious reductions of the above model

- If $\gamma_{sij} = 0$ the network consists of J independent queues and the equilibrium probability distribution is a J-fold product of the probability distribution for a finite capacity multi-service queue.
- If cell capacity is infinite the entire network acts as a single infinite server queue since no user is ever rejected.

B. Mean-Field Model

The Markov model gives a full cell-by-cell description of the network dynamics but the number of equations is too large to compute the stationary distribution even for a relatively small network, due to the exponential growth of the state space with the number of cells J. The mean-field approximation provides a more tractable description of the equilibrium distribution. In mean-field approximation the influence of the *specific* state of the network on a given cell is replaced by the influence of the *averaged* network state with respect to a hypothesised equilibrium probability distribution. The probability distributions of individual cells are further assumed to be independent and the stationary probability distribution for the network takes on a product form $P = \prod_{j\in J} P_j(X_j, \bar{X}_{-j})$,where $X_j = (X_{sj}, s = 1, \ldots, S)$ is the vector describing the number of users of each class present at the cell $j \in J$ and $\bar{X}_{-j}$ is the expected state of the network without the j-th cell with respect to P. In passing to the mean-field approximation we are making an assumption that the system can be approximated in this way, which we confirm later with simulations.

$$
\begin{aligned}
P_j(X_j,\bar{X}_{-j})\sum_{s\in S}\left(\lambda_{sj}+\bar{\Lambda}_{sj}(X)+X_{sj}\left(\mu_s+\Gamma_{sj}\right)\right) = \\
\sum_{s\in S}P_j(X_j+1_s)(X_{sj}+1)\left(\mu_s+\Gamma_{sj}\right) \qquad (3)\\
+\sum_{s\in S}P_j(X_j-1_s)\left(\lambda_{sj}+\bar{\Lambda}_{sj}(X)\right)
\end{aligned}
$$

where $\bar{\Lambda}_{sj}(X)=E[X_{si}\gamma_{sij}]$

The number of equations is now proportional to the number of cells J in the network but the equations are nonlinear and so impossible to solve explicitly. Non-linearity of the mean-field equations permits multiple solutions, or multiple equilibrium probability distributions. This, however, does not contradict the ergodicity of the original Markov model, rather it indicates that the stationary distribution for the full Markov model is a combination of equilibrium distributions of the mean-field approximation.

Assuming $P_j=Q$ are all identical, as might be expected for a homogeneous network, we can write down just S equations for $\bar{X}_s$, the expected number of users of class s per cell.

$$
\bar{X}_s=E_Q[X_s] \qquad (4)
$$

where $E_Q[]$ is expectation with respect to the equilibrium probability distribution $Q=Q_{\bar{X}}$. The later can be computed explicitly as the equilibrium distribution for a single multi-service class finite capacity queue because of its reversibility [3]

$$
\begin{aligned}
Q_{\bar{X}_s}(X) &= Z(\bar{X}_s)^{-1}\frac{\prod_s\rho_s(\bar{X}_s)^{X_s}}{\prod_s X_s!}\\
Z(\bar{X}_s) &= \sum_{X\in\mathcal{X}}\frac{\prod_s\rho_s(\bar{X}_s)^{X_s}}{\prod_s X_s!}
\end{aligned}
$$

where ρ_s are mean aggregate loads including intra- as well as extra-cell arrivals

$$
\rho_s(\bar{X}_s)=\frac{\lambda_s+\Gamma_s\bar{X}_s}{\mu_s+\Gamma_s}
$$

C. Quality of service

Performance of a service network can be quantified in a number of different ways. The operator of the network is most likely to be interested in maximizing the pay-off, which may be reasonably expected to be a linear function of the average number of users in the system

$$
\phi(\rho)=\sum_{s\in S}\sum_{j\in J}\bar{X}_{sj}(\rho)p_s
$$

where $\rho=\{\rho_{sj}:s\in S,j\in J\}$ and p_s is the profit per user of class s who is successfully serviced. For simplicity we assume that incomplete service brings zero revenue. A measure of performance which is likely to be of concern to network's customers, and so again, to its operator, is quality of service, i.e. the likelihood of successful service completion, Q_{sj}.

