Title

A performance bound on dynamic channel allocation in cellular systems: equal load

Permalink
https://escholarship.org/uc/item/09m1f56n

Journal
IEEE Transactions on Vehicular Technology, 43(2)

Authors
Jordan, Scott
Khan, Asad

Publication Date
1994-05-01

DOI
10.1109/25.293652

Peer reviewed

A Performance Bound on Dynamic Channel Allocation in Cellular Systems: Equal Load

Scott Jordan, *Member, IEEE,* and Asad Khan, *Student Member, IEEE*

***Abstract*— We model a cellular network as a more general multiple service, multiple resource system. We define the "state" of the system as the number of calls currently carried in each cell. We restrict ourselves to channel allocation policies that place restrictions on the global state of the system, are allowed immediate global channel reallocation, and ignore handoffs. Maximum packing and fixed allocation are considered as special cases of such policies. Under uniform load conditions, we prove that throughput is increasing and concave with respect to increases in load or capacity, under maximum packing or fixed allocation. We propose that the optimal policy, in the considered class, varies from maximum packing at low loads to fixed allocation at high loads. This policy is often impractical to implement, but can be considered as a performance bound on practical systems. The analytical results are investigated numerically using a simple seven cell linear network.**

I. Introduction

WIRELESS services are one of the strongest growth areas in telecommunications today. Cellular voice service is well established as a high-end service in most areas, but demand is increasing at 30% per year. Personal communications services (PCS) are expected to be introduced in the next few years as a mass market phone service. Wireless data services are appearing in the form of wireless LANs and modems. Capacity, however, is now a critical issue for all of these services.

Carriers are considering cell splitting, allocation of new spectrum, alternative multiple access architectures and dynamic channel allocation, as possible methods to increase capacity. We focus in this paper on channel allocation techniques. A great deal of recent research has proposed and investigated a broad range of channel allocation schemes. See [21] for a good overview. These schemes are based on the following concepts:

•*Permanent Channel Assignment:* Some of the channels are permanently assigned to each given cell, in accordance with frequency reuse constraints. The current cellular system permanently assigns all channels. The lack of sharing between nearby cells results in a lack of efficiency that inspired all other techniques.

•*Channel Borrowing by Request:* Some of the channels may be borrowed from an adjacent cell when all of the permanent channels are occupied.

•*Adjustment of Parameters according to load:* The number of permanent channels and/or borrowable channels may be reassigned periodically according to spatial inequities in load.

By combining these concepts in different ways, a large number of dynamic channel allocation schemes are possible. The current system, *fixed channel allocation* (FCA), permanently assigns all channels. *Simple channel borrowing* schemes [2], [23], incorporate the first types of sharing, but differ according to channel locking methods. *Hybrid strategies* [9] assign some permanent channels, and allow the rest to be borrowed. *Borrowing with channel ordering* suggests a hybrid scheme wherein the ratio of fixed to borrowable channels varies according to load. All of these methods have variants which add some channel reordering to reduce inefficiency at high load. *Flexible strategies* [20] permanently allocate some channels, but reserve the rest in a central pool to be assigned to needy cells on a temporary basis. They differ according to the times at which and the basis on which additional channels are assigned. More dynamic assignment strategies [1], [3], [23], assign channels for the duration of an individual call only, on the basis of a calculated cost function which may include many and varied factors.

Most of the studies on dynamic channel allocation are based on simulation models; in contrast, theoretical studies number relatively few. Kelly [10], [11] studies benefits of "maximum packing" (MP) dynamic allocation over fixed allocation, providing a capacity upper bound for some schemes. Hajek [4] bounds blocking probabilities for a similar system. Sivarajan [19] derives a "Shannon type bound" for a single service class, and Xu [22] studies a particular hybrid allocation. All of these studies ignore handoffs entirely. Kumar *et al.* [13] compare dynamic and fixed allocation using the notion of stochastic dominance, and in incorporating handoffs, they also derive conditions for which dynamic schemes perform better, for the case of uniform traffic and well defined cells.

Many claims have been presented as to the advantage of one scheme over another. It is as yet unclear to what extent and in what circumstances each scheme increases capacity of the system. See [16], [17], [18], for attempts to bound and compare a few schemes.

Questions remain. How does each dynamic channel allocation scheme produce its capacity gains? What are the basic trade-offs that are occuring? Why do some only work well

Manuscript received February 24, 1993; revised March 24, 1993 and July 1, 1993. This work was supported by the National Science Foundation under Grant NCR-9109223 and by the Ameritech Foundation.

The authors are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208-3118 USA.

IEEE Log Number 9213790.

under certain traffic patterns? Can they be combined? What is the value of additional information about the state of nearby cells? What is the best possible use of bandwidth? What will channel allocation look like for integrated services?

In this paper, we attempt to provide some insight into some of these issues by considering a class of channel allocation policies that includes MP and FCA. To do so we embed the cellular channel allocation problem within a more general Multiple Service, Multiple Resource (MSMR) model. This unifies all dynamic channel allocation strategies that use comparable "state" information into one context in which they can be compared. The restrictions of this model, however, are significant. Immediate global channel reassignment is assumed, and handoffs are ignored. *The resulting optimal allocation strategy for this model, therefore, can only be considered to be an upper bound for practical systems.*

We show that MP and FCA result in increasing and concave throughput with respect to increasing load or capacity, in any symmetric cellular network under equal loads. We propose that the optimal policy in this class progresses from MP at low loads to FCA at high loads. We then consider a simple seven cell linear network in order to numerically investigate competition for channels between neighboring cells and the variation of the optimal policy in this class with load.

In Sections II and III we introduce a cellular allocation model and the underlying multiple service, multiple resource system model. Section IV analyzes the variation of throughput under maximum packing, fixed and optimal allocations for this model with changes in load or capacity. In Sections V and VI, we discuss the effect of a cell upon its neighbors, and use this to show how the optimal control for this model progresses from maximum packing toward fixed allocation as the load increases in a homogeneous traffic scenario, in our simple seven cell system.

II. A Cellular Channel Allocation Model

The channel allocation scheme used in the current cellular system splits the available spectrum into several (assumed here to be 7) segments, and the geographical service region into hexagonal cells. Each cell is assigned 1 of the 7 segments in a manner in which cells assigned the same segment are several cell diameters apart. Each segment is then split into smaller frequency slots called channels, each large enough to accommodate one phone call.

Two types of interference can occur. First, calls placed in two adjacent channels can interfere with one another. This is solved by using a guard band in each channel and filters on the signal, at the cost of additional complexity and bandwidth.

Second, calls placed on the same channel in nearby cells can interfere with one another. The magnitude of the interference is determined by power level and distance between the two customers sharing the same channel. In the current system, the segmenting process insures a minimum distance between co-channel customers sufficient to guarantee low interference. Cell splitting, as mentioned above, results in a greater efficiency by lowering the power level and shortening the permissable re-use distance. This has a practical limit however.

This segmenting system is overly restrictive and hence wasteful of capacity. Given a fixed cell size and power level, the basic frequency reuse constraint is:

- No channel can be used in cells closer than some specified distance.

Segmenting, however, can result in a call being blocked when all channels dedicated to the originating cell are busy *even though other channels are available within the frequency reuse area.* Every dynamic channel allocation scheme, each in its own way, relaxes the strict segmenting rules in an attempt to allow nearby cells to share the available bandwidth in a manner that is efficient and that satisfies the interference constraints. The gain of each scheme is the increased utilization of allocated bandwidth. The cost is increased complexity and the requirement of more global information at call start-up or handoff times.

Each scheme's gains (or losses) compared to the existing fixed channel allocation rule depend upon its particular mechanism and thus upon the spatial traffic demand. They are therefore hard to compare. Similarly, since each scheme requires different information from nearby cells, and uses this information in different ways, the costs of each scheme are hard to compare.

