On Simple Algorithms for Dynamic Load Balancing

Murat Alanyali and Bruce Hajek

Coordinated Science Laboratory and the
Department of Electrical and Computer Engineering
University of Illinois, Urbana, Illinois 61801

Abstract

This paper focuses on the dynamic resource allocation problem. In communication networks such as wireless or circuit switching networks, resources correspond to base stations or links, and the load corresponds to the calls in the network. The principle of load balancing is examined for dynamic resource allocation. The question of interest is the performance of simple allocation strategies which can be implemented online. Either finite capacity constraints on resources or mobility of users can be incorporated into the setup. The load balancing problem is formulated as an optimal control problem, and variants of a simple "Least Loaded Routing" policy are shown to be asymptotically optimal for large call arrival rates.

1 Introduction

Dynamic resource allocation problems have been of considerable interest in communication networks recently, particularly in the context of dynamic routing in switching networks, and transmission scheduling in wireless networks. The generic dynamic resource allocation problem consists of customers which arrive at different times, each requiring a certain amount of service from a pool of resources. The control variable of the problem is the "assignment policy", which dictates by which resource each customer will be served. Oftentimes in applications, resources have finite capacities, hence the main objective of the assignment policy is to guarantee low blocking probability. On the other hand, in some applications such as spread spectrum mobile radio networks, there are no sharp capacity constraints, and the goal becomes to dynamically balance the load. In either case, one wants the assignment algorithm to have low complexity, require little information about the network state, and be robust to changes in the traffic parameters.

An instance of resource allocation arises in a wireless network, as pictured in Figure 1. The range of a base station is a geographical neighborhood of that station, and each user in the system has to be assigned to one of its neighboring stations. For a number of practical reasons (see [AcaNa94]), stations usually have overlapping ranges. Hence a question of interest is with which station to serve the user when there are multiple alternatives.

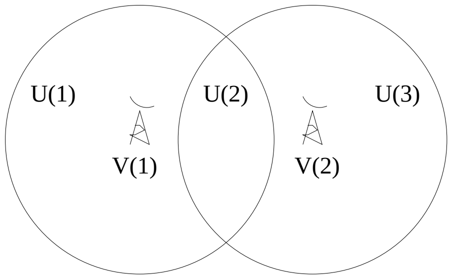

Figure 1: A typical wireless network with overlapping neighborhoods.

Another instance of resource allocation can arise in the "multiparented" circuit switching network structure ([KeGiT93]) illustrated in Figure 1. The network contains a connected core network. Each user is connected to two nodes of the core network, hence for any pair of users, there are multiple paths through the core network. When a call is initiated between a pair of users, a dynamic routing policy is needed to allocate the path on which the call will be routed.

Load balancing is a possible guiding principle for resource allocation, whereby the load is allocated across resources as evenly as possible. In some situations, load balancing may not be desirable. For example, suppose sixteen users are to be assigned to two base stations, such that each base station can handle four users perfectly, but a base station fails completely if five or more users are assigned to it. Then it is better to allocate four users to one station and twelve to the other (so at least four users are satisfied), than to balance the load by allocating eight users to each station (so that no users are satisfied).

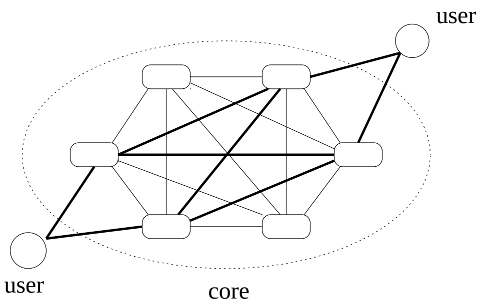

Figure 2: A multiparented circuit switching network.

It is well known that load balancing *can* be an effective allocation strategy, when the cost is convex (or the reward concave) as a function of the allocated loads. For example, $x^2+y^2+z^2$ is minimized over probability vectors (x, y, z) by $x = y = z = 1/3$. This is connected with the convexity of the function $f(x) = x^2$. For the wireless network example above, the total reward at a station with j users assigned to it could be j if $j \leq 4$ and zero if $j \geq 5$. This reward is not concave in j. If the base stations exert some form of rate control, then the reward function for a station might be changed to $f(j) = \min(j, 4)$ for $j \geq 0$. This new reward function is concave, and balancing the load leads to maximizing the total reward.

While the objective of resource allocation depends on the application, often the underlying idea is load balancing ([ChRaN89], [GaWHP93]). There is a rich literature on load balancing, and both static and dynamic versions of the load balancing problem have been studied by numerous authors ([LiuSi88], [WilRe93], and references therein). Most of this work focuses on *algorithms* for load balancing, and studies the performance thru simulations or under simplifying assumptions.

The purpose of this paper is to study certain load balancing strategies for *dynamic* allocation problems in which the load consists of calls that arrive to and depart from the system with certain rates. The main results of the paper are that, in the limit of high arrival rates, (1) a simple myopic strategy is asymptotically optimal in the sense of minimizing a long-term average quadratic cost, (2) a variant of the same strategy achieves the minimum blocking probability in the case of resources with finite capacities, and (3) the system state can be approximated by a system with fluid traffic, satisfying certain *fluid equations*. The fluid equations are essentially linear differential equations with boundary constraints, and they offer an analytic tool for performance analysis of load balancing strategies. Throughout the paper the emphasis is on communication networks, however the methods and results can be applied elsewhere.

