Analysis of Simple Algorithms for Dynamic Load Balancing

Murat Alanyali and Bruce Hajek

University of Illinois at Urbana Champaign

The principle of load balancing is examined for dynamic resource allocation subject to certain constraints. The emphasis is on the performance of simple allocation strategies which can be implemented on-line. Either finite capacity constraints on resources or migration of load can be incorporated into the setup. The load balancing problem is formulated as a stochastic optimal control problem. Variants of a "Least Load Routing" policy are shown to lead to a fluid type limit and to be asymptotically optimal.

Key words: Dynamic resource allocation, load balancing, fluid equations, loss networks, least load routing.

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

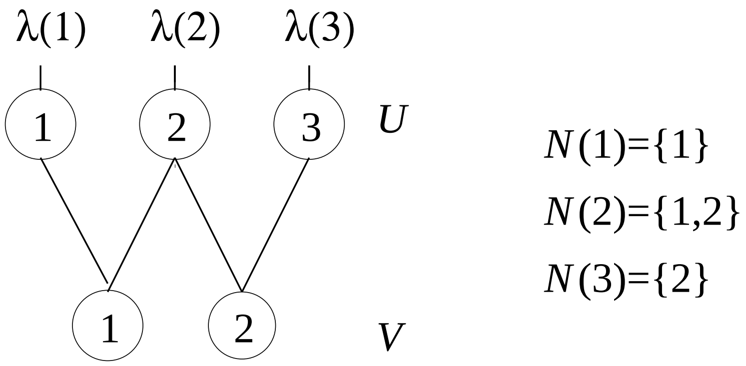

Figure 1: A typical load sharing network (U, V, N).

1 Introduction

Economic pressures and reliability considerations generally lead to communication networks with load sharing capabilities and highlight resource allocation as a fundamental issue in network design. The objective of resource allocation in such systems is oftentimes consumer satisfaction, which may translate to minimizing consumer blocking or achieving fairness by load balancing. Regardless of its objective, an essential aspect of a resource allocation policy is implementability: Practical considerations require allocation policies to have low complexity, require little information about the network state, and be robust to changes in the traffic parameters. This paper concerns trade-offs implied by these requirements.

In this paper the mathematical abstraction of a load sharing network is a triple (U, V, N). Here U is a finite set of *consumer types*, V is a finite set of *locations*, and $(N(u) \subset V : u \in U)$ is a set of *neighborhoods* (see Figure 1 for an example). A *demand* for this network is a vector $(\lambda(u) : u \in U)$ of positive numbers. In a dynamic setting $\lambda(u)$ denotes the arrival rate of *type u consumers*. Each consumer is served, starting immediately upon its arrival, for the duration of its *holding time*. The neighborhood $N(u)$ denotes the locations that are available to type u consumers, in the sense that each such consumer can be served only at a location within $N(u)$. An *allocation policy* is an algorithm that assigns consumers to locations within their respective neighborhoods. The *load* at a location is the number of consumers at the location.

Load balancing is a possible guiding principle for resource allocation, whereby the load is allo-

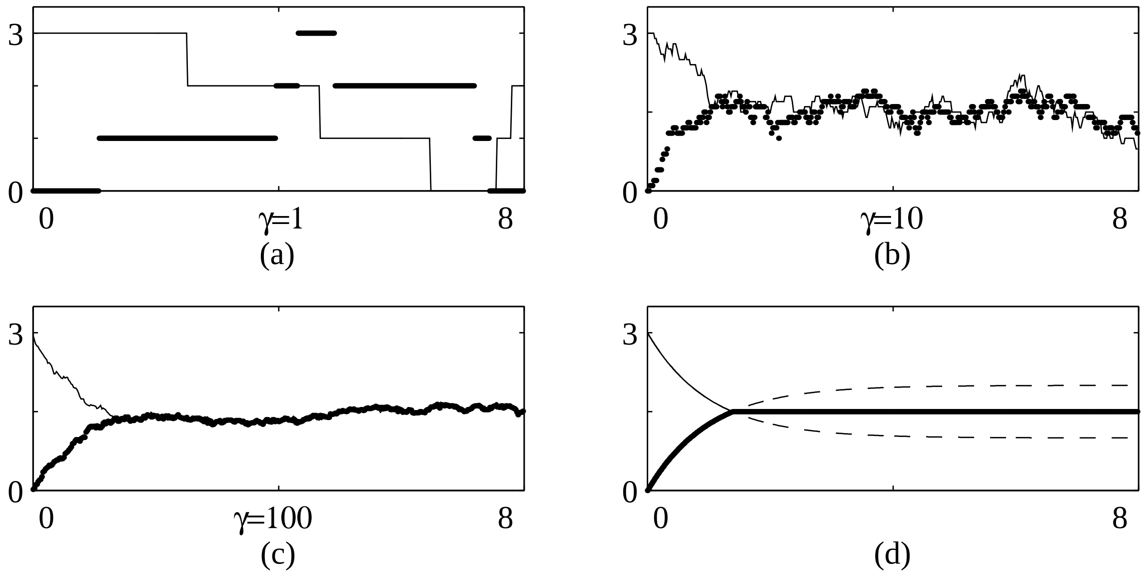

Figure 2: Load under the Least Load Routing policy.

cated across locations as evenly as possible (see for example, Chiu et al. (1989), and Ganger et al. (1993)). There is a rich literature on load balancing, and both static and dynamic versions of the problem have been studied by numerous authors (e.g. Liu and Silvester (1988), Willebeck-LeMair and Reeves (1993), and references therein). Most of this work focuses on *algorithms* for load balancing, and studies the performance through simulations or under simplifying assumptions.

A reasonable allocation policy for dynamic load balancing is the "Least Load Routing"(LLR) policy, which assigns each consumer to a location with the least load in the associated neighborhood. To observe a typical behavior of the network load under the LLR policy, consider the load sharing network of Figure 1. Suppose that the network demand is $\lambda = (\gamma, \gamma, \gamma)$ so that the arrivals of consumers of each type form a Poisson process of rate γ. Each consumer remains in the network for an exponentially distributed amount of time, with unit mean. Finally, suppose that initially location 1 has zero load, whereas location 2 has load 3γ. Figures 2(a)-2(c) depict typical sample paths of the normalized load, defined as the load divided by γ, at the two locations for $\gamma = 1, 10, 100$, for the time interval $[0, 8]$. In the limit as γ goes to infinity, the normalized load converges to the deterministic trajectory depicted in Figure 2(d). Deterministic descriptions of this sort are commonly referred to as *fluid limit approximations.*

This work focuses primarily on the optimality properties of the LLR allocation policy (and variants) implied by the corresponding fluid limits. The main results of the paper are that in the heavy traffic regime, (1) the LLR policy is asymptotically optimal in the sense of minimizing a long term average cost (Theorem 7.1, and similarly for a model including migration, Theorem 9.1), (2) variants of the same policy achieve the minimum blocking probability in the case of locations with finite capacities (Theorem 8.1). As discussed in Section 3, the Bernoulli Splitting policy, in which consumers are assigned to locations at random upon arrival, also shares these optimality properties, if the splitting probabilities are suitably chosen.

A fluid limit analysis is used in the paper, along the lines of Hunt and Kurtz (1994). That paper focuses on trunk reservation policies for networks with finite capacities, and considerable analytical difficulties arise. Asymptotic optimality of trunk reservation strategies for a single resource location is established in Hunt and Laws (1995). Fluid methods are used in Hunt and Laws (1993) to examine trunk reservation in a different regime, namely large numbers of locations with fixed capacities. A related paper on dynamic load balancing is that of Winston (1977), which shows that for routing to two queues, the send-to-the-shorter queue policy is optimal in a strong sense. Previous work dealing with load balancing in networks is that of Section 7.4 of Bertsekas and Tsitsiklis (1989), which considers an algorithm in which load is shifted around a network without constraints, in an asynchronous fashion, based on possibly delayed reports of load at the neighboring nodes. It is shown that the load at all locations converges to the average load.

The outline of the rest of the paper is as follows: Section 2 gives some preliminary results regarding *static* load balancing. Section 3 defines the basic dynamic model, in which consumers remain stationary in the network until departure, and locations have infinite capacities. The dynamic resource allocation problem is formulated as a stochastic optimal control problem with a long term average 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 the LLR policy. Sections 5 and 6 identify the fluid limit approximation of the network load under LLR as the solution to certain integral equations with boundary constraints. This solution converges to an optimal point in equilibrium, and Section 7 exploits this fact to establish the asymptotic optimality of LLR. Section 8 considers the case in which locations have finite capacities, and the resource allocation problem is defined as the minimization of blocking probability. It is shown that a class of *Least Relative Load Routing* (LRLR) policies asymptotically achieve the smallest blocking probability for large arrival rates. A connection with trunk reservation policies is discussed. Section 9 generalizes the results of Section 7 by considering an infinite capacity network in which consumers can migrate. Section 9 can be read independently of Section 8. A summary of conclusions and final remarks are collected in Section 10.