We can construct a model, however, that encompasses many channel schemes in a common framework. In general, we need to keep track of which channels are occupied in each cell in the network, and the age of each call. If we assume, however, that immediate global channel reassignment is possible to satisfy co-channel interference constraints, we can simply keep track of the total number of calls in each cell, rather than the specific channel used [3]. In addition, if we assume that call durations are exponentially distributed, the ages of each call are not required.

We therefore define the state of the system as $x = (x_1, \cdots, x_N)$, where N is the total number of cells in the system and x_i is the total number of active calls in cell i. The basic frequency reuse constraints impose restrictions upon the values the state x can take on, and hence define a state space "Z". For an appropriate choice of the minimum reuse distance, these constraints for the state space Z are

$$\sum_{j \in C_i} x_j \leq M \qquad \forall C_i \tag{1}$$

where C_i are overlapping clusters over the state space Z and M is the maximum number of channels available to each cluster [3].

The Maximum Packing strategy accepts a call request if and only if there exists a global reassignment of the existing calls and the new call to satisfy the basic frequency reuse constraints. Since this policy shares all channels among all cells, we will alternatively call this policy Complete Sharing (CS) to stress the resource usage. MP is equivalent to accepting a call if and only if the resulting state remains in the state space $Z_{CS} = Z$ as defined in (1) [3].

Fixed channel allocation, on the other hand, accepts a call request if and only if there is a free channel among the subset

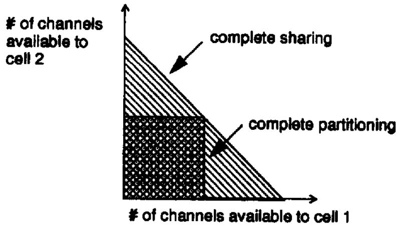
# of channels available to cell 2
complete sharing
complete partitioning
# of channels available to cell 1

Fig. 1. The state space Z for CS and CP policies.

of channels permanently assigned to the corresponding cell. Since this policy permanently divides all channels among cells in a cluster, we will alternatively call this policy Complete Partitioning (CP) to stress the resource usage. FCA is equivalent to accepting a call if and only if the resulting state both remains in the state space Z as defined in (1), and satisfies

$$x_i \leq \frac{M}{C} \qquad \forall i \tag{2}$$

where C is the number of cells per cluster. Equation (2) defines a reduced state space $Z_{CP} \subset Z_{CS}$. (See Fig. 1.)

In this paper we restrict ourselves to a class of channel allocation policies that admit a new call if and only if the resulting state would be in a state space $Y \subset Z$. The particular policy is given by the definition of Y. By necessity, Y must be coordinate convex, i.e., if $x \in Y \,\&\, x_i \geq 1$, then $(x_1, \cdots, x_i - 1, \cdots, x_N) \in Y$. Furthermore, all such policies lie in between CP and CS, i.e., $Z_{CP} \subseteq Y \subseteq Z_{CS}$.

We assume that immediate global channel reassignment is achievable for any policy in this class, and do not consider the computational burden imposed. *Since this burden is overwhelming for any practical system, the optimal channel allocation policy in the considered class is only an upper bound for such systems.* Furthermore, there exist policies that base the decision to accept a new call upon not only the resulting state, but also upon the *current* state, or upon the particular channels occupied. These policies are not in the class considered.

By modeling each call attempt in each cell as a service request, and the channels available as a pool of resources, we can view this cellular network as a specific example of a Multiple Service, Multiple Resource (MSMR) system [5]. The cellular system is modeled as a multidimensional time reversible Markov chain in which the state is the number of calls in progress in each cell.

The strengths of this model are that both basic frequency reuse constraints and any additional dynamic channel allocation constraints are incorporated in a unifed manner. Therefore, competing strategies can be equitably compared, and the differences between them easily understood. The optimal control policy gives us both an exact upper bound on the maximum achievable throughput of policies in the considered class and insight into how increased performance is gained.

The principal weakness of this model is that it ignores handoffs. This is necessary to achieve a tractable form for the stationary distribution and for the optimal control. Computational considerations limit the size of the state space for which we can easily calculate optimal policies under specific loads; this is thus a significant secondary weakness of the model.

In the next few sections, we use this model to compare CS (maximum packing), CP (fixed), and optimal control strategies in the considered class. First we review the general MSMR model upon which our cellular model is based. Then we investigate what can be analytically stated about these policies in the most general case. Finally, we numerically compare these 3 strategies for a simple celluar system.

III. The MSMR Model

In this section, we will review the MSMR model and some relevant MSMR results [5]–[8].

Model: Consider a system that offers n types of services. Each service requires a set of resources (dependent upon the service type) to process. If these resources are available then the system manager may choose to accept a service request, and then processing starts immediately. If the necessary resources are unavailable, of if the system manager denies the request, then the request is lost to the system.

Service requests arrive as independent Poisson processes. Each request occupies each resource that it needs for the same amount of time, and releases these resources simultaneously upon service completion. This amount of time is exponentially distributed, and independent of other service times.

We model this system as a Markov chain. Adopt the following notation

$\lambda \equiv (\lambda_1, \cdots, \lambda_n)$,	the rates of incoming service requests.
$\mu \equiv (\mu_1, \cdots, \mu_n)$,	the rates of service.
$\rho \equiv (\rho_1, \cdots, \rho_n)$,	the loads, given by $\rho_i = \lambda_i/\mu_i$.
$L \equiv (L_1, \cdots, L_n)$,	the rates of *accepted* service requests (throughput).
x_i,	the number of type i requests being processed.
$x \equiv (x_1, \cdots, x_n)$,	the state of the system.
$Z \equiv \{x \mid x\text{is feasible,}$ i.e., x can be simultaneously processed with available resources$\}$,	the state space.
$F_i \equiv \{x \mid x \in Z\,\text{but}\,(x_1, \cdots, x_i + 1, \cdots, x_n) \notin Z\}$	the full set w.r.t. service type i.
$E_i \equiv \{x \mid x \in Z\,\text{but}\,(x_1, \cdots, x_i - 1, \cdots, x_n) \notin Z\}$,	the empty set w.r.t. service type i.
$\pi(x)$,	the steady state probabilities.
$S \equiv \{1, \cdots, n\}$,	the set of all service types.
$x \uparrow I \equiv (y_j)$, $y_j = \begin{cases} x_j, & j \notin I \\ 0, & j \in I \end{cases}$ for$I \subset S$	a projection function.
$C_Z(x \uparrow I) \equiv \{y \in Z \mid y \uparrow I = x \uparrow I\}$,	a $\lvert I\rvert$ dimensional cross section of Z.
ri,	revenue generated by servicing request type i, per unit of time.

Our assumptions regarding the arrival and departure processes give us a Markov chain on state space Z with transition rates

$$r_{xy} = \begin{cases} \lambda_i, & \text{if } x \notin F_i \text{ and } y = (x_1, \cdots, x_i + 1, \cdots, x_n) \\ x_i u_i, & \text{if } x \notin E_i \text{ and } y = (x_1, \cdots, x_i - 1, \cdots, x_n) \\ 0, & \text{else.} \end{cases}$$

Assume that service completion is never blocked. This implies that the state space Z is *coordinate convex*, i.e., if $x \in Z \,\&\, x_i \geq 1$, then $(x_1, \cdots, x_i - 1, \cdots, x_n) \in Z$.