We derive an expression for the likelihood of successful service completion for the s-th service class user Q_{sj} in terms of the blocking probabilities B_{sj}, where B_{sj} stands for the likelihood that the j-th cell will block an arriving user of the corresponding class. B_{sj} can in turn be expressed in terms of $\bar{X}_{sj}$, which can be computed numerically from the mean-field equations (3),

$$
B_{sj}=1-\bar{X}_{sj}/\rho_{sj} \qquad (5)
$$

where $\rho_{sj}=(\lambda_{sj}+\bar{\Lambda}_{sj}(X))/(\mu_s+\Gamma_{sj})$.

Let the probability of successful service completion for an s-th service class user originating at the j-th cell be denoted by Q_{sj} then

$$
Q_{sj}=\sum_{\omega\in\Omega_j}P[\text{call unblocked on }\omega]P[\omega]
$$

where Ω_j is the collection of all paths through the network based at j, $\omega=\{j,j_1,j_2,j_3,\dots,j_k\}$, $j_l\in J$. Because the blocking probabilities of cells are independent (by mean-field assumption) the probability of navigating a given path successfully is equal to the probability of success in k independent trials with randomly distributed times. It can be shown that

$$
Q_{sj}=\sum_{\omega\in\Omega_j}\frac{\prod_{i=0}^{|\omega|}(1-B_{s\omega_i})\prod_{i=0}^{|\omega|-1}\gamma_{s\omega_i\omega_{i+1}}}{\mu_s^{-1}\prod_{i=0}^{|\omega|}\mu_s+\Gamma_{s\omega_i}} \qquad (6)
$$

In a homogeneous network, i.e. one in which all the rates λ_{sj} and γ_{sij} are equal, the equation (6) simplifies further if we assume in addition that equilibrium probability distribution is also homogeneous. In this case the $B_{sj}=B_s$ and the probability of successful completion is

$$
Q_s=\frac{\mu_s(1-B_s)}{\mu_s+\Gamma_sB_s}.
$$

The last formula can be rewritten in terms of the easily computable average number of users per cell $\bar{X}_s$

$$
Q_s=\frac{\bar{X}_s}{\lambda_s/\mu_s}, \qquad (7)
$$

which is can be understood as the conservation law: the average number of customers in the system is equal to the load times the proportion of the customers serviced. *Thus, in the homogeneous case in equilibrium each cell behaves as an isolated queue.*