The outline of the paper is as follows: Section 2 gives the notation used in the paper, and some preliminary results about *static* load balancing. In Section 3 we define the basic stochastic, dynamic model, where calls remain stationary in the network until departure, and resources have infinite capacities. The dynamic allocation problem is formulated as an optimal control problem with a longterm average quadratic cost, which is to be minimized within a set of practical controls. The results for the static load balancing policy are used to provide a lower bound on the performance of arbitrary controls. Section 4 describes one simple control, namely the *Least Loaded Routing* (LLR) policy. The LLR policy assigns each call to the location with the lightest load possible. Sections 5 and 6 are devoted to the study of LLR for large arrival rates. Towards this end, Section 5 studies the *fluid limit process*. The fluid limit process is a continuous state process satisfying the fluid equations. The fluid limit process converges to an optimal point in equilibrium, and Section 6 uses this fact to prove the asymptotic optimality of LLR. Section 7 considers the case in which resources have finite capacities, and defines the resource allocation problem as minimization of blocking probability. It is shown that the *Least Relatively Loaded Routing* (LRLR) policy asymptotically achieves the best blocking probability for large arrival rates. Section 8 extends the optimality of LLR to the more general model in which calls are allowed to migrate while they are in the system. Convergence of the network process and the asymptotic optimality of LLR are confirmed in Section 9 via simulation results. Finally, Section 10 contains a summary of conclusions.

Throughout the paper, $\|\cdot\|_{sup}$ denotes the supremum norm, unsubscripted variables denote a whole process, and variables with subscripts denote the value of a process at particular times. Proofs of most of the results are not included in the paper, but are included in a full version under preparation.

2 Notation and the Static Problem

In this paper, the abstraction of the load balancing setting is the *consumer demand network*, denoted by (U, V, N). A consumer demand network (see Figure 3 for an example) is essentially a bipartite graph. U denotes a set of *consumer types*, V denotes a set of *locations*, and $(N(u) \subset V : u \in U)$ is a set of *neighborhoods*. A *demand* for this network is a vector $(\lambda(u) : u \in U) \in R_+^U$. Here R_+^U is the set of nonnegative real valued vectors with the index set U. An *assignment*, given by $(a_{u,v} : u \in U, v \in V)$, is *admissible* if $a \geq 0$ and $a_{u,v} = 0$ whenever $v \in N(u)^c$. An admissible assignment a *satisfies* demand λ if $\sum_v a_{u,v} = \lambda(u)$ for all $u \in U$. The *load* at

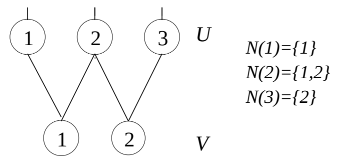

Figure 3: The consumer demand network corresponding to the wireless network example of Figure 1.

location $v \in V$ corresponding to assignment a is given by $q(v) = \sum_u a_{u,v}$, and $q = (q(v) : v \in V)$ is called the *load vector*.

Given a demand λ, let $\mathcal{A}_\lambda$ be the set of admissible assignments satisfying λ. Let $\Phi : R^V \to R$ be a strictly convex, twice differentiable function which is symmetric in its coordinates. Then the *Static Load Balancing Problem* (SLB) is defined as

$$SLB(\lambda, \Phi) \ : \ Minimize(\Phi(q) \ : \ a \in \mathcal{A}_\lambda).$$

The static load balancing problem was studied by Hajek in [Hajek90]. The following characterization of the solutions to the SLB provides considerable insight for the dynamic case, and establishes the intimate connection between load balancing and convexity :

Lemma 1 *There exists a solution to $SLB(\lambda, \Phi)$, and any solution yields the same load vector. Furthermore, an admissible assignment a which satisfies λ is a solution if and only if for all $u \in U$, and all $v \in N(u)$*

$$a_{u,v} = 0 \quad whenever \quad q(v) > m_u(q) \tag{1}$$

where $m_u(q) = \min_{v \in N(u)} q(v)$.

Note that the assignments satisfying (1) are universal, in the sense that they solve SLB for a large set of cost functions. The following lemma states one of the desirable effects of load balancing. (It is stated in [Hajek90] in case Φ is the sum of scalar convex functions.)

Lemma 2 *The solutions to $SLB(\lambda, \Phi)$ minimize the maximum load $q_{max} = \max_{v \in V} q(v)$, over all admissible assignments satisfying the demand λ.*

Let $\Psi(\lambda)$ be the value of $SLB(\lambda, \Phi)$, i.e. $\Psi(\lambda) = \Phi(q)$, where q is the unique load vector corresponding to the solutions of $SLB(\lambda, \Phi)$. The following lemma is used in the next section to establish a lower bound for the performance of dynamic load balancing policies:

Lemma 3 *Ψ is convex.*

3 The Basic Dynamic Model

Given a consumer demand network (U, V, N), and a demand vector λ, consider the following dynamic setting: For every consumer type $u \in U$, there is an independent Poisson arrival stream of rate $\lambda(u)$. Refer to the arrivals of consumer type u as *type u calls*. Upon arrival, each call is to be assigned to a location within the neighborhood associated with its type. We will assume that each location has infinite capacity, therefore the network can accommodate every call immediately. This assumption is relaxed in Section 7. Each call stays in the network for an exponentially distributed length of time with unit mean, independent of the past history. In this basic model it is assumed that calls do not change their locations until they depart from the system. This assumption is relaxed in Section 8, where a call is allowed to migrate in the sense that its type may change.