The Markov chain is time reversible with stationary distribution

$$\pi(x) = \pi(0) \prod_{i=1}^{n} \frac{\rho_i^{x_i}}{x_i!} \qquad \text{where } \pi(0) \equiv \frac{1}{\displaystyle\sum_{x \in Z} \prod_{i=1}^{n} \frac{\rho_i^{x_i}}{x_i!}}. \quad (3)$$

Our measure of performance is average total throughput

$$R = \sum_i \mu_i E X_i$$

We note that under equal load conditions maximizing total throughput is equivalent to minimizing the blocking probability

$$R = \sum_i L_i = \sum_i \lambda_i [1 - P(B_i)].$$

When not explicit, we will write R_Z to indicate that the average throughput is to be taken over the state space Z. The first relevant result describes interaction between services sharing resources [5]:

Theorem 3.1: The sensitivity of throughput to arrival rate is given by

$$\frac{\partial L_i}{\partial \lambda_j} = \begin{cases} \dfrac{\mu_i}{\lambda_j} \operatorname{cov}(x_i, x_j), & \text{if } i \neq j \\ \dfrac{\mu_i}{\lambda_i} \operatorname{var}(x_i), & \text{if } i = j. \end{cases}$$

Pairs of services can be classified as complements of substitutes, according to the sign of the associated covariance. The sign of the covariance, in turn, is affected by the resources in common between the services *and* by other services that compete with the pair in question. In the next section, we will use this to investigate contention between neighboring cells for channels.

The second relevant result describes the optimal policy. We consider policies that correspond to *restricting the state to some subset of the original space*. We have only a weak characterization of the optimal c.c. policy for the Markov chain model (3) [6]. A much stronger characterization can be obtained by approximating the discrete product form distribution (3) by a continuous product form distribution (4)

$$\pi(x) \equiv K \prod_{i=1}^{n} f_i(x_i) \qquad f_i(x_i) \text{ continuous on } x_i \geq 0. \quad (4)$$

The state space is now a continuous region given by $Ax \leq b$. The state is x, a vector of length n of nonnegative real numbers; A is a $m \times n$ matrix of nonnegative real numbers representing resource usage for each service type; b is a vector of length m of positive real numbers representing total numbers of each resource type.

We require that each additional resource added to the system produces a nonnegative but decreasing return to the optimal revenue. This is equivalent to the following concavity property on parallel cross sections of the optimal subset of the state space [7]:

Property P3: For any two cross sections C_1 and C_2,

$$C_1 \equiv C_Z(x \uparrow I) \,\&\, C_2 \equiv C_Z((x + \beta e_j) \uparrow I), \quad j \notin I,$$

$$R_{[\alpha C_1 + (1-\alpha) C_2]} \geq \alpha R_{[C_1]} + (1 - \alpha) R_{[C_2]} \quad \forall i \in I$$

where

$$\alpha C_1 + (1 - \alpha) C_2 \equiv C_Z((x + (1 - \alpha)\beta e_j) \uparrow I).$$

The optimal control policy is given by a particular convex subset of the state space. The form of this subset is described in the next theorem:

Theorem 3.2: If X has a distribution on $Ax \leq b$ given by (4) satisfying $P3$, then the optimal c.c. subset $Z^* \subseteq Z$ is defined by

$$x \in Z^* \text{ if } x \in Z \,\&\, x \uparrow I \in G_I^* \,\forall I \subset S.$$

The regions G_I^* tell when to deny service requests and are further characterized in [7]. Note that optimal policy for the continuous model is convex, whereas the optimal policy for the discrete model may not be. This allows a simpler approach to algorithmically find the optimal policy for (4), and although the optimal policies for (3) and (4) may differ greatly, the difference in the throughputs they produce are guaranteed to be small [8].

IV. Form of Optimal Policies and Throughput

In this section, we analytically compare CS, CP, and optimal policies for the model presented in Section II. We are particularly interested in the performance of each at different loads. We restrict ourselves to the case where each cell contributes an equal load to the system. Denote the total number of channels by M, the number of cells by N, and the cluster size by C.

The first 4 theorems characterize the variation in performance of CS and CP with respect to capacity or load.

Theorem 4.1: The total throughput under a complete partitioning (fixed allocation) policy is increasing and concave with respect to increases in capacity.

Proof: Under a fixed allocation, the total number of available channels, M, is divided equally among cells in a cluster, with each cell allowed access to $n = M/C$ channels. Each cell now acts as a $M/M/n/n$ server. The throughput of this system has been shown to be increasing and concave with respect to increases in capacity [14]. Therefore the total throughput is similarly increasing and concave with respect to increases in capacity.

Theorem 4.2: The total throughput under a complete partitioning (fixed allocation) policy is increasing and concave with respect to increases in equal cell loads. The minimum throughput is zero, and the supremum is $\mu MN/C$.

Proof: The throughput of a $M/M/n/n$ server has been proven to be increasing and concave with respect to increases in load in [12]. Therefore the total throughput is similarly increasing and concave with respect to increases in equal cell loads. By the Erlang-B blocking formula, the throughput at 0 load is 0, and the supremum is $\mu n = \mu M/C$ per cell.

Theorem 4.3: The total throughput under a complete sharing (maximum packing) policy, on a symmetric cellular system, is increasing and concave with respect to increases in capacity.

Proof: Denote the set of channels in a symmetric cellular system as CH_0. Define L_M to be the total throughput under a complete sharing policy in such a system with M channels. Compare this system to one with $M+1$ channels per overlapping cluster. Denote the set of incremental channels as CH_1, and without loss of generality assume that calls are assigned to channels in CH_1 if and only if no channels are available in $CH_0. L_{M+1} - L_M$ is thus equal to the throughput achieved by the incremental channels CH_1. These channels, however, just see as arrivals the overflow from CH_0. From these call requests, CH_1 accepts as many calls as possible, and let the rest go as its own overflow. Since CH_1 serves *some* arrivals, $L_{M+1} - L_M$ is positive and thus L_M is increasing in M; and also the overflow from CH_1 contains fewer arrivals than the overflow from CH_0.

Similarly, compare the system with $M+1$ channels to the system with $M+2$ channels. Denote the set of new incremental channels by $CH_2. L_{M+2} - L_{M+1}$ is equal to the throughput achieved by CH_2. These channels, however, just see as arrivals the overflow from CH_1, which as we stated above contains fewer arrivals than the overflow from CH_0. From this overflow, CH_2 accepts as many calls as possible, but this must be on average fewer than CH_1 accept, due to the independence of arrivals. Therefore $L_{M+2} - L_{M+1} < L_{M+1} - L_M$, and L_M is concave in M.

Theorem 4.4: The total throughput under a complete sharing (maximum packing) policy, on a symmetric cellular system, in increasing and concave with respect to increases in equal cell loads. The minimum throughput is zero, and the supremum is $\mu MN/C$.

Proof: Consider a cellular system presented with equal cell loads of ρ, under a maximum packing strategy. Denote the resulting throughput by L_ρ. Compare this system to the system resulting from adding in infinitesimal load of $\delta\rho$ to each cell. Denote the resulting throughout by $L_{\rho+\delta\rho}$.

Since call requests arrive to each cell as independent Poisson processes, we can, without loss of generality, consider this incremental load to be presented by Poisson processes of rate $\delta\lambda = (\delta\rho)\mu$ to each cell, where these processes are independent both of each other and of the original load. Denote the Poisson processes corresponding to the original load by P_0 and those corresponding to the incremental load by P_1.

We wish to examine in detail the incremental throughput obtained through the incremental load. Consider a realization of the cellular system with load $\rho, X_\rho(t)$. Denote the corresponding realization for the system with load $\rho + \delta\rho$ by $X_{\rho+\delta\rho}(t)$. Construct a new realization $X'_{\rho+\delta\rho}(t)$ as follows. Start with $X_\rho(t)$, and add in each arrival in P_1 *that does not prevent a future arrival in* P_0. Note that $X'_{\rho+\delta\rho}(t)$ is thus composed of $X_\rho(t)$ plus additional calls dropped into open time slots, and that these additional calls correspond to the incremental throughput.

Compare$X_{\rho+\delta\rho}(t)$ and $X'_{\rho+\delta\rho}(t)$. Consider the first arrival in P_1 accepted in $X_{\rho+\delta\rho}(t)$. If the call completes before blocking a call from P_0, then it is present in both realizations. If the call is still resident when a call in P_0 requests service, and if the new call can only be accepted by terminating the old call, then the two realizations differ: $X_{\rho+\delta\rho}(t)$ contains the old call whereas $X'_{\rho+\delta\rho}(t)$ contains the new call.