III. Metastability

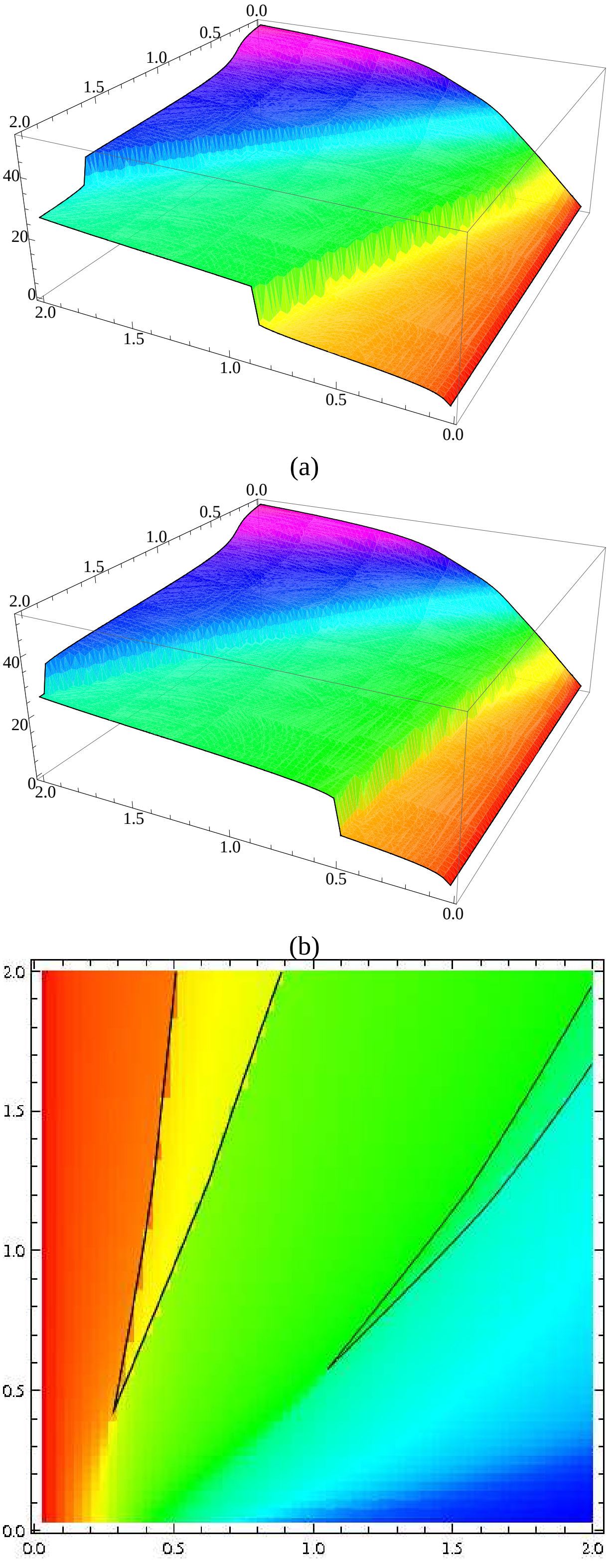
In this section we construct a phase diagram for the special case of the homogeneous network with two service classes by solving the mean-field equations (4) over a grid of parameter values. Solutions of the mean-field equations yield exactly the potential metastable equilibria of the network. Interpolating between the solutions on the grid yields a slice of the equilibrium manifold of the system. That is, the manifold of metastable equilibria parametrized by the network parameters such as arrival, departure and migration rates, etc. Since the network is homogeneous and the individual cell distributions are assumed to be identical, an equilibrium (as well as any other state) of the network is described by a pair of numbers corresponding to the expected number of users of each service class per cell. The full equilibrium manifold thus has dimension twice that of the parameter space which is hard to visualize so instead we plot its projection on to one of the service classes.

The parameters of the model can be roughly subdivided into two groups — fixed and variable parameters. Fixed parameters such as cell capacity, service class demands etc. describe internal and relatively static properties of the network. On the other hand, variable parameters, e.g. traffic load, vary over time and can to some extent be regulated by the operator. Since in practice network behavior under a shifting traffic load is of most immediate concern, we choose the service class loads as axes of the phase diagram, i.e. loads from the two service classes as fractions of the total cell capacity, $\rho_1 = \lambda_1/\mu_1 \times b_1/C$ and $\rho_2 = \lambda_2/\mu_2 \times b_2/C$.

Figure 1(a) and (b) shows the metastable equilibrium manifold for the first service class. The reason two plots are needed to describe this manifold is that it is not a single valued function of the chosen parameters and so can take on multiple values at a single point. Plot (a) correspond to the upper boundary and plot (b) to the lower boundary of the convex hull of this manifold.

In order to understand the structure of the equilibrium manifold it is helpful to note that the equilibrium converges to that of an infinite server queue in the limit of $\lambda_s/\mu_s \to 0$. *That is, the network system has a unique equilibrium provided the migration is slow relative to the lifespan of the users and/or the load is close to zero.* This is reflected by the single single-valuedness of the equilibrium manifold (Fig. 1(a) and (b), note that the origin is the far right bottom corner). As the load on the system increases the equilibrium manifold develops a fold. Under the vertical projection the fold covers every point of the parameter space below thrice, yielding three possible metastable equilibrium states for the same set of parameter values. The upper and lower branches, plots (a) and (b) in Figure 1 respectively, of the fold correspond to the metastable equilibrium states and the middle branch (not shown) to the unstable equilibrium state. The phase diagram in Figure 1 (c) was obtained by averaging the two plots (a) and (b), which colors the metastable regions in shades intermediate between those of the neighboring phases. Boundaries of the folds under projection (outlined in black) divide the parameter space into regions with distinct equilibria or phases. Between these lines (corresponding to regions below the folds) neighboring phases coexist giving rise to metastability. For unusually high loads folds can pile up on top of each other producing regions with three or more metastable states.