2 Preliminaries: The Static Load Balancing Problem

This section concerns the *static* load balancing problem, which plays an important role in the discussion of the dynamic load balancing problem of Section 3. The notation and results of this section follow Hajek (1990), though only separable cost functions are considered in that paper.

Given a load sharing network (U, V, N), we say that an *assignment* a, given by $(a_{u,v} : u \in U, v \in V)$, is *admissible* if $a \geq 0$ and $a_{u,v} = 0$ whenever $v \in N(u)^c$. Given a demand vector λ, an admissible assignment a *satisfies* demand λ if $\sum_v a_{u,v} = \lambda(u)$ for all $u \in U$. The *load* at 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.*

Let $\mathcal{A}_\lambda$ be the set of admissible assignments that satisfy demand λ. Let $\Phi : R^V \to R$ be a strictly convex, differentiable function which is symmetric in its arguments. The *static load balancing problem* (SLB) is defined as

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

The proofs of the following three lemmas can be found in the Appendix.

Lemma 2.1 *There exists a solution to $SLB(\lambda, \Phi)$. An assignment $a \in \mathcal{A}_\lambda$ is a solution if and only if for all $u \in U$, and all $v \in N(u)$*

$$a_{u,v} = 0 \quad whenever \quad q(v) > m_u(q), \tag{2.1}$$

where $m_u(q) = \min_{v \in N(u)} q(v)$. Furthermore, all such assignments yield the same load vector.

Lemma 2.1 implies that if $a_{u,v} > 0$ and $a_{u,v'} > 0$, then $x(v) = x(v')$. Thus, U and V can both be partitioned according to load levels, as indicated in the next lemma.

Lemma 2.2 *There exists a unique partition $\{V_1, V_2, \cdots, V_J\}$ of V, and a unique partition $\{U_1, U_2, \cdots, U_J\}$ of U such that for any assignment a satisfying condition (2.1), and the corresponding load vector q,*

$$q(v) = q(v') \qquad v, v' \in V_i \qquad i = 1, 2, \cdots, J \tag{2.2}$$

$$q(v) < q(v') \quad v \in V_i, \;\; v' \in V_j \quad i < j \tag{2.3}$$

$$a_{u,v} = 0 \quad v \in V_i, \;\; u \in U_j \quad i > j \tag{2.4}$$

$$N(u) \cap V_i = \emptyset \qquad u \in U_j \qquad i < j. \tag{2.5}$$

Let the function $\Psi : R_+^U \to R$ denote the value of SLB as a function of the demand vector, i.e., $\Psi(\lambda) = \Phi(q)$, where q is the unique load vector corresponding to the solutions of $SLB(\lambda, \Phi)$. The following lemma holds:

Lemma 2.3 *The function Ψ is convex.*

While this paper focuses on dynamic resource allocation, we comment briefly on the classification of problem SLB. The cost is convex and the constraint set is that of the basic assignment problem, well known to be a special case of the network flow problem, which in turn is a special case of a submodular system of constraints. As seen from Lemma 2.1, the solutions to SLB are the same for all Φ satisfying the specified assumptions, so that in solving SLB Φ can be taken to be quadratic and separable. Further, SLB with integer constraints can be solved by suitably rounding off solutions to SLB. See Hajek (1990) for discussion of SLB in particular, and Ibaraki and Katoh (1988) for polynomial time algorithms for the more general setting of constraints based on a submodular system.

3 The Basic Dynamic Load Balancing Model

Given a load sharing network (U, V, N), a demand vector λ, and a positive number γ, consider the following stochastic description of the network dynamics: For each $u \in U$ consumers of type u arrive according to a Poisson process of rate $\gamma\lambda(u)$, the processes for different types of arrivals being independent. In this section we assume that each location has infinite capacity; therefore, the network can accommodate every consumer immediately. This assumption is relaxed in Section 8, in which finite capacities are imposed on the locations. Each consumer has a holding time that is exponentially distributed with unit mean, independent of the past history. In the basic model it is also assumed that consumers do not change their types until they depart from the system. This assumption is relaxed in Section 9, which introduces a model such that consumers can migrate in the sense that their types change. Optimal repacking and Bernoulli splitting policies are discussed at the end of this section.

Let $X_t(v)$ denote the load at location $v \in V$ at time t, and set $X_t = (X_t(v) : v \in V)$. The consumer arrival and departure times, together with the allocation policy and an initial condition, determine the load process $X = (X_t : t \geq 0)$. In this paper, scaled loads will be considered; so in addition to the assumptions on Φ stated in Section 2, it is assumed that $\Phi(cx) = c^p\Phi(x)$ for all $c > 0$ and $x \in R^V$ for some $p > 1$. This implies that $\Psi(c\lambda) = c^p\Psi(\lambda)$ for $\lambda \in R_+^U$. A possible example for Φ is $\Phi(x) = \sum_v (x(v))^p$. The performance measure for an allocation policy π is the long-term average cost J_γ^π, defined by

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

The *dynamic load balancing problem* is to determine the set of allocation policies that minimize J_γ^π.

Let $L_t(u)$ denote the number of type $u \in U$ consumers in the network at time t, and set $L_t = (L_t(u) : u \in U)$. Note that consumer arrivals and departures, and hence the process $L = (L_t : t \geq 0)$, are not affected by the assignment decisions. Therefore, the value of problem $SLB(L_t, \Phi)$ yields a lower bound on the instantaneous cost at time t under *any* allocation policy. The process $(L_t(u) : t \geq 0)$ for fixed u is an $M/M/\infty$ queue length process with load factor $\gamma\lambda(u)$; hence the equilibrium distribution of L is described by a vector $(L_\infty(u) : u \in U)$ of Poisson random variables with mean vector $\gamma\lambda$. This implies the following lower bound on the cost of general allocation policies:

$$\begin{aligned} J_\gamma^\pi &\geq \liminf_{T\to\infty} E\left[\frac{1}{T}\int_0^T \Psi(L_t)dt \,|X_0 = x_0\right] \\ &= E\left[\Psi(L_\infty)\right] \\ &\geq \Psi(\gamma\lambda) \;=\; \gamma^p\Psi(\lambda). \end{aligned} \tag{3.1}$$

Here, the second inequality follows by Lemma 2.3 and Jensen's inequality.

While LLR is the allocation policy concentrated on in this paper, and it is defined in the next section, *Optimal repacking* (OR) and *Bernoulli splitting* (BS) are alternative allocation policies. *Repacking* a consumer entails changing its assigned location. The OR policy is a brute force approach whereby at each time t, the consumers in the network are repacked so as to solve the problem $SLB(L_t, \Phi)$, with the additional constraint that the assignment a have integer coordinates. This policy minimizes J_γ^π over all policies π. However it requires repacking of consumers, which, in some applications, may not be feasible due to operational constraints. Furthermore it can be implemented at a cost of $O(|V||U| + |V|^2)$ computations per consumer arrival and consumer departure, which may be impractical for large networks.

The BS policy, on the other hand, is a randomized nonrepacking policy under which each arriving type u consumer is assigned to location v with probability $a_{u,v}/\lambda(u)$, where a is an optimal assignment for the static problem $SLB(\lambda, \Phi)$. Since independent splitting of Poisson processes and merging of independent Poisson processes again yield Poisson processes, the load at any location under BS is an $M/M/\infty$ queue process. The limit of normalized cost $\gamma^{-p}J_\gamma^{BS}$ as $\gamma \to \infty$ is $\Psi(\lambda)$, which in view of (3.1) is the minimum possible normalized cost. Later sections of this paper establish the same asymptotic optimality property for the LLR policy. Thus, BS has the same asymptotic optimality property that is established for LLR. The BS policy also has the same asymptotic optimality properties that are established for LLR for networks with finite capacities and for networks with migration, as long as a solution to the corresponding static allocation problem is used to specify the routing probabilities in BS. However the BS policy explicitly uses the arrival rates; thus it is not robust to changes in the traffic parameters. Furthermore analysis more sensitive than the fluid limit approach taken in this paper can discriminate between the performances of the

BS and the LLR policies. For example Alanyali and Hajek (1996) considers a large-deviations type analysis of a simple network of three consumer types and two locations, and establishes that the BS policy has a higher overflow rate than the LLR policy.

4 Least Load Routing

The LLR policy, in the context of the basic dynamic load balancing model, is defined by the following assignment rule:

- When a type u consumer 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 consumer is assigned at random to one such location, each location having equal probability.