We wish to examine the difference in total throughput resulting from this decision. Fix the cell, i, of the incremental call in question and fix the time s, of the conflict with the new call. Consider all such realization pairs that conflict at time s over acceptance of an incremental call in cell i. Denote by $Y_{s,i}$ those realizations that admit the incremental call, and by $Y'_{s,i}$ those that admit the new call instead. Call durations are memoryless, therefore all realization will contain the corresponding call they admit for the same expected remaining time after s. Furthermore, cell loads are equal on a symmetric cellular space, therefore the expected number of future calls accepted in $Y_{s,i}$ equals the expected number of future calls accepted in $Y'_{s,i}$. We conclude that, on average, $X_{\rho+\delta\rho}(t)$ and $X'_{\rho+\delta\rho}(t)$ obtain the same throughput $L_{\rho+\delta\rho}$.

We can therefore consider the incremental throughput $L_{\rho+\delta\rho} - L_\rho$ as due to calls in $X'_{\rho+\delta\rho}(t)$ not present in $X_\rho(t)$. Since some calls from the incremental load are accepted in $X'_{\rho+\delta\rho}(t), L_{\rho+\delta\rho} - L_\rho$ is positive and, thus, L_ρ is increasing in ρ. We also note that in each realization the open time slots in $X'_{\rho+\delta\rho}(t)$ are fewer than in the corresponding $X_\rho(t)$.

Similarly, now construct corresponding realizations $X'_{\rho+\delta\rho}(t)$ by adding additional incremental Poisson processes, P_2, of rate $d\lambda$ to each $X'_{\rho+\delta\rho}(t)$. Examine the corresponding incremental achieved throughput $L_{\rho+2\delta\rho} - L_{\rho+\delta\rho}$. This additional throughput is the result of calls in P_2 accepted into open time slots in $X'_{\rho+\delta\rho}(t)$. These open time slots, however, are, as we stated above, fewer in number than those present in $X_\rho(t)$. Since all arrivals in P_1 and P_2 are independent from each other and from the system state, those from P_2 accepted in $X'_{\rho+2\delta\rho}(t)$ must on average be fewer than those from P_1 accepted in $X'_{\rho+\delta\rho}(t)$. Therefore, $L_{\rho+2\delta\rho} - L_{\rho+\delta\rho} < L_{\rho+\delta\rho} - L_\rho$, and L_ρ is concave in ρ.

The range of achievable throughputs is

$$\left\{\sum_{l} \mu_i x_i, x \in Z\right\}$$

and the minimum is 0 at zero load, and the supremum is $\mu MN/C$, obtained as the load tends to infinity [5].

The next theorem compares CS and CP at different loads.

Theorem 4.5: Consider a symmetric cellular system under equal loads. At low loads, the total throughput under complete sharing is *higher* than under complete partitioning. At high loads, the total throughput under completer sharing is *lower* than under complete partitioning. Pursuant to theorems 4.2 and 4.4, there therefore exists a unique crossover point of the two throughput versus load curves.

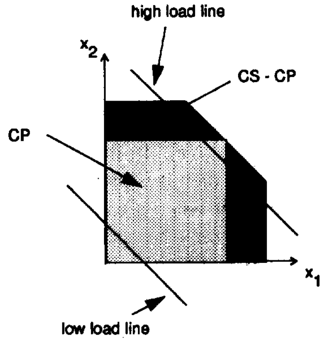

Fig. 2. Pictorial explaining the difference between CS and CP.

Proof: Denote Z_{CS} as the state space corresponding to the maximum packing strategy, and $Z_{CP} \subset Z_{CS}$ as the state space corresponding to the fixed allocation strategy. Denote R_{CS} and R_{CP} as the corresponding total throughputs, and R_{CS-CP} as the average total throughput on the set of states $Z_{CS} - Z_{CP}$. Since the Markov chain is time reversible, $R_{CS} > R_{CP}$ if and only if $R_{CS-CP} > R_{CP}$ or equivalently if $R_{CS} > R_{CS-CP}$. A two dimensional pictorial[1] is shown in Fig. 2.

The load line is given by $\mu \Sigma x_i = R$, and hence passes through the points at which the instantaneous total throughput equals the average total throughput of the system. The expected throughput while in the region Z_{CS-CP} is greater than the expected throughput while in $Z_{CP}(R_{CS-CP} > R_{CP})$ if and only if the expected point on the region Z_{CS-CP} is above the load line $\mu \Sigma x_i = R_{CP}$.

At low loads, the total average throughput per cell, under CP, is approximately but strictly less than λ. Therefore,

$$R_{CP} \leq \lambda N.$$

On the region Z_{CS-CP} however, all states have an instantaneous throughput of at least $\mu M/C$. Therefore,

$$R_{CS-CP} \geq \mu \frac{M}{C}.$$

Therefore, for low enough loads $(\lambda < \mu M/CN), R_{CS-CP} > R_{CP}$ and hence $R_{CS} > R_{CP}$.

Under high loads, both CS and CP achieve a total throughput of almost $\mu MN/C$. To compare the two, we investigate the probability density functions of instantaneous throughput for CS and for CP.

For high loads, both density functions achieve a maximum at $\mu MN/C$, and fall off as the instantaneous throughput decreases. Therefore the policy that results in the density function with the higher derivative near $\mu MN/C$ will have the higher *average* throughput.

Both CS and CP achieve their maximum instantaneous throughput at the state $(x_i = M/C \, \forall i)$. Compare the states at which each obtain an instantaneous throughput of $\mu\,(MN/C - 1)$. Under CP, $\{(x_j = M/C - 1, x_i = M/C \;\; \forall \;\; i \neq j)$ for any $j\}$ achieves this. Under CS, not only does $\{(x_i = M/C - 1, x_i = M/C \;\; \forall \;\; i \neq j)$ for any $j\}$ achieve this, but so do other states such as $(x_j = M/C - 1, x_k = M/C - 1, x_1 = M/C + 1, x_i = M/C \;\; \forall \;\; i \neq j, k, l)$ where $(j, k, l) \subseteq C_m$ for some m.

Since the Markov chain is time reversible, this implies that

$$\frac{P\left(\text{instantaneous throughput} = \mu \frac{MN}{C}\right)}{P\left(\text{instantaneous throughput} = \mu\left(\frac{MN}{C} - 1\right)\right)}$$

is greater under CP than under CS.

Therefore, under sufficiently high load, $R_{CS} < R_{CP}$.

Now since both CS and CP policies start out with zero throughput at zero load, and both total throughput versus load curves are increasing and concave, these results imply that there exists a unique crossover point λ^*, such that $R_{CS} > R_{CP}$ for $\lambda < \lambda^*$ and $R_{CS} < R_{CP}$ for $\lambda > \lambda^*$.

Examples are shown later in this paper.

At low loads, both policies achieve throughput close to the offered load, but CS obtains a lower blocking probability because the fundamental re-use constraint is farther out on the tail of the density of instantaneous load than the constraint imposed by CP. At high loads, both policies achieve throughput close to the capacity of the cellular system, but CP obtains a lower blocking probability because it more often avoids states where the instantaneous throughput is suboptimal.

At a moderate load, it is natural to ask if it might be valuable to combine these two strategies by reserving some of the channels for each cell, and sharing the remainder among the cluster. Indeed, several policies have been proposed along these lines. In this paper we consider any policy that accepts or denies call requests by restricting the global state of the cellular system.

We would expect that such optimal policies would depend upon the system load. At low loads, the optimal policy should resemble CS; at high loads, it should resemble CP. The gain achieved by the optimal policy over CS and CP should be highest at moderate loads, where a mixture is most desirable.