Let the *load* at location $v \in V$ be the number of calls at that location. Given the arrival and departure times, the load process is defined by the assignment rule of calls to locations. A *call assignment policy* is an algorithm which assigns new arrivals to locations within their respective neighborhoods. An assignment policy is said to be *load conserving* if it does not drop any calls, i.e. assigns every call to one of the locations.

Let $X_t(v)$ be the load at location $v \in V$ at time t under some call assignment policy π. The performance measure for π is the long term average cost J^π, defined by

$$J^\pi = \liminf_{T \to \infty} E\left[\frac{1}{T}\int_0^T \Phi(X_t)dt \,|X_0 = x_0\right],$$

where $\Phi(x) = \sum_v x^2(v)$ for $x \in R^V$. The *dynamic load balancing problem* is to determine the set of load conserving call assignment policies that minimize J^π.

A wide class of convex functions would be appropriate for the instantaneous cost in order to focus on load balancing, in the sense that certain asymptotic optimality properties remain true for any such cost function. In this paper we concentrate on the quadratic instantaneous cost for convenience.

Let $L_t(u)$ be the number of calls of type $u \in U$ in the network at time t. Note that call arrivals and departures, and hence the process L, are not affected by the assignment decisions. Therefore minimizing the instantaneous cost at every t by solving $SLB(L_t, \Phi)$ with integral assignments yields a lower bound to the cost of any load conserving policy. $L_t(u)$ has the same distribution as the number in a $M/M/\infty$ queue with load factor $\lambda(u)$ and with the same initial distribution. Hence in equilibrium, $(L(u) : u \in U)$ is a vector of Poisson random variables, with mean vector λ.

Let the function $\Psi_{IC} : Z_+^U \rightarrow R_+$ be defined by $\Psi_{IC}(m) = \Phi(q_{IC})$, where q_{IC} is the load vector corresponding to the solutions of $SLB(m, \Phi)$ with integral assignments. Note that for any $m \in Z_+^U$, $\Psi_{IC}(m) \geq \Psi(m)$, therefore for any load conserving policy π,

$$\begin{aligned} J^\pi &\geq \liminf_{T\to\infty} E\left[\frac{1}{T}\int_0^T \Psi_{IC}(L_t)dt\,|X_0 = x_0\right] \\ &= E\left[\Psi_{IC}(L)\right] \\ &\geq E\left[\Psi(L)\right] \\ &\geq \Psi(\lambda). \end{aligned} \tag{2}$$

where the last inequality follows by Lemma 3 and Jensen's inequality.

If repacking is allowed, the optimal policy is to rearrange all the calls in the network upon every arrival and departure so as to solve $SLB(L_t, \Phi)$ with integral assignments. This policy achieves $J^\pi = E[\Psi_{IC}(L)]$ since $\Phi(X_t) = \Psi_{IC}(L_t)$ for all t. Repacking can be implemented at a cost of $O(|V||U| + |V|^2)$ computations per call arrival and call departure, which can be too high for applications requiring fast network response such as mobile radio networks. Furthermore frequent repacking of calls may not be feasible or may cause significant performance degradation due to hand-off overhead ([AcaNa94]). Hence the approach taken in this paper is to study the relative performance of more practical policies, which are much simpler in terms of computational complexity and required information about the network state.

4 Least Loaded Routing

The particular call assignment policy considered in the basic model is the "Least Loaded Routing" (LLR) policy, which is defined as follows:

- When a type u call arrives it is assigned to a location $v \in N(u)$ with the minimum load. If multiple locations achieve the minimum in $N(u)$, the assignment is made randomly among such locations, each having equal probability.

LLR mimics a gradient descent method for the static load balancing problem, under the fixed assignments of already existing calls. It is a myopic assignment policy, and is not necessarily optimal for finite arrival rates. LLR has been studied by a number of authors, and has been shown to have a poor worst case performance relative to the optimal repacking policy [AzaBK92].

LLR is a nonrepacking policy, and costs $|N(u)|$ comparisons per call arrival of type u. Another desirable feature of LLR is that it can be implemented in a distributed manner, by using one independent assignment agent per consumer type. Each arrival can be assigned to a location based on partial information about the network state. Furthermore the traffic parameters do not play a role in the assignment algorithm, which endows LLR with robustness.

Under LLR, the load process $X = ((X_t(v) : v \in V) : t \geq 0)$ is Markov on the state space Z_+^V. Therefore for a given consumer demand network, one can, at least in theory, compute the equilibrium distribution of X, and therefore the cost incurred by LLR. However obtaining an expression for the cost of LLR for arbitrary networks through an expression for the equilibrium distribution is not tractable. We therefore resort to fluid model approximations, and evaluate the performance of LLR for any network in heavy traffic. In particular, we study the performance of LLR relative to the performance of the optimal assignment policy when the call arrival rates are very large.