The LLR policy is a nonrepacking policy and costs $|N(u)|$ comparisons per consumer arrival of type $u \in U$. 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, LLR is robust with respect to the network demand. On the other hand, LLR is a myopic allocation policy and is not necessarily optimal for finite arrival rates. The LLR policy has been studied by a number of authors and has been shown to have a poor worst-case performance relative to the optimal nonrepacking policy (see Azar et al. (1992)).

Under LLR, the load process X is Markov on the state space Z_+^V. For $v \in V$, define the operator $T_v \,:\, Z_+^V \to Z_+^V$ as

$$(T_v x)(v') = \begin{cases} x(v') + 1 & \text{if } v' = v, \\ x(v') & \text{else} \end{cases}$$

for all $x \in Z_+^V$. Then the off-diagonal entries of the generator matrix of X are given by

$$Q(x,y) = \begin{cases} \sum_{u \in N^{-1}(v)} \gamma\lambda(u) \frac{I\{x(v)=m_u(x)\}}{\sum_{v' \in N(u)} I\{x(v')=m_u(x)\}} & \text{if } y = T_v x \\ x(v) & \text{if } y = T_v^{-1} x \\ 0 & \text{else,} \end{cases} \tag{4.1}$$

where $N^{-1}(v) = \{u \in U : v \in N(u)\}$.

In principle, given a load sharing network, one can compute the equilibrium distribution of X and thereby the cost incurred under the LLR policy. However, it is computationally intractable to obtain an expression for the cost of LLR for arbitrary networks through an expression for the equilibrium distribution. As an alternative approach, we study the network for large values of the parameter γ and, by obtaining fluid limit approximations, evaluate the performance of the LLR policy for arbitrary network topologies.

5 Convergence

This section addresses the weak convergence of the network load as γ tends to infinity. The main result, Lemma 5.3, characterizes the possible weak limits of the load process, properly normalized, as solutions to certain fluid equations.

Let the *normalized load process* X^γ be defined as $X^\gamma = \gamma^{-1}X$, where X denotes the network load under the LLR policy. Assume the existence of a finite number K such that $E[\sum_v X_0^\gamma(v)] \le K$ for all γ. For $x \in R^V$ define

$$a_{u,v}(x) = \frac{I\{x(v) = m_u(x)\}}{\sum_{v' \in N(u)} I\{x(v') = m_u(x)\}}\lambda(u),$$

and let

$$A_{u,v}^\gamma(t) = \int_0^t a_{u,v}(X_s^\gamma)ds. \tag{5.1}$$

For each $v \in V$, the drift of the process $X^\gamma(v)$ at time t, given that $X_t^\gamma = x$, is $\gamma^{-1}\sum_{y \in R^V}(y(v) - \gamma x(v))$ $Q(\gamma x, y)$, which by (4.1) is given by $(\sum_{u \in N^{-1}(v)} a_{u,v}(x)) - x(v)$. Therefore, the process $M^\gamma(v)$ defined implicitly by

$$X_t^\gamma(v) = X_0^\gamma(v) + M_t^\gamma(v) + \sum_{u \in N^{-1}(v)} A_{u,v}^\gamma(t) - \int_0^t X_s^\gamma(v)ds \tag{5.2}$$

is a local martingale with $M_0^\gamma(v) = 0$. (See Section 4.7.B and Problem 4.11.15 of Ethier and Kurtz (1986)).

Lemma 5.1 *For $v \in V$, $M^\gamma(v)$ is a square integrable martingale, and*

$$E[(M_t^\gamma(v))^2] \leq \frac{1}{\gamma}(2t\sum_u \lambda(u) + K). \tag{5.3}$$

Proof. Let $\tau_n = \inf\{t : M_t^\gamma(v) \geq n\}$. Since the local martingale $M^\gamma(v)$ has jumps of size γ^{-1}, the process $M_{t\wedge\tau_n}^\gamma(v)$ is bounded and hence is a square integrable martingale. Thus,

$$\begin{aligned} E[(M_{t\wedge\tau_n}^\gamma(v))^2] &= E[[M^\gamma(v)]_{t\wedge\tau_n}] \\ &\leq \frac{1}{\gamma^2}E[\text{number of jumps of } X^\gamma(v) \text{ in } [0,t]] \\ &\leq \frac{1}{\gamma}(2t\sum_u \lambda(u) + K), \end{aligned}$$

where $[M^\gamma(v)]$ is the quadratic variation process of $M^\gamma(v)$. Fatou's Lemma implies (5.3). Finally (5.3) implies that $M^\gamma(v)$ over any finite interval is uniformly integrable; hence it is a martingale. $\square$

Remark 5.1 *By Doob's L^2 inequality and Lemma 5.1,*

$$E[\sup_{0\leq s\leq t}(M_s^\gamma(v))^2] \leq 4E\left[(M_t^\gamma(v))^2\right] = O(\gamma^{-1}).$$

Therefore, $M^\gamma(v) \Longrightarrow 0$ for all $v \in V$.

Lemma 5.2 *(Tightness) If the sequence $(X_0^\gamma : \gamma > 0)$ is tight, then the sequence of processes $((X^\gamma, A^\gamma) : \gamma > 0)$ is tight.*

Proof. By Proposition 3.2.4 of Ethier and Kurtz (1986) it suffices to show the tightness of (X^γ) and (A^γ) separately. Towards this end, the observation

$$\begin{aligned} A_{u,v}^\gamma(0) &= 0, \\ 0 \leq A_{u,v}^\gamma(t) - A_{u,v}^\gamma(s) &\leq (t-s)\lambda(u) \end{aligned}$$

for $t \geq s$, yields the tightness of (A^γ) (Ethier and Kurtz (1986), Theorem 3.7.2). This, along with Remark 5.1 and the representation (5.2), implies that to establish tightness of (X^γ) it suffices to establish the tightness of $(\int_0^\cdot X_s^\gamma(v)ds)$. Observe that if $0 \leq a, b \leq t$, then

$$\left| \int_0^a X_s^\gamma(v)ds - \int_0^b X_s^\gamma(v)ds \right| \leq |a-b| \left(\sup_{0\leq s\leq t} X_s^\gamma(v) \right).$$

Note also that

$$\begin{aligned} E[\sup_{0\leq s\leq t} X_s^\gamma(v)] &\leq E[\frac{1}{\gamma}(\text{total number of arrivals in [0,t]}) + \sum_v X_0^\gamma(v)] \\ &\leq t\sum_u \lambda(u) + K, \end{aligned}$$

thus Markov's inequality yields that for each $\eta > 0$

$$P\left(\sup_{0\leq s\leq t} X_s^\gamma(v) > \frac{t\sum_u \lambda(u) + K}{\eta} \right) < \eta, \tag{5.4}$$

and the desired result follows by Theorem 3.7.2 of Ethier and Kurtz (1986). □

Lemma 5.3 *(Convergence of Subsequences and Fluid Equations) If $X_0^\gamma \Longrightarrow x_0$, then every subsequence $(X^{\gamma_n}, A^{\gamma_n})$ has a further subsequence $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}})$ such that $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}}) \Longrightarrow (x, A)$, where (x, A) satisfies the following fluid equations :*

$$x_t(v) = x_0(v) + \sum_{u\in N^{-1}(v)} A_{u,v}(t) - \int_0^t x_s(v)ds \tag{5.5}$$

$$A_{u,v}(0) = 0, \ \ A_{u,v}(t) \text{ nondecreasing}, \ \sum_{v\in N(u)} A_{u,v}(t) = \lambda(u)t, \tag{5.6}$$

$$\int_0^t I\{x_s(v) > m_u(x_s)\} dA_{u,v}(s) = 0. \tag{5.7}$$

Proof. Let (γ_{n_k}) be a subsequence of $\gamma_n \to \infty$ such that $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}})$ converges weakly. Let (x, A) denote the limit. By Skorokhod's theorem (Ethier and Kurtz (1986), Theorem 3.1.8), the processes can be constructed on the same probability space such that the convergence is almost everywhere. The limit x is continuous with probability one and the convergence is uniform on compact time sets (Ethier and Kurtz (1986), Lemmas 3.10.2 and 3.10.1 respectively); therefore,

$$\lim_{n\to\infty}\int_0^t X^{\gamma_n}(s)ds = \int_0^t x(s)ds$$

with probability one. Since $M^{\gamma_n} \to 0$, (x, A) satisfies Equation (5.5) with the initial condition x_0. By (5.1), $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}})$ satisfies conditions (5.6) and (5.7) for all k. The relation (5.6) defines a closed subset in the Skorokhod topology, hence it is satisfied by the limit A. Since

$$\int_0^t I\{X_s^\gamma(v) > m_u(X_s^\gamma)\} dA_{u,v}^\gamma(s) = 0,$$

it follows that

$$\int_0^t (X_s^\gamma(v) - m_u(X_s^\gamma)) \wedge 1 \; dA_{u,v}^\gamma(s) = 0. \tag{5.8}$$

By Lemma 2.4 of Williams and Dai (1995), we can take a limit in (5.8) along the subsequence (γ_{n_k}) so that (x, A) satisfies (5.7). This establishes the lemma. □

6 Analysis of the Fluid Limit

In this section we concentrate on the solutions to the fluid equations (5.5)-(5.7), existence of which is known due to Lemma 5.3. In particular, via a monotonicity argument, Lemma 6.2 establishes that there is a unique load trajectory that solves the fluid equations, and Lemma 6.4 identifies the limit point of this trajectory. We start with a remark.