This result could be formalized if the throughput of the optimal policy were also increasing and concave with respect to capacity and load. However, since the optimal control is imposed upon a discrete state space, the amount of control imposed is a staircase function of capacity or load, and hence concavity is lost.

We therefore instead investigate the variation of throughput with capacity and load for the *continuous approximation* to the discrete space, where the control imposed can be a continuous function. The difference between the optimal throughput in the continuous and discrete cases should be small. If the throughput curves under the optimal policy, complete sharing and complete partitioning are all fairly smooth, then we expect that the gain of the optimal policy over either complete sharing or complete partitioning is positive and unimodal, i.e., the largest gains occur at intermediate loads.

Proposition 4.6: For the continuous state space model, the total throughput under the optimal policies, on a symmetric

[1] Due to the difficulties of capturing the nature of a N dimensional space on a two dimensional page, this pictorial is not a project. It is intended to capture the relevant characteristics.

cellular system, is increasing and concave with respect to increases in capacity.

Reasoning: Consider increasing the capacity from M channels per cluster to $M + 1$ channels per cluster. Denote the corresponding throughputs as L'_M and L'_{M+1}. We are guaranteed that the new optimal policy generates no less revenue than the old optimal policy, $L'_{M+1} \geq L'_M$, since the capability to restrict use of the additional channels is within the class of control policies we are considering. Furthermore, we expect that the symmetric cellular system will obey *decreasing returns,* namely that the additional revenue generated by a marginal increase in capacity will be decreasing.

In fact, we expect that

Property DR1: $0 < L'_{M+1} - L'_M < L'_M - L'_{M-1} < \mu \frac{N}{C}$.

Now consider the variation of this incremental throughput with arrival rate. Since the additional throughput is an increasing function of the overflow, we also expect that:

Property DR2: $P(B_i)$ is an increasing function of the arrival rate.

Property DR3: $L'_M - L'_{M-1}$ is an increasing function of the arrival rate.

Theorem 4.7: For the continuous state space model, if properties DR1, DR2, and DR3 hold, then the total throughput under the optimal policies is increasing and concave with respect to increases in equal cell loads. The minimum throughput is zero, and the supremum is $\mu MN/C$.

Proof: We can relate the sensitivity of throughput to arrival rate, and the sensitivity of throughput to capacity, using a result from [10]

$$\frac{\partial R}{\partial \lambda_i} = (1 - P(B_i))\left(r_i - \frac{\Delta R}{\Delta F_i}\right)$$

where

$$\frac{\Delta R}{\Delta F_i} \equiv R_Z - R_{Z-F_i}.$$

We wish to investigate an incremental increase in the common load. Suppose

$$\lambda_i = \lambda \quad \forall i$$
$$\frac{\Delta R}{\Delta F} \equiv L'_M - L'_{M-1}.$$

Then, using symmetry

$$\frac{\partial R}{\partial \lambda} = \sum_{i=1}^{N} \frac{\partial R}{\partial \lambda_i} = (1 - P\left(r_i - \frac{\Delta R}{\Delta F_i}\right)$$
$$= (1 - P(B))\left(\mu \frac{N}{C} - \frac{\Delta R}{\Delta F}\right).$$

Note that $1 - P(B)$ is positive. Also, by DR1, $\mu(N/C) - (\Delta R/\Delta F)$ is positive. Therefore $(\partial R/\partial \lambda)$ is positive, and thus the total throughput is increasing with load.

Consider the variation of $(\partial R/\partial \lambda)$ with λ. By DR2, $1 - P(B)$ is decreasing as λ increases. By DR3, $\mu(N/C) - (\Delta R/\Delta F)$ is also decreasing as λ increases. Therefore, $(\partial R/\partial \lambda)$ is decreasing as λ increases, namely total throughput is concave with load.

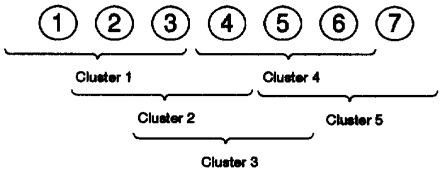

Fig. 3. A 7 cell linear network.

Finally, the range of achievable throughputs is

$$\left\{ \sum_i \mu_i x_i, x \in Z \right\}$$

and the minimum is 0 at zero load, and the supremum is $\mu MN/C$, obtained as the load tends to infinity [5].

In the next two sections, we will examine in greater detail how the optimal policy in the considered class achieves greater throughput than CS or CP, through a numerical example of a simple cellular system.

V. Sensitivity to the State of the Neighborhood

In the last section, we argued that the optimal policy in a symmetric cellular system with equal loads should progress from CS at low loads to CP at high loads, and should achieve its largest gains over these competing strategies at intermediate loads. In this section, we investigate in more detail how gains in throughput are achieved at various loads. To do this, we will investigate the effect of traffic in one cell upon neighboring cells by looking at the sensitivity of throughput in cell i, L_i, to the offered demand, λ_j, in nearby cells.

Sensitivity is a measure of interaction between services sharing resources. Pair of services are either *complements* or *substitutes,* as defined in terms of Theorem 3.1. If $E[X_j|X_i]$ increases with X_i, the covariance is positive. This means that rate of change of throughput of type j service with the increase in arrival rate of type i service is positive. Services i and j in this case are termed complements. Otherwise, if $E[X_j|X_i]$ decreases with X_i, their covariance is negative. This from Theorem 3.1 results in a decrease in throughput of type j service with an increase in rate of request for service i, and i and j are termed substitutes.

We consider a 7 cell linear (one-dimensional) network[2], with a 3 cell reuse cluster (Fig. 3). A complete sharing dynamic allocation scheme is applied.

Load is increased in cell 1 on the left edge and its effect on throughput in all the cells is plotted. Changes in throughput in each cell are shown for various loading conditions (Fig. 4).

Increasing the load of one service increases its throughput, which we know from Theorem 3.1, since cov $(X_i, X_i) = \text{var}(X_i) \geq 0$.

Thus the throughput of cell 1 increases with the increase in its request rate. All other services whose throughput increases are cell 1's complements while those cells whose throughput decreases are the substitutes of cell 1.

[2] The number of cells is kept small due to computational considerations in the optimal policy algorithms used in the next section.

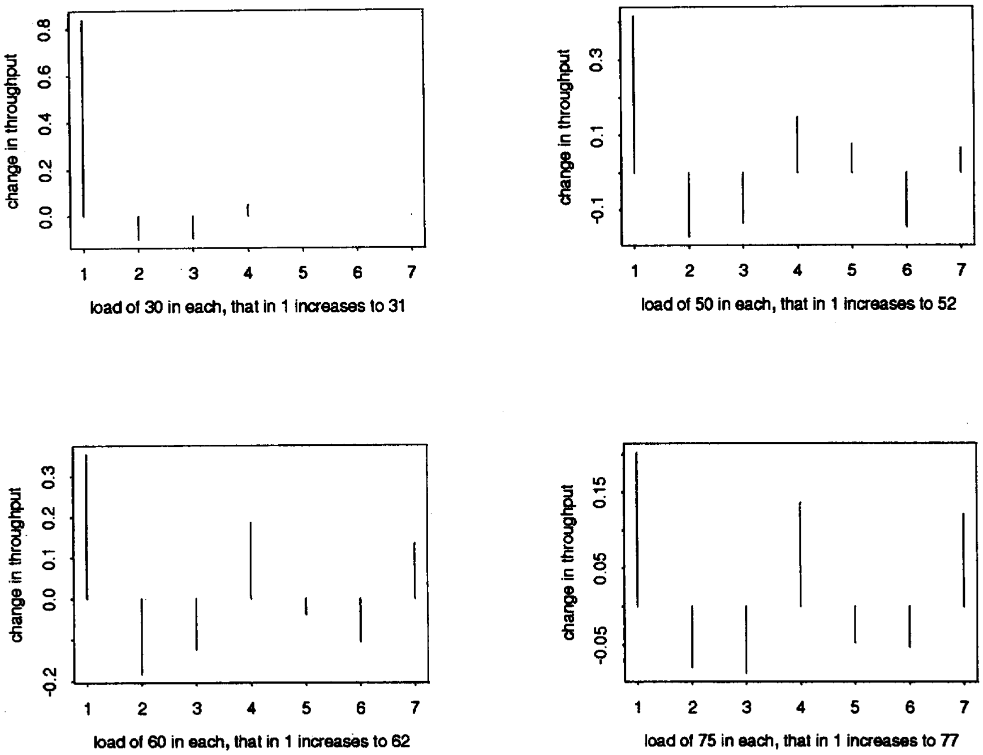

Fig. 4. Sensitivity of throughput in neighboring cells.