5 The Fluid Limit and its Properties

Taking the fluid limit of a sequence of processes entails use of the law of large numbers for such sequences. In the dynamic load balancing setting, given a fixed load vector $\lambda_0(u)$, we set $\lambda = \gamma\lambda_0$, and investigate the weak limits of processes of interest as $\gamma \rightarrow \infty$. In particular, let X be the load process on (U, V, N) under LLR when the demand is λ, and define the *normalized load process* $X^\gamma = \gamma^{-1}X$. If $X_0^\gamma \Longrightarrow X_0$, then $X^\gamma \Longrightarrow x$, such that (x, A), for some A, satisfies the following equations:

Basic Fluid Equations

$$x_t(v) = x_0(v) + \sum_{u\in N^{-1}(v)} A_{u,v}(t) - \int_0^t x_s(v)ds \tag{3}$$

$$A_{u,v}(0) = 0, \;\; A_{u,v}(t) \text{ nondecreasing,} \tag{4}$$

$$\sum_{v\in N(u)} A_{u,v}(t) = \lambda_0(u)t, \tag{5}$$

$$\int_0^t I\{x_s(v) > m_u(x_s)\}dA_{u,v}(s) = 0 \tag{6}$$

with the random initial condition X_0.

Note that in the fluid analogy, $A_{u,v}(t)$ corresponds to the amount of fluid received by location v from consumer u up until time t. By (3), $x_t(v)$ is the load at location v at time t under assignment regime A and linear discharge rate. Condition (5) is a "conservation of fluid" equation, stating that the total fluid arrived at type u by time t is equal to the total fluid that u has assigned to the locations by t. Conditions (3)-(5) are satisfied under any load conserving policy. Condition (6) is the impact of the LLR policy, which states that

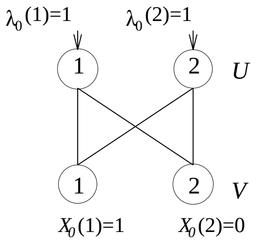

Figure 4: The consumer demand network of Example 1.

location v does not receive any fluid from consumer u unless the load in v achieves the minimum in the neighborhood of u.

Properties of Solutions to the Fluid Equations

Monotonicity : Suppose (x, A) and $(\tilde{x}, \tilde{A})$ are two solutions to the fluid equations with $x_0(v) \geq \tilde{x}_0(v)\ \forall v$. Then $x_t(v) \geq \tilde{x}_t(v)\ \forall v$ and $\forall t \geq 0$.

Uniqueness of Load Trajectory : If (x, A) and $(\tilde{x}, \tilde{A})$ are two solutions to the fluid equations (3)-(6) with $x_0 = \tilde{x}_0$, then $x_t = \tilde{x}_t$ for all $t \geq 0$. Note that uniqueness of the load trajectory follows easily by applying the monotonicity property twice with $\tilde{x}_0 \leq x_0$ and $\tilde{x}_0 \geq x_0$.

Insensitivity to Initial State : Let (x, A) be a solution to the fluid equations with $x_0(v) \geq 0$ for all v. Then,

$$\|x_t - q\|_{sup} \leq e^{-t}\left(\|q\|_{sup} \vee \sum_v x_0(v)\right),$$

where q is the unique load vector corresponding to the solutions of $SLB(\lambda_0, \Phi)$.

Let a be a solution to $SLB(\lambda_0, \Phi)$. It is easy to verify that $\left(\{q(v)(1-e^{-t})\}, \{a_{u,v}t\}\right)$ is a solution to the fluid equations with zero initial state, and that as $t \to \infty$, the load converges exponentially to q. The insensitivity property, which is handy in proving the convergence of equilibrium distributions, implies that starting from *any* initial state, x_t converges to q exponentially fast.

To illustrate the solutions of the fluid equations (3)-(6) for a typical consumer demand network, consider the following example :

Example 1 *Consider the consumer demand network of Figure 4. At time $t = 0$, $x_0(1) = 1 > 0 = x_0(2)$, therefore the fluid equations are satisfied only if both consumers assign all of their fluid to location 2. However in this assignment regime, $x(1)$ will decay to 0, and $x(2)$ will build up to 2 exponentially, equalizing at the "collapse time", $t_c = \ln 3/2$.*

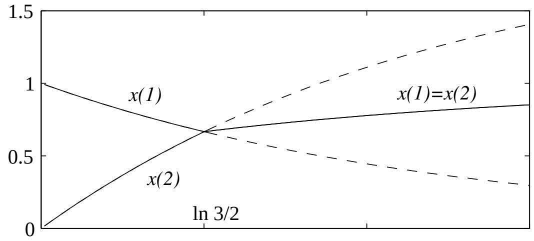

Figure 5: The solution to the fluid equations (3)-(6) for the example of Figure 4.

After the collapse time (3)-(6) are not violated if both consumers divide their incoming fluid equally between the two locations, and maintain $x_t(1) = x_t(2)$ for $t \geq t_c$. Therefore, the load x for a solution to the fluid equations is as given in Figure 5, and by the uniqueness property, it is the only solution.

Note that the fluid equations do not necessarily determine the function A uniquely. For a simple illustration, consider the network of Example 1 where $x_0(1) = x_0(2) = 0$. Let $A_{u,v}(t) = (1/2)t$, and $\tilde{A}_{u,v}(t) = I\{u = v\}t$. Both (x, A) and $(x, \tilde{A})$, where $x_t(v) = 1 - e^{-t}$ for $v = 1, 2$, satisfy the fluid equations with the zero initial state.