Computer simulations of the 7×7 cell network arranged in a hexagonal lattice with periodic boundary conditions confirm the qualitative and quantitative correctness of the mean-field approximation. Figure 2 shows plots of the simulated average numbers of users per cell at two points of the parameter space — one non-metastable and one metastable. Figure 2(a) gives the time series for a non-metastable (according to mean-field

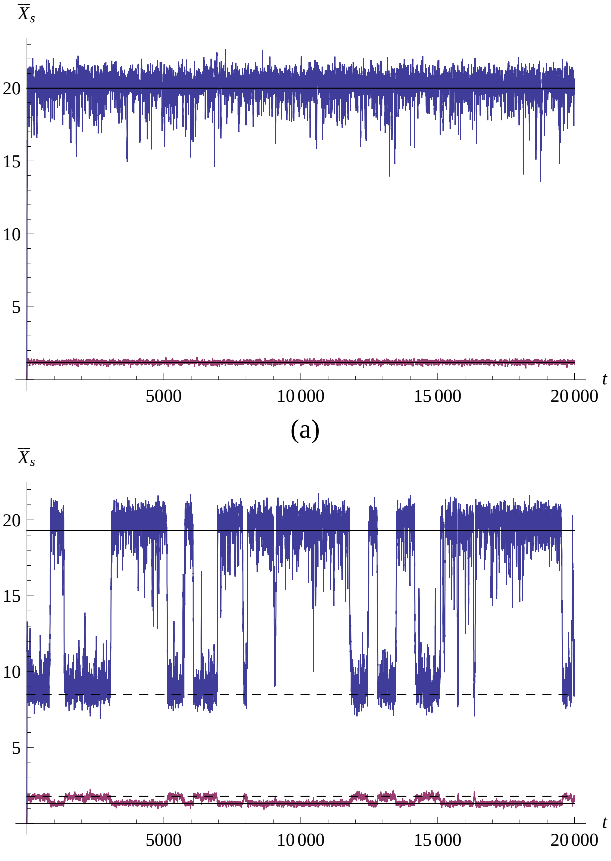


(a)

(b)

(c)

Fig. 1. (a) & (b) Convex hulls of the lower and upper branches respectively of the equilibrium manifold for service class 1. (c) Phase diagram constructed by projecting the equilibrium manifold by averaging. Regions with phase coexistence are outlined in black.($C = 64$, $b_1 = 1$, $b_2 = 18$, $\mu_1 = 1$, $\mu_2 = .5$, $\gamma_1 = 64$, $\gamma_2 = 1$.)



(a)

(b)

Fig. 2. Network parameters are as in Figure 1 (a)$\rho_1 = 0.5$, $\rho_2 = 1.0$ (b)$\rho_1 = 0.5$, $\rho_2 = 1.35$. The horizontal lines indicate equilibrium states predicted by the mean-field approximation.

approximation) parameter vector. As expected the network remained in the steady state predicted by the mean-field approximation for the duration of the simulation. Figure 2(b) shows the time series for a metastable (according to mean-field approximation) parameter vector. Here the network spent substantial amounts of time in the two metastable states, predicted by the mean-field approximation, with rapid transitions between the two states typical of metastable systems. While no substitute for a regular validation procedure, repeated comparisons against simulation runs for different parameter vectors yielded the same good agreement with the mean-field approximation.

We note that analysis of simulation data also confirms the correctness of formula (7). Figure 3 shows the fraction of the users successfully serviced in a sliding 50 time unit window and the mean-field predicted values of the service completion probabilities Q_s.