Remark 6.1 *Equation (5.6) implies that $A_{u,v}$ has a density $a_{u,v}$ such that $\sum_{v\in N(u)} a_{u,v}(t) = \lambda(u)$ for almost all $t \geq 0$. Therefore, x and A are almost everywhere differentiable, and whenever the derivatives exist, $\dot{A}_{u,v}(t) = a_{u,v}(t)$, $\dot{x}_t(v) = (\sum_u a_{u,v}(t)) - x_t(v)$, and $I\{x_t(v) > m_u(x_t)\}a_{u,v}(t) = 0$.*

Lemma 6.1 *(Monotonicity) Suppose (x', A') and (x, A) are two solutions to the fluid equations (5.5)-(5.7) with $x_0'(v) \geq x_0(v)$ for all $v \in V$. Then $x_t'(v) \geq x_t(v)$ for all $v \in V$ and $t \geq 0$.*

Proof. To prove the claim by contradiction, suppose that the conclusion is false. Take $\epsilon > 0$ so that t_1 defined as follows is finite:

$$t_1 = \inf\{t \geq 0 : \ x_t(v) - x'_t(v) \geq \epsilon \text{ for some location } v \in V\}.$$

Since x'_t and x_t are continuous, the set F defined as follows is nonempty:

$$F = \{v \in V : \ x_{t_1}(v) = x'_{t_1}(v) + \epsilon\}.$$

Let $\epsilon' = \max\{x_{t_1}(v) - x'_{t_1}(v) : \ v \in F^c\}$ and ϵ_0, ϵ_1 be such that $\epsilon' < \epsilon_0 < \epsilon_1 < \epsilon$, and $\epsilon_1 > 0$. By the continuity of solutions, there exists t_0 with $0 \leq t_0 < t_1$ such that

$$x_s(v) - x'_s(v) \geq \epsilon_1 \text{ for } v \in F,\ s \in [t_0, t_1) \tag{6.1}$$

$$x_s(v) - x'_s(v) \leq \epsilon_0 \text{ for } v \in F^c,\ s \in [t_0, t_1). \tag{6.2}$$

Note that $\sum_{v \in F} x_{t_0}(v) - x'_{t_0}(v) < |F|\epsilon = \sum_{v \in F} x_{t_1}(v) - x'_{t_1}(v)$ so that

$$\sum_{v \in F} \left(x_{t_1}(v) - x_{t_0}(v)\right) > \sum_{v \in F} \left(x'_{t_1}(v) - x'_{t_0}(v)\right).$$

This, together with (5.5) and (6.1), implies the existence of a $u \in U$ such that

$$\int_{t_0}^{t_1} \sum_{v \in N(u) \cap F} a_{u,v}(s) ds > \int_{t_0}^{t_1} \sum_{v \in N(u) \cap F} a'_{u,v}(s) ds. \tag{6.3}$$

By Remark 6.1, for almost all $s \in [t_0, t_1)$ such that the integrand of the left-hand side of (6.3) is positive,

$$\min_{v \in N(u) \cap F} x_s(v) \leq \min_{v \in N(u) \cap F^c} x_s(v). \tag{6.4}$$

In view of (6.1) and (6.2), this implies that

$$\min_{v \in N(u) \cap F} x'_s(v) < \min_{v \in N(u) \cap F^c} x'_s(v). \tag{6.5}$$

Thus, for almost all such s, the integrand of the right-hand side of (6.3) equals $\lambda(u)$, which is an upper bound to the integrand of the left-hand side. This contradicts (6.3) and hence also the existence of t_1 for any $\epsilon > 0$. $\square$

Lemma 6.2 *(Uniqueness of Load Trajectory) If (x, A) and (x', A') are two solutions to the fluid equations (5.5)-(5.7) with $x_0 = x'_0$, then $x_t = x'_t$ for all $t \geq 0$.*

Proof. Use Lemma 6.1 twice with $x'_0 \leq x_0$ and $x'_0 \geq x_0$. □

Remark 6.2 *Note that the fluid equations and the initial state x_0 do not necessarily determine A uniquely. For a simple illustration, suppose that $V = U = \{0, 1\}$, $N(u) = V$ for $u \in \{0, 1\}$, and $\lambda = (1, 1)$. Let $A_{u,v}(t) = t/2$, and $\tilde{A}_{u,v}(t) = I\{u = v\}t$. Both (x, A) and $(x, \tilde{A})$ satisfy the fluid equations with $x_0(v) = 0$ and $x_t(v) = 1 - e^{-t}$ for all v.*

The uniqueness result of Lemma 6.2 removes the need to pass to a subsequence for the convergence of X^γ in Lemma 5.3.

Corollary 6.1 *If $X_0^\gamma \Longrightarrow x_0$, then $X^\gamma \Longrightarrow x$, where for some process A, (x, A) is a solution of the fluid equations (5.5)-(5.7) with the initial condition x_0.*

Let a be an assignment that solves the static problem $SLB(\lambda, \Phi)$ with the corresponding load q. It is easy to verify that $(q(1 - e^{-t}), at)$ is a solution to the fluid equations with zero initial state and that this solution converges to q exponentially fast as $t \to \infty$. The next two lemmas show that starting from *any* initial state x_t converges to q exponentially fast.

Lemma 6.3 *Let (U, V, N) be an arbitrary load sharing network. For any (x, A) that satisfies (5.5) and (5.6),*

$$\sum_v x_t(v) = \sum_v x_0(v) e^{-t} + \sum_u \lambda(u)(1 - e^{-t}).$$

In particular, $\lim_{t \to \infty} \sum_v x_t(v) = \sum_u \lambda(u)$ uniformly for all x_0 in bounded subsets of R^V.

Proof. Equations (5.5) and (5.6) yield the integral equation

$$\sum_v x_t(v) = \sum_v x_0(v) + t \sum_u \lambda(u) - \int_0^t \sum_v x_s(v) ds,$$

which yields the desired result. □

Lemma 6.4 *(Insensitivity to Initial State) Let (x, A) be a solution to the fluid equations with $x_0 \geq 0$. Then*

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

where $\|\cdot\|_{sup}$ denotes the supremum norm. In particular, $\lim_{t\to\infty}\|x_t - q\|_{sup} = 0$ uniformly for all x_0 in bounded subsets of R_+^V.

Proof. Let $v \in V$ be arbitrary. By Lemma 6.1, $x_t(v) \geq q(v)(1-e^{-t})$; thus $0 \leq x_t(v) - q(v)(1-e^{-t}) \leq \sum_{v'}(x_t(v') - q(v')(1-e^{-t}))$. By Lemma 6.3, $\sum_{v'}(x_t(v') - q(v')(1-e^{-t})) = \sum_{v'} x_0(v')e^{-t}$; therefore,

$$-q(v)e^{-t} \leq x_t(v) - q(v) \leq (-q(v) + \sum_{v'} x_0(v'))e^{-t}.$$

This establishes the lemma. □

7 Asymptotic Optimality of Least Load Routing

This section establishes the asymptotic optimality of LLR for the optimal control problem formulated in Section 3. In Section 6 it was shown that the finite dimensional distributions of the normalized load process converge as $\gamma \to \infty$, and the limit process converges to an optimal point q as $t \to \infty$. Lemma 7.2 establishes the convergence of the equilibrium distribution of the normalized load process to the deterministic distribution concentrated at q. These facts are used to prove Theorem 7.1 on the asymptotic optimality of LLR.

In what follows, P_μ denotes the distribution of the process X^γ when X_0^γ has distribution μ. Also, μ_0 is the deterministic distribution concentrated at the zero state, and μ_t^γ denotes the distribution of X_t^γ given $X_0^\gamma = 0$. We start with an auxiliary lemma.