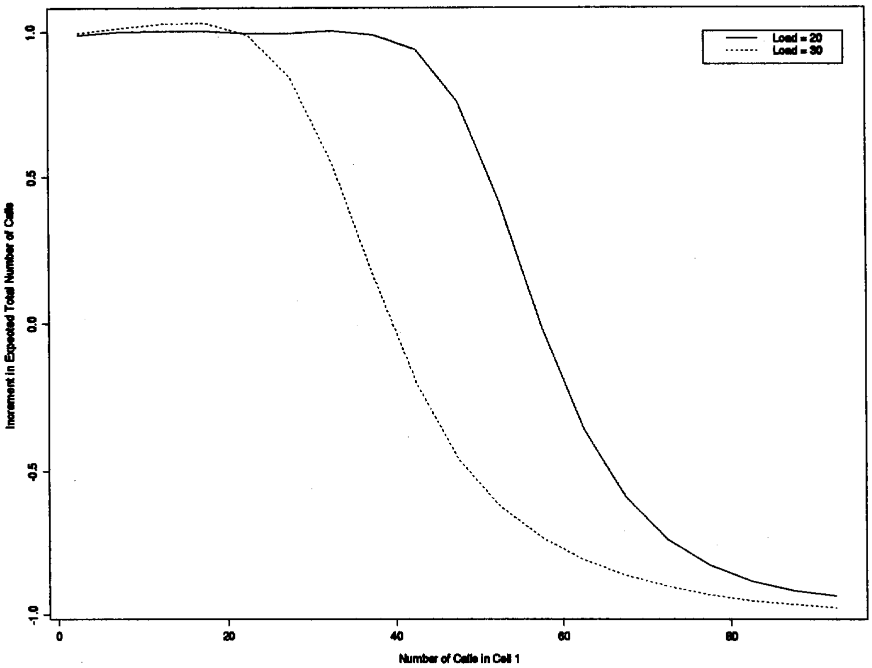

Fig. 5. Value of an additional call in cell 1.

First, note that throughput in cells 2 and 3 decrease for all loads. Cells in the same cluster are thus substitutes. This is caused by the strong competition for channels within a cluster. Any increase in effective throughput in a cell, resulting from an increase in request rate for that cell, is at the expense of throughputs of other cells in the same cluster.

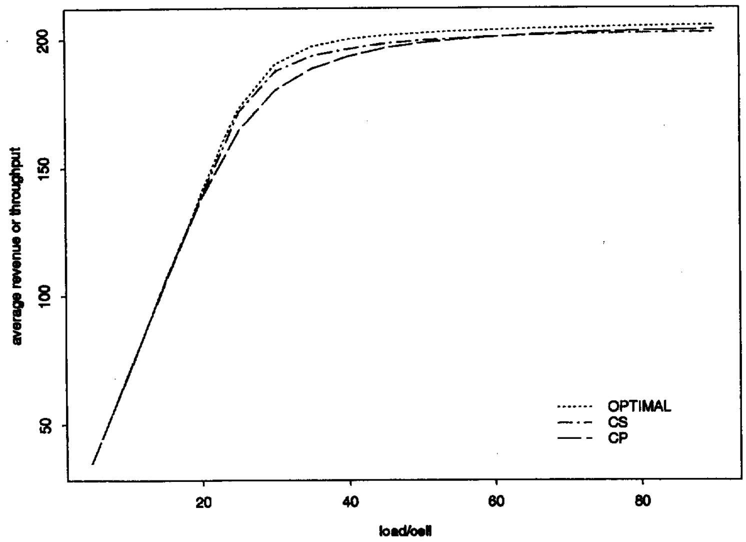

Fig. 6. Total throughput for competing channel allocation schemes.

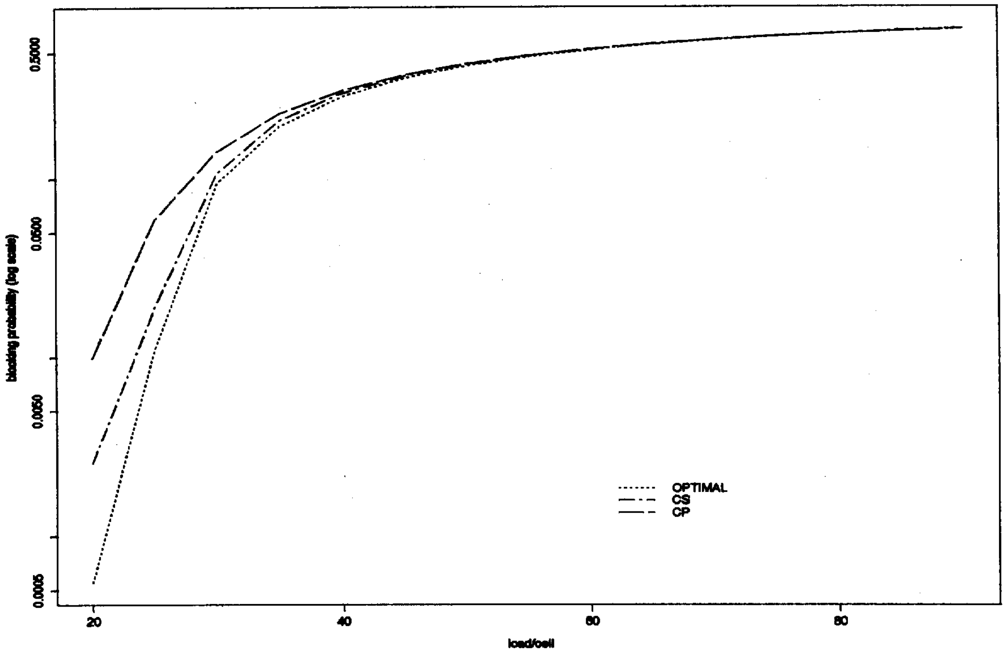

Fig. 7. Blocking probability for competing channel allocation schemes.

Second, note that throughput increases in cell 4 in all cases. Cells 2, 3, and 4 form a reuse cluster. As throughput of cells 2 and 3 decrease due to an increase in throughput of cell 1, the result is an effective increase in throughput of cell 4, as loads remain the same in all 3 cells of the cluster. Thus an increase in throughput of cell 1 *indirectly* effects an increase in throughput of cell 4. Services separated by 1 reuse distance (in this case 3 cells) are thus complements.

The effect of an increase in traffic in cell 1 upon cells more than 3 cells away are dependent upon load, both in sign and magnitude. For low loads, when there are enough idle channels, any incremental change in the load of cell 1 has a *large* effect upon the throughput of cell 1, but this sensitivity damps *quickly* with distance from that cell. At high loads, an incremental change in traffic in cell 1 produces a *small* change in throughput there, but this change damps *slowly* with distance. The signs and magnitudes of sensitivities of cells greater than 3 diameters apart thus depend on the particular load pattern.