IV. Implications of Metastability for Performance

The operator can control the load of each service class on the network through admission control. Resource reservation method of admission control has been shown to be effective in keeping the network out of the multi-phase regions in [1],

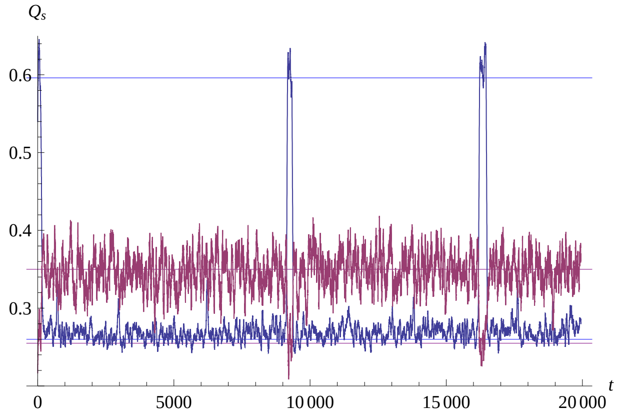


Fig. 3. Mean-field predicted and simulated service completion probability Q_s for network parameters as in Figure 1.(a) and $\rho_1 = .5C$, $\rho_2 = 1.5C$

but it can also simultaneously be used to maximize revenue by steering the network toward the maximum of ϕ. The time-dependent problem of revenue maximization is thus to find $\rho^* = \{\rho^*_{sj} : s \in S, j \in J\}$ solving the constrained maximization problem

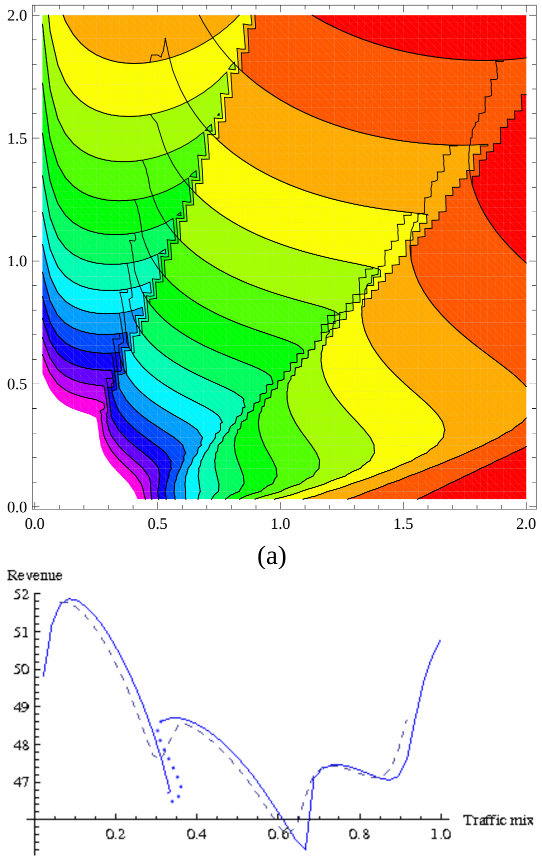
$$\begin{aligned} &\text{Maximize } \phi(r) && (8)\\ &\text{Constraint : } r_{sj} \leq \rho_{sj},\ \sum_s r_{sj} \leq (1+o_j)C_j\ \forall\ j \in J \end{aligned}$$

where ρ_{sj} are the current traffic loads, and o_j is the maximum permitted overload margin for cell j.

Figure 4(a) shows the level sets of ϕ for a spatially homogeneous network with two service classes, and with $p_1 = 1$ and $p_2 = 20$ (the network parameters are as in Figure 1). Because, the equilibrium manifold is multivalued as a function of ρ_1 and ρ_2, ϕ also takes multiple values in the regions with coexisting phases. The figure thus contains two sets of level curves corresponding to the upper and lower branches of the equilibrium manifold, which coincide on the single phase region. The next figure (Figure 4(b)) shows the restriction of the graph of ϕ to the line $\rho_1 + \rho_2 = 1.5$, which is the boundary of the constraint region if ρ_1 and ρ_2 are greater than 1.5 and $o_j = 0.5$. From Figure 4(a) it is clear that the maximum of the optimization problem (8) will be achieved along this line.