Lemma 7.1 *Given $\epsilon > 0$, there exists a γ_ϵ such that whenever $\gamma > \gamma_\epsilon$,*

$$P_{\mu_0}\left(\sum_v X_t^\gamma(v) \leq \epsilon + \sum_u \lambda(u)\right) \geq 1 - \epsilon \quad \textit{for any } t \geq 0.$$

Proof. Starting from the zero state, the total load in the system at any time $t > 0$ is stochastically dominated by a Poisson random variable with mean $\gamma \sum_u \lambda(u)$. Chebychev's inequality yields the desired result. $\square$

Lemma 7.2 *(Convergence of Equilibrium Distributions) Let q be the unique load vector corresponding to solutions of $SLB(\lambda, \Phi)$ and ν be the distribution of the equilibrium load X_∞^γ. Then for all $\epsilon > 0$,*

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

Proof. Let $\epsilon > 0$ be fixed. Note that $\nu(\|X_\infty^\gamma - q\|_{sup} > \epsilon) \leq \liminf_{T\to\infty} P_{\mu_0}(\|X_T^\gamma - q\|_{sup} > \epsilon)$ so that it suffices to show that

$$P_{\mu_0}(\|X_T^\gamma - q\|_{sup} > \epsilon) < \epsilon$$

for all sufficiently large T and γ. Towards this end, appeal to Lemma 6.4 to fix θ so that $\|x_t - q\|_{sup} < \epsilon/2$ whenever $t \geq \theta$ and $\sum_v x_0(v) < \epsilon + \sum_u \lambda(u)$. By the time homogeneous Markov property of X^γ,

$$P_{\mu_0}\left(\|X_T^\gamma - q\|_{sup} > \epsilon\right) = P_{\mu_{T-\theta}^\gamma}\left(\|X_\theta^\gamma - q\|_{sup} > \epsilon\right)$$

whenever $T \geq \theta$. Therefore, to prove the lemma, it suffices to establish the following claim: There exists a γ_ϵ such that whenever $\gamma > \gamma_\epsilon$,

$$P_{\mu_{T-\theta}^\gamma}\left(\|X_\theta^\gamma - q\|_{sup} > \epsilon\right) < \epsilon \quad \text{for all } T > \theta. \tag{7.1}$$

To argue by contradiction, suppose that the claim is false. Then one can construct a sequence $(\tilde{\mu}^\xi)$ with $\xi \to \infty$, such that $\tilde{\mu}^\xi = \mu_{t(\xi)}^\xi$ for some choice of $t(\xi) > 0$, and

$$P_{\tilde{\mu}^\xi}\left(\|X_\theta^\xi - q\|_{sup} > \epsilon\right) \geq \epsilon. \tag{7.2}$$

By Lemma 7.1, $(\tilde{\mu}^\xi)$ is tight; therefore, by Lemma 5.3, there exists a subsequence $\xi_n \to \infty$ such that if $X_0^{\xi_n} \sim P_{\tilde{\mu}^{\xi_n}}$, then $X^{\xi_n} \Longrightarrow x$, for some x as in Lemma 5.3. Hence there exists an n_ϵ such that

$$P_{\tilde{\mu}^{\xi_n}}\left(\|X_\theta^{\xi_n} - x_\theta\|_{sup} > \epsilon/2\right) < \epsilon/2 \tag{7.3}$$

whenever $n > n_\epsilon$. However, by the choice of θ and Lemma 7.1, for all ξ_n sufficiently large,

$$P_{\tilde{\mu}^{\xi_n}}\left(\|x_\theta - q\|_{sup} > \epsilon/2\right) < \epsilon/2. \tag{7.4}$$

Observations (7.3) and (7.4) contradict (7.2), hence proving (7.1), which establishes the lemma. □

Theorem 7.1 *(Asymptotic Optimality of LLR) Given an allocation policy π, let J_γ^π denote the cost under π when the network demand is $\gamma\lambda$. Then*

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

Proof. Note that in equilibrium $\gamma X_\infty^\gamma(v)$ is stochastically dominated by a Poisson random variable with mean $\gamma \sum_u \lambda(u)$, for all $v \in V$. Consequently, $E\left[(X_\infty^\gamma(v))^{p+1}\right]$ is bounded independently of γ. This, along with the observation that $\Phi(X_\infty^\gamma) \leq \max\{\Phi(u) : \|u\|_{sup} \leq 1\}\|X_\infty^\gamma\|_{sup}^p$, implies that $(\Phi(X_\infty^\gamma) : \gamma > 0)$ is uniformly integrable. Thus, by Lemma 7.2,

$$\begin{aligned} \lim_{\gamma\to\infty} \gamma^{-p} J_\gamma^{LLR} &= \lim_{\gamma\to\infty} E\left[\Phi(X_\infty^\gamma)\right] \\ &= \Phi(q) \;=\; \Psi(\lambda) \;>\; 0. \end{aligned}$$

Inequality (3.1) implies that for any allocation policy π, $\gamma^{-p} J_\gamma^\pi \geq \Psi(\lambda)$ for all $\gamma > 0$. This proves the theorem. □

8 Finite Capacities

This section considers a variation of the basic model in which each location has a finite capacity. Namely, we assume that the load of a location cannot exceed its capacity, and arrivals to the congested neighborhoods are dropped. In this setting, a natural objective for the allocation policy is to minimize the percentage of consumers dropped in the system. We concentrate on a broad class of practical allocation policies, namely the *least* relative *load routing* policies, in which new consumers are assigned to the location with the least relative load. The relative load of a location is defined by applying a normalization function to the actual load. Theorem 8.1 establishes that such policies asymptotically achieve the smallest loss probability for large arrival rates. We then provide stronger results on two members of this class, namely the *least ratio routing* (LRR) and the *maximum residual capacity routing* (MRCR) policies.

The use of a binary valued normalization function would model trunk reservation strategies, studied in a similar context by Hunt and Kurtz (1994). However, we require the normalization functions to be strictly increasing; thus, trunk reservation strategies are not covered in this paper. In doing so, we avoid the pathologies associated with trunk reservations in heavy traffic encountered by Hunt and Kurtz (1994), and can therefore establish optimality results. The drawback of our approach is that more feedback information about the network state is required to implement the allocation policies. The work of Hunt and Laws (1995) resolves the problems set in Hunt and Kurtz (1994) for the case of a single resource location.

To describe the dynamic model of interest, let a *capacity vector* $\kappa = (\kappa(v) \ : \ v \in V)$ be a vector of positive numbers. Given a load sharing network (U, V, N), a capacity vector κ, and a load vector λ, consider the limiting regime of Section 5. Suppose that in each system indexed by γ, each location v has capacity $\lfloor \gamma\kappa(v) \rfloor$. A location is called *full* if its load and capacity are equal, and a consumer is *lost* if, upon its arrival, all of the locations in its neighborhood are full. Lost consumers cannot be assigned later; hence they are treated by the system as if they never arrived. The problem of interest is to find allocation policies that minimize the *loss probability* $P_\gamma(Loss)$, which is defined as

$$P_\gamma(Loss) \quad = \quad \liminf_{t\to\infty} \frac{E[\text{number of consumers lost in } [0,t]]}{t\gamma \sum_u \lambda(u)}.$$

We start with some definitions regarding an associated static problem. Given a capacity vector κ, define $\mathcal{B}_{\lambda,\kappa}$ as the set of admissible assignment vectors a such that $\sum_v a_{u,v} \leq \lambda(u)$ for all u, and $q(v) \leq \kappa(v)$ for all v, where q denotes the load vector determined by assignment a. The *static load packing problem* (SLP) is the simple assignment problem defined as

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

Towards the end of characterizing certain solutions to SLP, we have the following definition:

Definition 8.1 *A function $f : R^V \times V \to R$ is called a* normalization function *if for all $v \in V$, the real valued function $f(\cdot, v)$ has the following three properties:*

(i) $f(q, v)$ *depends on q only through $q(v)$,*

(ii) $f(q, v)$ *is a strictly increasing and continuously differentiable function of $q(v)$, such that $\partial f(q, v)/\partial q(v) \geq \delta$ for some $\delta > 0$,*

(iii) $f(q, v) = 0$ *when $q(v) = \kappa(v)$.*

The interpretation of $f(q, v)$ for $0 \leq q(v) \leq \kappa(v)$ is that larger values of $f(q, v)$ (i.e. values closer to zero) represent heavier load. Further, if $f(q, v) > f(q, v')$ then v is considered more heavily loaded than v', under the normalizing function f. Two examples to be considered in more detail later are $f(q, v) = -1 + q(v)/\kappa(v)$ (corresponds to least ratio loading) and $f(q, v) = q(v) - \kappa(v)$ (corresponds to maximum residual capacity loading).

Consider the following two conditions on a generic assignment a, where q denotes the load vector corresponding to a, and $m_u(f(q)) = \min_{v \in N(u)} f(q, v)$:

Condition 8.1 $a_{u,v} = 0$ *whenever* $f(q, v) > m_u(f(q))$.

Condition 8.2 $\sum_v a_{u,v} < \lambda(u)$ *only if* $f(q, v) = 0$ *for all* $v \in N(u)$.