6 Asymptotic Optimality of LLR

The main result of this section is that for large arrival rates, LLR performs as well as any load conserving assignment policy. By the process level convergence result of Section 5, X^γ converges to x on *finite* time intervals as $\gamma \to \infty$, and in turn by the insensitivity property of solutions, x_t converges to q as time t tends to *infinity*. Therefore for any $\epsilon > 0$,

$$\lim_{t\to\infty} \lim_{\gamma\to\infty} P\left(\|X_t^\gamma - q\|_{sup} > \epsilon\right) = 0. \qquad (7)$$

A technical point in the proof of the asymptotic optimality of LLR is to establish the convergence of the equilibrium distributions of the load process. This amounts to showing that (7) holds under the interchange of limits. The result is Lemma 4. The proof relies on convergence of X^γ to x over large, fixed-length time intervals, uniformly over the initial times of such intervals.

Lemma 4 *(Convergence of Equilibrium Distribution) Let q be the unique load vector corresponding to solutions of $SLB(\lambda_0, \Phi)$, and ν be the equilibrium distribution of X^γ. Then $\forall \epsilon > 0$*

$$\lim_{\gamma\to\infty} \nu\left(\|X^\gamma - q\|_{sup} > \epsilon\right) = 0.$$

Note that by ergodicity, the cost of any stationary call assignment policy is determined by the equilibrium distribution of the corresponding load process. Lemma 4 describes the limiting equilibrium distribution of X^γ, in sufficient detail to imply the asymptotic optimality of LLR :

Theorem 1 *Suppose $\lambda_0(u) \neq 0$ for some $u \in U$. Let π be any load conserving policy with the corresponding cost J_γ^π. Let J_γ^{LLR} be the cost of the LLR policy. Then*

$$\liminf_{\gamma\to\infty} \gamma^{-2} J_\gamma^\pi \geq \lim_{\gamma\to\infty} \gamma^{-2} J_\gamma^{LLR} = \Psi(\lambda_0) > 0$$

Proof Being dominated by the normalized total load process, $(X^\gamma(v)^2 : \gamma \geq 0)$ is uniformly integrable. Thus by Lemma 4,

$$\begin{aligned}
\lim_{\gamma\to\infty} \gamma^{-2} J_\gamma^{LLR} &= \lim_{\gamma\to\infty} E\left[\sum_v (X^\gamma(v))^2\right] \\
&= E\left[\sum_v q^2(v)\right] \\
&= \Psi(\lambda_0).
\end{aligned}$$

However by (2), for any load conserving policy π,

$$\begin{aligned}
\gamma^{-2} J_\gamma^\pi &\geq \gamma^{-2}\Psi(\gamma\lambda_0) \\
&= \Psi(\lambda_0)
\end{aligned}$$

for all $\gamma > 0$. This proves the theorem. □

Notice that for any load conserving policy, the total load in the network in equilibrium is Poisson with mean $\gamma\sum_u \lambda_0(u)$. Thus the cost incurred by any such policy will grow unboundedly as $\gamma \to \infty$. In this regime, LLR asymptotically achieves the best *normalized* cost by keeping the ratios of the loads at locations at their respective optimal values.

7 Locations with Finite Capacities

In this section we consider a variation of the basic model in which each location has a finite capacity. Namely, at any time, the number of calls assigned to a location cannot exceed the capacity of the location, and the location is called *full* if the capacity is achieved. A call is *lost* if, upon its arrival, all the locations in its neighborhood are full. Lost calls vanish immediately, hence do not contribute to the system state. The goal of the assignment policy is to minimize the percentage of calls lost in the system.

As in the infinite capacity case, consideration of a static loss minimization problem provides considerable intuition. Towards this end, let a *capacity vector* $\kappa = (\kappa(v) : v \in V)$ be a vector of nonnegative numbers. Define $\mathcal{B}_{\lambda_0,\kappa}$ as the set of admissible assignment vectors a such that $\sum_v a_{u,v} \leq \lambda_0(u)$ for all u, and $q(v) \leq \kappa(v)$, where q denotes the load vector determined by a. The *Static Load Packing Problem* (*SLP*) is the simple assignment problem defined as

$$SLP(\lambda_0, \kappa) \ : \ Maximize(\sum_v q(v) \ : \ a \in \mathcal{B}_{\lambda_0,\kappa}).$$

The dynamic model is described analytically as follows: Given a consumer demand network (U, V, N), a capacity vector κ, and a load vector λ_0, consider the limiting regime of Section 5. Suppose that in each system indexed by γ, each location v has capacity $\gamma\kappa(v)$. A call is lost if it cannot be assigned immediately. Lost calls cannot be assigned later, hence are treated by the system as if they never arrived. The solution of $SLP(\lambda_0, \kappa)$ provides a lower bound for the loss probability of *any* assignment policy:

Lemma 5 *For any assignment policy π, and any $\gamma > 0$,*

$$P_\gamma^\pi(Loss) \geq 1 - \frac{\sum_v q(v)}{\sum_u \lambda_0(u)},$$

where q is a load vector corresponding to a solution of $SLP(\lambda_0, \kappa)$.

Let $l_v(q)$ denote the *relative load* of location v, defined as $l_v(q) = q(v)/\kappa(v)$. This section considers a variation of LLR, namely the "Least *Relatively* Loaded Routing" (LRLR) policy, which is defined as follows

- Upon arrival, a call of type u is assigned to a location $v \in N(u)$ with the minimum relative load, provided that the minimum relative load is less than one. Otherwise, all the locations in $N(u)$ are full, and the call is lost.