Notice that while the extremes in Figure 4(b), corresponding to the exclusion of one or the other of the service classes, are global maxima of the revenue function they may not be acceptable, since they correspond to exceptionally poor quality of service for the excluded service class. Recall that by equation (7) likelihood of successful service is proportional to the average number of service class users in the system.

One crucial observation that can be gleaned from Figure 4(b) is that some local solutions of the above optimization problem may be "cliff hangers", in the sense that they are located close to the phase transition boundary, where a small perturbation in the network state can precipitate a dramatic drop in revenue. Since in practice the network state cannot be perfectly controlled and small fluctuations are inevitable, the



(a)

(b)

Fig. 4. (a) Level curves of ϕ for $p_1 = 1$ and $p_2 = 20$ and network parameters as in Figure 1 (Sawtooth edges are numerical artifacts). (b) Restriction of the revenue manifold to $\rho_1 + \rho_2 = 1.5$ with the x-axis showing the fraction of class 1 users (the dotted portion indicates values for the unstable equilibrium not observed in practice). Dashed graph shows the convolution of the upper branch of the revenue manifold with a Gaussian kernel $g(x) = 1/\sqrt{2\pi}e^{-x^2/2}$ simulating small deviations .

expected revenue of a “cliff hanger” state may be substantially smaller than predicted by the simple minded model. A more accurate picture of revenue may be given by convolving the above graph with a Gaussian kernel, simulating small deviations from the equilibrium. The maxima of ϕ shift, relative to the original graph, when small deviations are introduced (dashed graph in Figure 4(b)).

For practical applications this means that stability margins of revenue maximizing states must be considered carefully when considering optimal control of such service networks. Admission control may not be of much help in improving stability margins significantly because transitions between phases may occur due to a sudden mass exodus of users from the system (perhaps through cell failure) just as well as by a sudden surge of users into the system. Furthermore, once the network enters a new phase it may be difficult to return it to the previous one without at least a temporary decrease in the quality of service. A careful analysis of revenue loss due to system “resets”, following such unwanted phase transitions, versus their likelihood must be carried out in order to determine the true revenue maximizing states.

V. Conclusion

We used the mean-field approximation to derive equations for the metastable states of a large-scale loss network. Using numerical solutions of these equations we constructed a phase diagram for a homogeneous network with two service classes and verified its correctness by running computer simulations for representative parameter values. We observed that metastability arises through a folding of the equilibrium manifold when the traffic load approaches network capacity; at higher traffic loads the load-space is divided into regions with distinct phases that overlap along wedge-shaped metastable regions. Mean-field approximation was then used to derive a formula describing likelihood of successful service and this too was verified by computer simulations. Finally, we observe that states maximizing a linear revenue function may have rather small stability margins, leading to decreased profits due to undesirable phase transitions caused by inevitable chance fluctuations in the network state.

Metastability could potentially be harnessed and put to good use if a reliable way to compute stability margins of the metastable equilibria is discovered. Call admission control could then be used to maintain the network in a desirable metastable state whose quality of service can be estimated using equation (6). Further research is, also, necessary to understand the relationship of inhomogeneity in user traffic and network parameters, and of network topology to stability.

References

[1] Nelson Antunes, Christine Fricker, Philippe Robert, and Danielle Tibi. Metastability of cdma cellular systems. In *MobiCom ’06: Proceedings of the 12th annual international conference on Mobile computing and networking*, pages 206–214, New York, NY, USA, 2006. ACM.

[2] Nelson Antunes, Christine Fricker, Philippe Robert, and Danielle Tibi. Stochastic networks with multiple stable points. *ArXiv:math.PR/0601296*, 2006.

[3] Kelly Frank. *Reversibility and Stochastic Networks*. Wiley, Chichsester, 1994.