Let the function $\Phi : R^V \to R$ be defined as $\Phi(q) = \sum_v \int_0^{q(v)} f(\tilde{q}, v) d\tilde{q}(v)$. Note that Φ is convex, however, not necessarily symmetric in its arguments. The following three lemmas are proved in the Appendix. The first lemma concerns a static load *balancing* problem, the second concerns a connection between static load balancing and load packing, and the third gives a sufficient condition for optimality in $SLP(\lambda, \kappa)$.

Lemma 8.1 *(Load Balancing) There exists a solution to $SLB(\lambda, \Phi)$. An admissible assignment $\tilde{a}$ which satisfies demand λ solves the $SLB(\lambda, \Phi)$ if and only if $\tilde{a}$ satisfies Condition 8.1 . Furthermore, all such assignments yield the same load vector.*

Lemma 8.2 *(Truncation) Let $\tilde{a}$ solve $SLB(\lambda, \Phi)$ with the corresponding load vector $\tilde{q}$, and let a be the assignment defined by*

$$a_{u,v} = \tilde{a}_{u,v} \left(1 \wedge \frac{\kappa(v)}{\tilde{q}(v)} \right).$$

Then $a \in \mathcal{B}_{\lambda,\kappa}$ and a satisfies Conditions 8.1 and 8.2 with the corresponding load vector $q(v) = \tilde{q}(v) \wedge \kappa(v)$.

Lemma 8.3 *(Sufficiency) An assignment vector $a \in \mathcal{B}_{\lambda,\kappa}$ solves $SLP(\lambda, \kappa)$ if there exists a normalization function f such that both Conditions 8.1 and 8.2 hold. For a given normalization function, there exists an assignment $a \in \mathcal{B}_{\lambda,\kappa}$ that satisfies Conditions 8.1 and 8.2, and all such assignments yield the same load vector.*

To treat the lossy network in the context of the already existing theory, we introduce a new location v_L and define an extended load sharing network $(\hat{U}, \hat{V}, \hat{N})$ by $\hat{U} = U$, $\hat{V} = V \cup \{v_L\}$, and $\hat{N}(u) = N(u) \cup \{v_L\}$, where $\kappa(v_L) = \infty$. A load process $(X(v) : v \in V)$ corresponding to an allocation policy π can be extended to a load process $(X(v) : v \in \hat{V})$ on $(\hat{U}, \hat{V}, \hat{N})$ by letting $X_t(v_L)$ denote the number of blocked consumers that would have been in service at time t if they were not blocked. We continue to use X to denote the extended process, and let $X^\gamma(v) = \gamma^{-1} X(v)$ for all $v \in \hat{V}$.

The solution of $SLP(\lambda, \kappa)$ provides a lower bound for the loss probability of *any* allocation policy.

Lemma 8.4 *Let π be an arbitrary allocation policy, and let $P_\gamma^\pi(Loss)$ denote the loss probability under policy π. Then*

$$P_\gamma^\pi(Loss) \geq 1 - \frac{\sum_{v \in V} q(v)}{\sum_u \lambda(u)},$$

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

Proof. Consider the load process X on $(\hat{U}, \hat{V}, \hat{N})$, and assume, without loss of generality, that X starts with the zero initial state. Let

$$\begin{aligned} s_k &= \text{Holding time of the } k^{th} \text{ arrival to } v_L, \\ \beta(t) &= \text{Number of arrivals to } v_L \text{ in } [0, t). \end{aligned}$$

Note that $\beta(t)$ is the number of consumers lost by time t in the original system (U,V,N). By the construction of X,

$$\begin{aligned} E[\int_0^t \sum_{v\in\hat{V}} X_s(v)ds] - E[\int_0^t \sum_{v\in V} X_s(v))ds] &= E[\int_0^t X_s(v_L)ds] \\ &\leq E[\sum_{k=1}^{\beta(t)} s_k] \\ &= E[\beta(t)], \end{aligned} \tag{8.1}$$

where the last step follows by the independence of $(s_k : k \geq 1)$ and $\beta(t)$. For every s, $E[X_s^\gamma]$ is the load vector corresponding to an assignment in $\mathcal{B}_{\lambda,\kappa}$; therefore, by the choice of q, $\sum_{v\in V} E[X_s^\gamma(v)] \leq \sum_{v\in V} q(v)$ so that $E[\int_0^t \sum_{v\in V} X_s(v)ds] \leq t\gamma \sum_{v\in V} q(v)$. The total number of consumers in $\hat{V}$ is the number in an $M/M/\infty$ queue so that $E[\int_0^t \sum_{v\in\hat{V}} X(s)ds] = \gamma \sum_u \lambda(u) \int_0^t (1-e^{-s})ds$. These observations and rearranging inequality (8.1) yield

$$\frac{E[\beta(t)]}{t\gamma} \geq (1 - \frac{1-e^{-t}}{t}) \sum_u \lambda(u) - \sum_{v\in V} q(v).$$

The result follows by dividing each side by $\sum_u \lambda(u)$ and letting $t \to \infty$. □

Consider the randomized, nonrepacking policy which assigns each type u consumer to location v with probability $a_{u,v}/\lambda(u)$, where a is a solution to $SLP(\lambda,\kappa)$. It is easy to see that this policy achieves the lower bound established by Lemma 8.4, asymptotically as γ tends to infinity. However it has the same drawbacks as the randomized policies discussed in Section 3. In this section we focus on the class *least relative load routing* (LRLR) of allocation policies, and show that they also achieve the minimum consumer loss probability for large values of γ. Given a normalization function f, an LRLR policy is defined by the following assignment rule:

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

Consider the extended load process $(X(v) : v \in \hat{V})$ under an LRLR policy. Intuitively, this process is lossless, and it is also governed by LRLR with the normalization function extended by defining $f(q, v_L) = 0^-$. The process X^γ is Markov and has the following representation:

$$X_t^\gamma(v) = X_0^\gamma(v) + M_t^\gamma(v) + \sum_{u \in \hat{N}^{-1}(v)} A_{u,v}^\gamma(t) - \int_0^t X_s^\gamma(v) ds,$$

where

$$A_{u,v}^\gamma(t) = \begin{cases} \int_0^t \frac{I\{f(X_s,v)=m_u(f(X_s))\} I\{f(X_s,v)<0\}}{\sum_{v' \in N(u)} I\{f(X_s,v')=m_u(f(X_s))\}} \lambda(u) ds & \text{if } v \in V \\ \int_0^t I\{m_u(f(X_s)) = 0\} \lambda(u) ds & \text{if } v = v_L, \end{cases}$$

and $M^\gamma(v)$ is a local martingale with $M_0^\gamma(v) = 0$. Given that (X_0^γ) is tight, the methods of Section 5 can be applied to establish the tightness of (X^γ, A^γ) and characterize the limits of weakly convergent subsequences. Namely, the following lemma holds:

Lemma 8.5 *Suppose (X_0^γ) is tight. Then every subsequence $(X^{\gamma_n}, A^{\gamma_n})$ has a further subsequence $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}})$ such that $(X^{\gamma_{n_k}}, A^{\gamma_{n_k}}) \Longrightarrow (x, A)$, where (x, A) satisfies the following fluid equations:*

$$x_t(v) = x_0(v) + \sum_{u \in \hat{N}^{-1}(v)} A_{u,v}(t) - \int_0^t x_s(v) ds, \tag{8.2}$$

$$A_{u,v}(0) = 0, \quad A_{u,v}(t) \textit{ nondecreasing}, \tag{8.3}$$

$$0 \le x_t(v) \le \kappa(v), \qquad \sum_{v \in \hat{N}(u)} A_{u,v}(t) = \lambda(u) t, \tag{8.4}$$

$$\int_0^t I\{f(x_s, v) > m_u(f(x_s))\} dA_{u,v}(s) = 0 \qquad v \ne v_L, \tag{8.5}$$

$$\int_0^t I\{m_u(f(x_s)) < 0\} dA_{u,v_L}(s) = 0. \tag{8.6}$$

Call t a *regular point* of a function $g : R \to R$ if g is differentiable at t, and let $\dot{g}_t$ denote the derivative of g at a regular point t. The following lemma is proved in the Appendix.

Lemma 8.6 *Let g be an absolutely continuous function, and let $\alpha \in R$. Then $\{t : g_t = \alpha\} \subset \{t : \dot{g}_t = 0\} \cup N_\alpha$, where N_α is a set of Lebesgue measure zero.*

Lemma 8.7 *(Monotonicity) Let (x', A') and (x, A) be two solutions to the fluid equations (8.2)-(8.6). Then,* *(i) If $x'_0(v) \geq x_0(v)$ for all $v \in V$, then $x'_t(v) \geq x_t(v)$ for all $v \in V$ and $t \geq 0$.*

(ii) If in addition $x'_0(v_L) \geq x_0(v_L)$, then $x'_t(v_L) \geq x_t(v_L)$ for all $t \geq 0$.