To facilitate the analysis, we introduce an imaginary *loss location*, v_L, of infinite capacity, to which calls are assigned only if they would be lost otherwise. Note that the resulting system is lossless, and the system process restricted to V is the same as the original system. Hence the tools of Section 5 can be applied. To formalize this argument, define a new consumer demand network $(\hat{U}, \hat{V}, \hat{N})$ by $\hat{U} = U$, $\hat{V} = V \cup \{v_L\}$, $\hat{N}(u) = N(u) \cup \{v_L\}$. Let $\kappa(v_L) = \infty$, and $l_{v_L}(q) = 1$, and let $\hat{X}$ be the load process on $(\hat{U}, \hat{V}, \hat{N})$ under LRLR. Note that $(\hat{X}(v) : v \in V) = (X(v) : v \in V)$, where X is the load process on (U, V, N) under LRLR. Let $\hat{X}^\gamma = \gamma^{-1}\hat{X}$. If $\hat{X}_0^\gamma \Longrightarrow \hat{X}_0$, then $\hat{X}^\gamma \Longrightarrow \hat{x}$, such that $(\hat{x}, \hat{A})$, for some $\hat{A}$ satisfies the following equations:

Fluid Equations with Capacity Constraints

$$\hat{x}_t(v) = \hat{x}_0(v) + \sum_{u \in \hat{N}^{-1}(v)} \hat{A}_{u,v}(t) - \int_0^t \hat{x}_s(v) ds \qquad (8)$$

$$\hat{A}_{u,v}(0) = 0, \quad \hat{A}_{u,v}(t) \text{ nondecreasing}, \qquad (9)$$

$$0 \leq \hat{x}_t(v) \leq \kappa(v), \quad \sum_{v \in \hat{N}(u)} \hat{A}_{u,v}(t) = \lambda_0(u)t, \qquad (10)$$

$$\int_0^t I\{l_v(\hat{x}_s) > m_u(l(\hat{x}_s))\} d\hat{A}_{u,v}(s) = 0, \qquad (11)$$

$$\int_0^t I\{m_u(l(\hat{x}_s)) < 1\} d\hat{A}_{u,v_L}(s) = 0, \qquad (12)$$

with the initial condition $\hat{X}_0$, and $m_u(l) = \min_{v \in N(u)} l_v$.

The solutions of Equations (8)-(12) have the monotonicity property, and hence there is a unique $\hat{x}$ that solves (8)-(12). While the solution of $SLP(\lambda_0, \kappa)$ is not necessarily unique, there is a unique solution with a certain balance property, and $\hat{x}$ restricted to V converges to that solution for any initial condition:

Lemma 6 *There exists an assignment $a \in \mathcal{B}_{\lambda_0,\kappa}$ such that*

$$\begin{aligned} a_{u,v} = 0 \qquad & \textit{whenever} \qquad l_v(q) > m_u(l(q)), \\ \sum_v a_{u,v} < \lambda_0(u) \qquad & \textit{only if} \qquad q(v) = \kappa(v) \;\; \forall v \in N(u). \end{aligned}$$

Such assignment is a solution to $SLP(\lambda_0, \kappa)$, and its load vector q is uniquely determined. Furthermore if $(\hat{x}, \hat{A})$ is a solution to (8)-(12), then $\lim_{t\to\infty} \hat{x}_t(v) = q(v)$ for all $v \in V$.

By Lemma 4 of Section 6, the equilibrium distribution of $\hat{X}^\gamma$ also converges. Furthermore, the uniform integrability of $\hat{X}^\gamma$ implies the convergence of the moments, establishing the asymptotic optimality of LRLR:

Theorem 2 *Let q be a load vector corresponding to a solution of $SLP(\lambda_0, \kappa)$. If π is an arbitrary assignment policy,*

$$\liminf_{\gamma\to\infty} P_\gamma^\pi(Loss) \geq \lim_{\gamma\to\infty} P_\gamma^{LRLR}(Loss) = 1 - \frac{\sum_v q(v)}{\sum_u \lambda_0(u)}.$$

8 The Migration Model

In this section we generalize the basic model, by allowing calls to change their types while they are in the system, and by allowing type dependent departure rates.

Consider the wireless network of Section 1, where the type of a call is determined by its geographical location. If the users of the network are mobile, it is conceivable that a user can visit different neighborhoods while in the system. This corresponds to calls migrating among different *types* in the consumer demand network abstraction, which is not covered by the basic dynamic model. In this section we generalize the basic model so as to allow calls to change their types during their lifetimes, and include type dependent departure rates from the system. It is assumed that every time a call changes its type, it is treated like a new arrival, and reassigned, possibly to a new location, using LLR policy. Therefore in the migration model, calls are allowed to change their *locations* as well as their types during their lifetimes. Although this happens in a much less controlled manner than repacking, migration of calls gives the assignment algorithm more chances to balance the load, and the asymptotic optimality of LLR persists in the migration model.