Consider the implications of these sensitivities upon a channel allocation policy by focusing on the decision to accept or deny a call in cell 1. If the call is accepted, fewer calls on

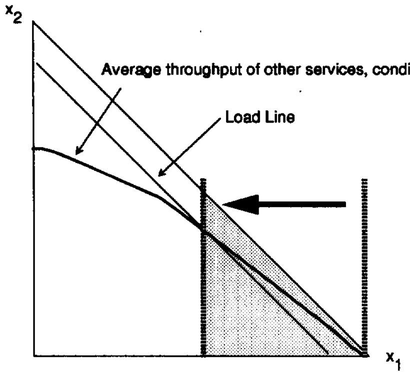
x_2
Average throughput of other services, conditioned on $X_1=x_1$
Load Line
x_1

Fig. 8. Placing and adjusting a new constraint.

average will be accepted in all clusters containing cell 1, but more calls on average will be accepted in cells 1 reuse distance away. The call in cell 1 should be accepted if it results in an average *increase* in the number of calls accepted in the whole cellular system.

From our discussion above, we expect that the effect of an acceptance of a call in cell 1 upon neighboring cells depends on the current state of the system and on the load. We therefore consider the *sum effect* of accepting a call in cell 1 upon the *total expected number* of accepted calls, conditioned on the number of existing calls in cell 1 and on the load. Results are displayed in Fig. 5 for a CS policy under a low load and under a high load[3].

We find that when cell 1 is currently carrying a *small* number of calls, accepting 1 more call in that cell results in an *increase* of almost 1 in the expected number of calls admitted in the cellular system. The effect of the additional call in cell 1 upon neighboring cells is small, since there are many idle channels. When cell 1 is currently carrying a *large* number of calls, however, accepting 1 more call in that cell results in a *decrease* of almost 1 in the expected number of calls admitted in the cellular system. Now, the effect of the additional call in cell 1 upon neighboring cells is large, and 1 additional call in cell 1 results in a drop of almost 1 call each, on average, in the two clusters on either side of cell 1.

This implies that the optimal policy accepts all calls when there are many idle channels, and blocks calls *when one cell is occupying too large a percentage of the cluster's capacity.* The form of the optimal policy for the considered class, therefore, will be to restrict the state space to a subset that does not include states near extreme points corresponding to domination of a cluster's capacity by a single cell. One such policy would be a mixed policy of the form

$$\sum_{j\in C_i} x_j \le M \qquad \forall C_i$$

$$x_i \le K \qquad \forall i \qquad \frac{M}{C} \le K \le M.$$

We also find that the transition in net gain from +1 to −1 occurs at a level of calls in cell 1 dependent upon the load in other cells. At higher loads, the transition from a gain to a shortfall occurs at a lower occupancy in cell 1. This is expected from Fig. 4, since at higher loads cell 1 becomes a stronger substitute with cells 2 and 3. This implies that the optimal policy is dependent on load. At lower loads, it restricts access to fewer states; at higher loads, it restricts access to more states. For instance, if the optimal policy were a mixed policy of the form above, K would decrease from M to M/C as the load increased from 0 to infinity.

The optimal policy in the considered class, therefore, resembles CS at low loads and CP at high loads.

VI. Optimal Control under Equal Loads

We consider a 7 cell linear network, with a 3 cell reuse cluster, as before (Fig. 3), but to counter edge effects we add clusters containing cells (6, 7, 1) and (7, 1, 2). This translates into a seven-dimensional MSMR model where the state equations are of the form $Ax \le b$, where the matrix A and vector b depend upon the dynamic channel allocation strategy.[4]

We initially compare complete sharing and complete partitioning channel allocation policies, under equal loads for all cells. Plots for both the blocking probability and average total throughput are shown (Figs. 6 and 7). Note that all throughput curves are increasing and concave, as expected. Under low to medium load conditions CS performs better than CP. However, at high loads CP approaches CS, and at very high loads does a little better than CS.

This behavior is as suggested by Theorem 4.5. At low and moderate loads the set of states Z_{CS-CP} has expected throughput above average. At high loads the same set of states Z_{CS-CP} has an expected value less than the expected value for Z_{CP}, and hence CP outperforms CS.

We now turn to the optimal policy for the model introduced in Section II.

Consider the two-dimensional state space Markov chain for a CS policy (Fig. 8). The load line passes through all states that generate an instantaneous throughput equal to the average throughput. States below the load line generate less than average total throughput, and hence we can do better by eliminating these states as long as the new state space is coordinate convex, preserving the product form distribution of the Markov chain.

The control algorithm adaptively adds constraints to a CS policy. Average total throughput on the whole state space, and on all boundaries and extreme points, are calculated. The routine places new constraints at extreme points with below average total throughput, and "pushes in" these constraints until the average total throughput on the new boundary equals that on the new state space.

The process therefore produces a new convex and coordinate convex state space that is near optimal, pursuant to Theorem 3.2.

[3] The results are for the same 7 cell linear network, but with additional clusters containing cells (6, 7, 1) and (7, 1, 2).

[4] The length of time required to find the optimal policy is substantial. Searching among convex policies and approximating the discrete distribution by a continuous one reduces the complexity, but the current software implementation of these algorithms still limits us to a small space dimension.

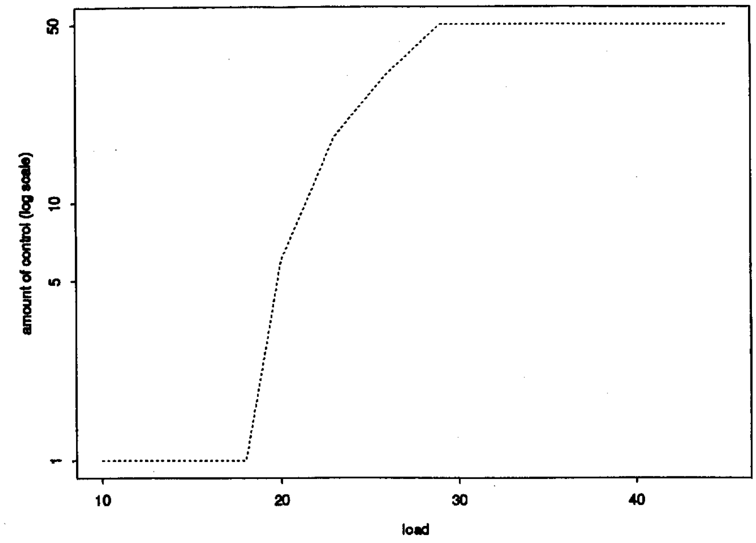

Fig. 9. Amount of control applied.

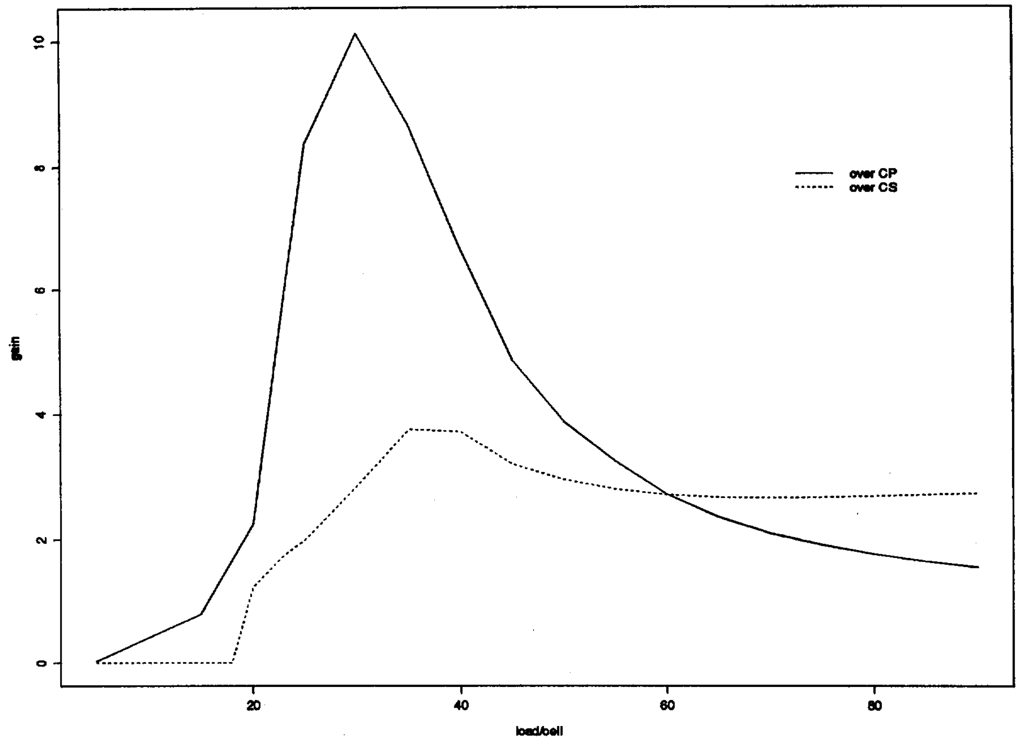

Fig. 10. Gain of optimal policy over CS and CP policies.