Proof. For (i), the proof of Lemma 6.1 applies directly by replacing F^c by $F^c \cap V$ and using $f(x_s, v)$ and $f(x'_s, v)$ in place of $x_s(v)$ and $x'_s(v)$ in inequalities (6.4) and (6.5), respectively.

To prove (ii), for each t define

$$F_t = \{v \in V : x_t(v) = \kappa(v)\}, \quad F'_t = \{v \in V : x'_t(v) = \kappa(v)\}, \tag{8.7}$$

$$G_t = \{u \in U : N(u) \subset F_t\}, \quad G'_t = \{u \in U : N(u) \subset F'_t\}. \tag{8.8}$$

By part (i), $F_t \subset F'_t$, and $G_t \subset G'_t$ for all t. Remark 6.1 applied to (8.2)-(8.6) and Lemma 8.6 with $q_t = x_t(v)$ and $\alpha = \kappa(v)$ yield that for almost all t,

$$\begin{aligned} (\sum_{u \in G_t} a_{u,v}(t)) - \kappa(v) = \dot{x}_t(v) = 0 \quad & \text{for all } v \in F_t, \\ (\sum_{u \in G'_t} a'_{u,v}(t)) - \kappa(v) = \dot{x}'_t(v) = 0 \quad & \text{for all } v \in F'_t. \end{aligned} \tag{8.9}$$

Therefore, for almost all t,

$$\begin{aligned} \dot{x}_t(v_L) &= \dot{x}_t(v_L) + \sum_{v \in F_t} \dot{x}_t(v) \\ &= \sum_{u \in G_t} \lambda(u) - (\sum_{v \in F_t} \kappa(v)) - x_t(v_L), \\ \dot{x}'_t(v_L) &= \dot{x}'_t(v_L) + \sum_{v \in F_t} \dot{x}'_t(v) \\ &\leq \sum_{u \in G_t} \lambda(u) - (\sum_{v \in F_t} \kappa(v)) - x'_t(v_L), \end{aligned}$$

where the inequality follows by (8.9) and the definitions (8.7) and (8.8). Hence if $e_t = x'_t(v_L) - x_t(v_L)$, then $\dot{e}_t \geq -e_t$ for almost all t. This, along with the hypothesis $e_0 \geq 0$, proves (ii). □

Corollary 8.1 *(Uniqueness) If (x, A) and (x', A') are two solutions to the fluid equations (8.2)-(8.6) with $x_0 = x'_0$, then $x_t = x'_t$ for all $t \geq 0$.*

We now concentrate on the properties of the unique trajectory x that corresponds to the solutions of the fluid equations (8.2)-(8.6). The proof of the following lemma can be found in the Appendix.

Lemma 8.8 *Let $g_t(i)$ be absolutely continuous, $i = 1, 2, \cdots, I$, and set $m_t = \min_i g_t(i)$. Then m is absolutely continuous, almost every t is a regular point for $g(1), \cdots, g(I), m$, and for all such t, $\dot{m}_t = \dot{g}_t(i)$ for all i such that $g_t(i) = m_t$.*

Note that by the continuous differentiability of f, there exists Δ such that $\frac{\partial f(q',v)}{\partial q'(v)} \leq \Delta$ whenever $0 \leq q'(v) \leq \kappa(v)$ for all $v \in V$. Let q be the optimal load vector corresponding to the assignments satisfying Conditions 8.1 and 8.2. Extend q to $\hat{V}$ by setting $q(v_L) = (\sum_u \lambda(u)) - \sum_{v \in V} q(v)$. The next two lemmas establish the convergence of x to the load vector q.

Lemma 8.9 *Given $\epsilon > 0$, there exists $\tau_0(\epsilon)$ such that for all $v \in \hat{V}$,*

$$x_t(v) \geq q(v) - \epsilon \tag{8.10}$$

whenever $t \geq \tau_0(\epsilon)$.

Proof. We first establish the inequality (8.10) for $v \in V$. Let $\{V_1, V_2, \cdots, V_J\}$ and $\{U_1, U_2, \cdots, U_J\}$ be the unique partitions of V and U, respectively, defined by Lemma 2.2 when condition (2.1) is replaced by Condition 8.1, and the vector q is replaced by $(f(q, v) : v \in V)$ in (2.2) and (2.3). Let $j \in \{1, 2, \cdots, J\}$ and define

$$\begin{aligned} m_t^j &= \inf_{v \in \bigcup_{i=j}^J V_i} f(x_t, v), \\ F_t^j &= \left\{ v \in \bigcup_{i=j}^J V_i \ : \ f(x_t, v) = m_t^j \right\}, \\ N^*(F_t^j) &= \{u : N(u) \cap F_t^j \neq \emptyset \text{ and } N(u) \cap (\bigcup_{i=1}^{j-1} V_i) = \emptyset\}, \end{aligned}$$

with the understanding that $\bigcup_{i=1}^0 V_i = \emptyset$. Let f_j denote the value such that $f(q, v) = f_j$ for all $v \in V_j$. Assume that t is a regular point of m^j such that $m_t^j < f_j - \epsilon\delta$. Then by the explanations indicated in parentheses,

$$
\begin{array}{rcll}
|F_t^j|\dot{m}_t^j & = & \sum_{v\in F_t^j} \dot{f}(x_t, v) & \text{(Lemma 8.8)} \\
& \geq & \delta \sum_{v\in F_t^j} \dot{x}_t(v) & \text{(Definition 8.1)} \\
& \geq & \delta(\sum_{u\in N^*(F_t^j)} \lambda(u) - \sum_{v\in F_t^j} x_t(v)) & \text{(Definition of } F_t^j \text{ and the fluid equations)} \\
& \geq & \delta(\sum_{u\in N^*(F_t^j)} \lambda(u) - \sum_{v\in F_t^j} (q(v) - \epsilon\delta/\Delta)) & \text{(Definition of } F_t^j \text{)} \\
& \geq & |F_t^j|\epsilon\delta^2/\Delta. & \text{(Definition of } N^* \text{ and } q \text{)}
\end{array}
$$

Therefore, if $t \geq \sup_v |f(0,v)|\Delta/\epsilon\delta^2$, then $m_t^j \geq f_j - \epsilon\delta$, so that $f(x_t, v) \geq f(q, v) - \epsilon\delta$ for $v \in V_j$, which in turn implies that $x_t(v) \geq q(v) - \epsilon$ for $v \in V_j$. Since j is arbitrary, (8.10) holds for all $t \geq \sup_v |f(0,v)|\Delta/\epsilon\delta^2$, and $v \in V$.

To complete the proof of the lemma, note that by Lemma 6.3, there exists a $\tau_J(\epsilon)$ such that $x_t(v_L) + \sum_{v\in V_J} x_t(v) \geq \sum_{u\in U_J} \lambda(u) - \epsilon$ for all $t \geq \tau_J(\epsilon)$. Therefore, for all such t, $x_t(v_L) \geq (\sum_{u\in U_J} \lambda(u) - \epsilon - \sum_{v\in V_J} \kappa(v))_+ \geq q(v_L) - \epsilon$. This proves (8.10) for $v = v_L$ and establishes the lemma with $\tau_0(\epsilon) = (\sup_v |f(0,v)|\Delta/\epsilon\delta^2) \vee \tau_J(\epsilon)$. □

Lemma 8.10 *(Insensitivity to Initial State) If (x, A) is a solution to the fluid equations (8.2)-(8.6), then* $\lim_{t\to\infty} \|x_t - q\|_{sup} = 0$ *uniformly for all x_0 in bounded subsets of $R_+^{\hat{V}}$.*

Proof. Fix $l > 0$ and let $\sum_{v\in\hat{V}} x_0(v) < l$. Given $\epsilon > 0$, set $\epsilon_0 = \epsilon/(|V| + 2)$. Appealing to Lemma 6.3, let $\tau_1(l, \epsilon_0)$ be such that for all $t \geq \tau_1(l, \epsilon_0)$, $|\sum_{v\in\hat{V}} x_t(v) - \sum_u \lambda(u)| < \epsilon_0$, or equivalently $|\sum_{v\in\hat{V}}(x_t(v) - q(v))| < \epsilon_0$. If $t > \tau_0(\epsilon_0) \vee \tau_1(l, \epsilon_0)$, then Lemma 8.9 implies that

$$\inf_{v\in\hat{V}} (x_t(v) - q(v)) \geq -\epsilon_0 \geq -\epsilon.$$

This in turn implies that

$$
\begin{aligned}
\sup_{v\in\hat{V}}(x_t(v) - q(v)) &\leq \sum_{v\in\hat{V}} (x_t(v) - q(v) + \epsilon_0) \\
&= \sum_{v\in\hat{V}} (x_t(v) - q(v)) + (|V| + 1)\epsilon_0 \\
&\leq (|V| + 2)\epsilon_0 = \epsilon,
\end{aligned}
$$

which yields the desired result. □

Lemma 8.11 *(Convergence of Equilibrium Distributions) Let q be the unique load vector corresponding to assignments satisfying Conditions 8.1 and 8.2, and ν be the distribution of the equilibrium load X_∞^γ. Then for all $\epsilon > 0$,*