To describe the migration model analytically, let $R = [r_{u,u'}]_{U\times U}$ be called a *routing matrix* if $r_{u,u'} \geq 0$ for $u \neq u'$, and $\sum_{u' \in U} r_{u,u'} \leq 0$ for all $u \in U$. Given a consumer demand network (U, V, N), an arrival rate vector λ_0, and a routing matrix R, consider the load balancing problem of Section 3. Suppose for all $u, u' \in U$ such that $u' \neq u$, each type u call transforms into a type u' call with rate $r_{u,u'}$, or departs from the system with rate $-\sum_{u'\in U} r_{u,u'}$. Each arrival is assigned to a location via LLR policy. In addition, when a call changes its type, it is reassigned using LLR. Its location may or may not change. We assume that R is nonsingular, so that every call eventually departs from the system. It can be verified by direct substitution that in equilibrium $(L(u) : u \in U)$ is a vector of independent Poisson random variables, with mean vector $\gamma\rho_0$, where $\rho_0 = -\lambda_0 R^{-1}$, so that the normalized cost of any load conserving policy is lower bounded by $\Psi(\rho_0)$.

Notice that in the migration model, the evolution of the load process X is determined by the composition of the load according to types. Hence, unlike the basic model of Section 3, the load process alone is not Markov. Therefore we introduce some additional notation, and look at the system process at a finer level :

Let the *contribution* of type u to location v at time t, $C_t(u, v)$, be the number of type u calls at location v at time t. Define $C_t^\gamma(u, v) = \gamma^{-1} C_t(u, v)$, and note that $X_t^\gamma(v) = \sum_u C_t^\gamma(u, v)$. Under LLR, C is Markov on the state space $Z_+^{U\times V}$. Methods of [EthKu86] can be used to show that if C_0^γ is tight, then C^γ is tight. Therefore any subsequence $\gamma_n \to \infty$ has a further subsequence γ_{n_k} such that the distribution of $C^{\gamma_{n_k}}$ converges in Prohorov metric. This is equivalent to weak convergence of $C^{\gamma_{n_k}}$, and the weak limit c, together with some A, satisfy the following equations:

Fluid Equations with Migration

$$c_t(u,v) = c_0(u,v) + A_{u,v}(t) + \int_0^t r_{u,u} c_s(u,v) ds \qquad (13)$$

$$A_{u,v}(0) = 0, \ \ A_{u,v}(t) \text{ nondecreasing}, \qquad (14)$$

$$\sum_{v \in N(u)} A_{u,v}(t) = \lambda_0(u)t + \sum_{u' \neq u} \int_0^t l_s(u') r_{u',u} ds, \qquad (15)$$

$$\int_0^t I\{x_s(v) > m_u(x_s)\} dA_{u,v}(s) = 0, \qquad (16)$$

where $x_t(v) = \sum_{u \in N(v)^{-1}} c_t(u,v)$, $l_t(u) = \sum_{v \in N(u)} c_t(u,v)$.

Since calls of different types may have different departure rates under the migration model, the departures do not necessarily help balancing the load. Furthermore the diversity of departure rates deprives us from the monotonicity property of the basic model, and can cause nonuniqueness of the load process corresponding to the solutions of (13)-(16). Since the initial state c_0 and the fluid equations (13)-(16) do not necessarily determine the load trajectory x uniquely, the load process X^γ does not necessarily have a weak limit. However all the load trajectories determined by (13)-(16) have the same limit point as $t \to \infty$:

Lemma 7 *(Global Convergence) Let x be the load trajectory corresponding to an arbitrary solution of the fluid equations (13)-(16). Then $\lim_{t\to\infty} x_t = q$, where q is the unique load vector corresponding to the solutions of $SLB(\rho_0, \Phi)$.*

Although the nonuniqueness of x introduces some extra complication, Lemma 7 is sufficient to prove the convergence of equilibrium distribution of X^γ to the optimal point q, which in turn establishes the asymptotic optimality of LLR in the migration model :

Theorem 3 *Suppose $\rho_0(u) \neq 0$ for some $u \in U$. Then, in the migration model, the equilibrium distribution of X^γ converges weakly to q, and for any load conserving policy π,*

$$\liminf_{\gamma\to\infty} \gamma^{-2} J^\pi_\gamma \geq \lim_{\gamma\to\infty} \gamma^{-2} J^{LLR}_\gamma = \Psi(\rho_0) > 0.$$

9 Simulations

This Section presents some simulation results to confirm the convergence of the load process, and to compare the performances of LLR and the optimal repacking policies. Towards this goal, we consider the consumer demand network of Figure 6, and study the basic model and the migration model separately.

The first example that we consider is the network of Figure 6 under the basic model. The demand vector is given by

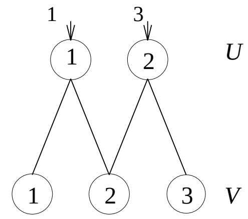

Figure 6: The simulated consumer demand network.

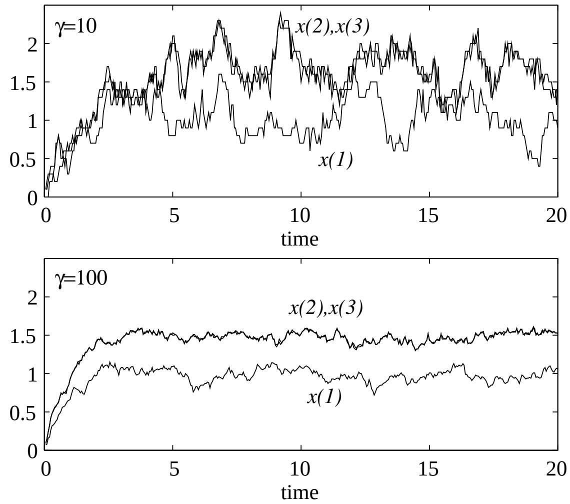

Figure 7: Sample paths of the normalized load process for $\gamma = 10$ and $\gamma = 100$.