The accuracy of the control algorithm depends on step sizes. Smaller step sizes result in more accurate results at the cost of greater execution time. Hence the results that we obtained are near optimal rather than optimal.

As seen from the plots for blocking probability and total throughput (Figs. 6 and 7), optimal control results in a gain in average total throughput and a decrease in blocking probability over both CS and CP, of the form expected from Theorem 4.7. At low loads CS is optimal, since all extreme points of the state space generate above average throughputs. At medium to high loads we get an improvement over CS as the algorithm restricts the maximum number of channels used by each cell through the addition of new constraints. As we have shown in Fig. 5, the throughput carried by the network decreases if a single cell carries too much traffic, and the control algorithm exactly avoids this scenario. These restrictions increase as the load in the network increases, with the state space finally approaching that for CP at high load.

We show the amount of control applied at various loads in Fig. 9. The abscissa is the difference between the maximum number of calls allowed in any cell under the optimal policy and the maximum number allowed under CS. The amount of control is the same for all cells as we have a symmetric network under equal load conditions.

No control is required at low loads. As the load increases, the amount of control applied increases, as discussed above. Fig. 10 shows the gain of the optimal policies over CS and CP policies, in terms of average total throughput. At low loads CS is the optimal policy, and hence there is no gain. As we increase the load, the control algorithm adds new constraints, cutting off sets of states generating less than average total throughput. As we get rid of these states, the average total throughput increases over that for CS.

For high loads the control algorithm restricts the state space to an extent that it almost looks like that for CP, resulting in decreasing differences between throughput from optimal and CP policies. Maximum gian is achieved at moderate load, as predicted by Theorem 4.7.

VII. Parting Thoughts

This study has been restricted to the homogeneous traffic case. Since dynamic channel allocation elicits its gains from instantaneous differences in perceived load among nearby cells, we expect these gains to increase in heterogeneous traffic situations. We will explore optimal channel allocation under unequal loads in future studies.

References

[1] D. C. Cox and D. O. Reudnik, "Increasing channel occupancy in large-scale mobile radio systems: dynamic channel assignments," *IEEE Trans. Veh. Technol.*, vol. 22, Nov. 1973.

[2] S. M. Elnoubi, R. Singh, and S. C. Gupta, "A new frequency channel assignment algorithm in high capacity mobile communication systems," *IEEE Trans. Veh. Technol.*, vol. 31, Aug. 1982.

[3] D. E. Everitt and D. Mansfield, "Performance analysis of cellular mobile communication systems with dynamic channel assignment," *IEEE J. Select Areas Commun.*, vol. 7, Oct. 1989.

[4] B. Hajek and A. Krishna, "Bounds on the accuracy of the reduced-load blocking formula in some simple circuit-switched networks," *Commun., Control, and Signal Processing,* 1990.

[5] S. Jordan and P. P. Varaiya, "Throughput in multiple service, multiple resource communications networks," *IEEE Trans. Commun.*, vol. 39, Aug. 1991.

[6] ——, "Control of multiple service, multiple resource communications networks," *InfoCom '91,* Bal Harbour, FL, Apr. 1991.

[7] ——, "A continuous model of mulitple service, multiple resource communications networks," *ICC '91,* Denver, CO, June 1991.

[8] S. Jordan, "Algorithms for resource allocation in multiple service, multiple resource communications networks," *Allerton Conference on Communication, Control and Computing,* Monticello, IL, Oct. 1991.

[9] T. J. Kahwa and N. D. Georganas, "A hybrid channel assignment scheme in large-scale, cellular-structured mobile communication systems," *IEEE Trans. Commun.*, vol. 26, Apr. 1978.

[10] F. P. Kelly, "Blocking probabilities in large circuit-switched networks," *Advances in Applied Probability,* vol. 18, 1986.

[11] ——, "Routing in circuit-switched networks: optimization, shadow prices and decentralization," *Advances in Applied Probability,* vol. 20, 1988.

[12] K. R. Krishnan, "The convexity of loss rate in an Erlang loss system and sojourn in an Erlang delay system with respect to arrival and service rates," *IEEE Trans. Commun.*, vol. 38, Sept. 1990.

[13] S. P. R. Kumar, H. W. Chung, and M. Lakshminarayan, "Dynamic channel allocation in cellular/wireless networks," in *Third Generation Wireless Information Networks,* S. Nanda and D. J. Goodman Eds. Norwood, MA: Kluwer, 1992.

[14] E. J. Messerli, "Proof of a convexity property of the Erlang B formula," *Bell System Tech. J.*, vol. 51, 1972.

[15] S. Nanda and D. J. Goodman Eds., *Third Generation Wireless Information Networks.* Norwood, MA: Kluwer, 1992.

[16] V. Prabhu and S. S. Rappaport, "Approximate analysis for dynamic channel assignment in large systems with cellular structure," *IEEE Trans. Commun.*, vol. 22, Oct. 1974.

[17] P. A. Raymond, "Performance analysis of cellular networks," *IEEE Trans. Commun.*, vol. 39, Dec. 1991.

[18] L. Schiff, "Traffic capacity of three types of common-user mobile radio communication systems," *IEEE Trans. Commun.*, vol. 18, Feb. 1970.

[19] K. N. Sivarajan, R. J. McEliece, and J. W. Ketchum, "Dynamic channel allocation in cellular radio," in *Proc. IEEE Conf. Veh. Technol.*, May 1990.

[20] J. Tajima and K. Imaura, "A strategy for flexible channel assignment in mobile communication systems"' *IEEE Trans. Veh. Technol.*, vol. 37, May 1988.

[21] S. Tekinay and B. Jabbari, "Handover and channel assignment in mobile cellular networks," *IEEE Commun. Mag.*, vol. 29, Nov. 1991.

[22] J. Xu and P. Mirchandani, "Virtual fixed channel allocation in radio-telephone systems," *ICC '92,* Chicago, IL, June 1992.

[23] M. Zhang and T. P. Yum, "Comparison of channel-assignment strategies in cellular mobile telephone systems," *IEEE Trans. Veh. Technol.*, vol. 38, Nov. 1989.

Scott Jordan (S'86–M'91) received the B.S./A.B., M.S., and Ph.D. degrees from the University of California, Berkeley, in 1985, 1987, and 1990, respectively in applied mathematics and electrical engineering and computer science.

He is currently an Assistant Professor of Electrical Engineering and Computer Science at Northwestern University. His teaching and research interests are the modeling and analysis of behavior, control, and pricing in computer/telecommunication networks, production, queueing, and other stochastic systems.

Asad Khan (S'91) received the B.S. degree from the University of Engineering and Technology, Lahore, Pakistan, in 1989, and the M.S. degree from Northwestern University, in 1992, both in electrical engineering.

From June 1989, to December 1990, he worked at Carrier Telephone Industries in Islamabad, Pakistan, where he was involved in software design and development of microcontroller based systems. Currently, he is a Ph.D. student at Northwestern University. His research is in traffic characterization and control of high speed networks.