	LLR	Optimal Repacking
$\gamma = 1$	8.3550	7.9099
$\gamma = 10$	5.7530	5.7219
$\gamma = 100$	5.5160	5.5159

Table 1: Comparison of the performances of LLR and the optimal repacking policy under the basic model. Lower bound : $\Psi(\lambda_0) = 5.5$.

	LLR	Optimal Repacking
$\gamma = 1$	4.5033	4.1763
$\gamma = 10$	2.8052	2.7656
$\gamma = 100$	2.6815	2.6798

Table 2: Comparison of the performances of LLR and the optimal repacking policy under the migration model. Lower bound : $\Psi(\rho_0) = 2.6729$.

$\lambda_0 = (1,3)$. Calls do not migrate, and each call departs from the system with rate 1, regardless of its type. The solution of the static load balancing problem SLB yields the load vector $q = (1, 3/2, 3/2)$, hence we do not expect to balance the load perfectly at all times in the dynamic problem under any load conserving policy. As a matter of fact, under LLR, at almost all times, consumer 1 assigns all its arrivals to location 1, and consumer 2 concentrates only on locations 2 and 3, balancing its load almost perfectly between these two locations. Typical sample paths of the normalized load process for $\gamma = 10$ and $\gamma = 100$ are given in Figure 7. The upper curve in each case depicts the nearly indistinguishable loads at locations 2 and 3. It can be observed easily that as γ increases the discontinuities of the sample paths disappear, and the normalized load process approaches the deterministic, continuous vector function $((1-e^{-t}), (3/2)(1-e^{-t}), (3/2)(1-e^{-t}))$ as determined by the fluid equations (3)-(6). The normalized costs of LLR and the optimal repacking policy for different values of γ are given in Table 1.

As the second example, consider the network of Figure 6 under the migration model, with the same demand vector λ_0. Let the routing matrix R be

$$\begin{bmatrix} -2 & 1 \\ 2 & -4 \end{bmatrix}$$

so that the migration and the departure rates are both 1 for type 1 calls, and both 2 for type 2 calls. In this case the mean equilibrium load vector is $\rho_0 = (10/6, 7/6)$, hence the load can be perfectly balanced in the network. Under LLR, consumers 1 and 2 split their arrivals so as to balance the load, and the normalized load process approaches the optimal load $q' = (17/18, 17/18, 17/18)$ as time goes to infinity. The normalized costs of LLR and the optimal repacking policy for different values of γ are given in Table 2.

10 Conclusion

This paper concentrates on the dynamic resource allocation problem and studies the performance of two practical call assignment policies, namely LLR and LRLR. When there are no capacity constraints on the resources, LLR is shown to asymptotically achieve the most balanced load in the sense of minimizing a certain cost. LLR is also robust to migration, provided that calls are reassigned according to LLR whenever their type changes. On the other hand, when the resources have finite capacities, LRLR asymptotically achieves the minimum possible blocking probability. The desirable aspects of the considered policies are low computational complexity, decentralized implementation, and robustness to call arrival and departure rates.

Acknowledgement : This work was supported in part by the National Science Foundation under contract NSF NCR 93-14253, and by a TUBITAK NATO Fellowship.

References

[AcaNa94] A. Acampora, and M. Naghshineh, *Control and Quality of Service Provisioning in High-Speed Microcellular Networks,* IEEE Personal Communications Magazine, Second Quarter 1994.

[AzaBK92] Y. Azar, A. Broder, and A. Karlin, *On-line Load Balancing,* Proceedings of 33^{rd} Annual Symposium on FOCS, pp 218-225, 1992.

[ChRaN89] G. M. Chiu, C. S. Raghavendra, and S. M. Ng, *Resource Allocation with Load Balancing Consideration in Distributed Systems,* Proceedings of IEEE Infocom 89, Vol. 2, pp. 758-65, 1989.

[EthKu86] S. Ethier, and T. Kurtz, *Markov Processes : Characterization and Convergence,* Wiley, 1986.

[GaWHP93] G. R. Ganger, B. L. Worthington, R. Y. Hou, and Y. N. Patt, *Disk Subsystem Load Balancing: Disk Striping vs. Conventional Data Placement,* Proc. of 26^{th} Hawaii Int. Conf. on System Sciences, Vol. 1, pp. 40-49, 1993.

[Hajek90] B. Hajek, *Performance of Global Load Balancing by Local Adjustment,* IEEE Trans. on Information Theory, Vol. 36, No. 6, pp. 1398-1414, November 1990.

[KeGiT93] F. P. Kelly, R. Gibbens, and S. Turner, *Dynamic Routing in Multiparented Networks,* IEEE/ACM Trans. Networking, Vol. 1, No. 2, April 1993.

[LiuSi88] H. T. Liu, and J. Silvester, *An Approximate Performance Model for Load-dependent Interactive Queues with Application to Load Balancing in Distributed Systems,* Proceedings of IEEE Infocom 88, pp. 956-65, 1988.

[WilRe93] M. H. Willebeck-LeMair, and A. P. Reeves, *Strategies for Dynamic Load Balancing on Highly Parallel Computers,* IEEE Trans. Parallel and Distributed Systems, Vol. 4, No. 9, pp. 979-93, September 1993.