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

Proof. The proof of Lemma 7.2 applies directly by using Lemmas 8.5 and 8.10 in place of Lemmas 5.3 and 6.4, respectively. □

Theorem 8.1 *(Asymptotic Optimality of LRLR) Let $P_\gamma^\pi(Loss)$ denote the loss probability of an arbitrary allocation policy π and $P_\gamma^{LRLR}(Loss)$ denote the loss probability of the LRLR policy for some normalization function f. Then*

$$\liminf_{\gamma\to\infty} P_\gamma^\pi(Loss) \geq \lim_{\gamma\to\infty} P_\gamma^{LRLR}(Loss) = 1 - \frac{\sum_{v\in V} q(v)}{\sum_u \lambda(u)},$$

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

Proof. The collection $(X_\infty^\gamma : \gamma > 0)$ is dominated by normalized Poisson random variables and is uniformly integrable. Therefore,

$$\begin{aligned}
\lim_{\gamma\to\infty} P_\gamma^{LRLR}(Loss) &= \lim_{\gamma\to\infty}\liminf_{t\to\infty} \frac{E[\text{number of consumers lost in } [0,t]]}{t\gamma\sum_u \lambda(u)}, \\
&= \frac{1}{\sum_u \lambda(u)} \lim_{\gamma\to\infty} E[X_\infty^\gamma(v_L)], \\
&= \frac{q(v_L)}{\sum_u \lambda(u)}, \\
&= 1 - \frac{\sum_{v\in V} q(v)}{\sum_u \lambda(u)},
\end{aligned}$$

where the second step follows by Little's Theorem, and the third step is a consequence of Lemma 8.11 and the uniform integrability of $(X_\infty^\gamma : \gamma > 0)$. The theorem now follows by Lemma 8.4. □

Having proven the optimality properties of generic LRLR policies, we now focus on two particular elements of this class, namely the least ratio routing and the maximum residual capacity routing policies. In the next two sections, we obtain a stronger version of Lemma 8.10 for these policies and provide explicit solutions of the fluid equations (8.2)-(8.6) for certain initial conditions.

8.1 Least ratio routing

In this section we focus on a particular least relative load routing policy, namely the *least ratio routing* (LRR). The LRR policy is defined as the LRLR policy associated with the normalization function $f(q,v) = -1 + q(v)/\kappa(v)$. Note that LRR assigns each consumer to the location with the least *load-to-capacity ratio*, $X_t^\gamma(v)/\kappa(v)$.

Let $\tilde{a}$ and $\tilde{q}$ be defined as in Lemma 8.2. Define the trajectories a and z as follows:

$$a_{u,v}(t) = \begin{cases} \tilde{a}_{u,v} & v \in V \quad t < \ln(\tilde{q}(v)/(\tilde{q}(v) - \kappa(v))_+), \\ \tilde{a}_{u,v}\kappa(v)/\tilde{q}(v) & v \in V \quad t \geq \ln(\tilde{q}(v)/(\tilde{q}(v) - \kappa(v))_+), \\ \sum_{v\in V}(\tilde{a}_{u,v} - a_{u,v}(t)) & v = v_L, \end{cases}$$

$$z_t(v) = \begin{cases} \tilde{q}(v)(1-e^{-t}) \wedge \kappa(v) & v \in V, \\ (1-e^{-t})\sum_u \lambda(u) - \sum_{v\in V} z_t(v) & v = v_L. \end{cases}$$

It is straightforward to verify that $(z, \int_0^\cdot a(s)ds)$ solves the fluid equations (8.2)-(8.6) with the zero initial state.

Lemma 8.12 *(Convergence Rate of LRR) Let q be the unique load vector corresponding to assignments satisfying Conditions 8.1 and 8.2. If (x, A) is a solution to the fluid equations (8.2)-(8.6), then*

$$\sup_{v\in V}|x_t(v) - q(v)| \leq e^{-t}\left(\sup_{v\in V} q(v) \vee \sum_{v\in V} x_0(v)\right).$$

Proof. By Lemma 8.7, for all $v \in V$, $x_t(v) \geq z_t(v)$. Also,

$$x_t(v) - z_t(v) \leq \sum_{v\in V}(x_t(v) - z_t(v)) = e^{-t}\sum_{v\in V} x_0(v).$$

These inequalities, together with the fact that

$$q(v)(1-e^{-t}) \leq z_t(v) \leq q(v)$$

for all $v \in V$, yield the desired result. □

8.2 Maximum residual capacity routing

The *maximum residual capacity routing* (MRCR) is defined as the LRLR policy associated with the normalization function $f(q,v) = q(v) - \kappa(v)$. Note that the MRCR policy assigns each consumer to the location with the maximum *residual capacity* defined as $\gamma\kappa(v) - X_t(v)$.

Let a be an optimal assignment satisfying Conditions 8.1 and 8.2, and let q be the load vector determined by a. Extend a to $(\hat{U}, \hat{V}, \hat{N})$ by setting $a_{u,v_L} = \lambda(u) - \sum_{v \in V} a_{u,v}$. Let z_0 denote the vector such that $z_0(v) = \kappa(v)$ for $v \in V$, and $0 \le z_0(v_L) < \infty$. Direct verification yields that (z, A), defined by

$$\begin{aligned} z_t(v) &= z_0(v)e^{-t} + q(v)(1 - e^{-t}) \\ A_{u,v}(t) &= a_{u,v}t, \end{aligned}$$

is a solution to the fluid equations (8.2)-(8.6) starting with the initial state z_0.

Lemma 8.13 *(Convergence Rate of MRCR) Let q be the unique load vector corresponding to assignments satisfying Conditions 8.1 and 8.2. If (x, A) is a solution to the fluid equations (8.2)-(8.6) with an arbitrary initial state x_0, then*

$$\sup_{v \in \hat{V}} |x_t(v) - q(v)| \le e^{-t}\left((q(v_L) \vee x_0(v_L)) + \sum_{v \in V} \kappa(v)\right).$$

Proof. Let z_0 be an initial state vector defined as $z_0(v) = \kappa(v)$ for $v \in V$ and $z_0(v_L) = x_0(v_L)$, and let z denote the load trajectory starting with z_0. By Lemma 8.7, for all $v \in \hat{V}$,

$$\begin{aligned} x_t(v) &\le z_t(v) \\ &= z_0(v)e^{-t} + q(v)(1 - e^{-t}). \end{aligned}$$

On the other hand, for any $v \in \hat{V}$,

$$\begin{aligned} z_t(v) - x_t(v) &\le \sum_{v \in \hat{V}} (z_t(v) - x_t(v)) \\ &= e^{-t} \sum_{v \in \hat{V}} (z_0(v) - x_0(v)) \\ &\le e^{-t} \sum_{v \in V} \kappa(v). \end{aligned}$$

Therefore,

$$(z_0(v) - q(v) - \sum_{v \in V} \kappa(v))e^{-t} \leq x_t(v) - q(v) \leq (z_0(v) - q(v))e^{-t},$$

and the desired result follows. □

9 The Migration Model

In this section we consider locations with infinite capacities and generalize the basic model of Section 3 by allowing consumers to change their types while they are in the system and also by including type-dependent departure rates. Towards this end, Lemma 9.1 identifies the weak limits of the network process as solutions to certain fluid equations. An example shows that the fluid equations do not necessarily uniquely determine the transient behavior of the load; nevertheless, by Lemma 9.5, the limit point is unique. These facts are used to establish Theorem 9.1 on the asymptotic optimality of LLR.

The analytical description of the migration model involves a *routing matrix* R, such that $R = [r_{u,u'}]_{U\times U}$, where $r_{u,u'} \geq 0$ for $u \neq u'$, and $\sum_{u'\in U} r_{u,u'} \leq 0$ for all $u \in U$. Given a load sharing network (U, V, N), an arrival rate vector λ, 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 consumer transforms into a type u' consumer with rate $r_{u,u'}$ or departs from the system with rate $-\sum_{u'\in U} r_{u,u'}$. This implies that $-r_{u,u}$ is the rate that a type u consumer changes by either transforming to another type or leaving the system. Each arrival is assigned to a location via the LLR policy. In addition, when a consumer changes its type, it is reassigned using LLR. Its location may or may not change. We assume that R is nonsingular, so that every consumer eventually departs from the system. Let $L_t(u)$ continue to denote the number of type u consumers in the network at time t. It can be verified by direct substitution that the equilibrium vector $(L_\infty(u) : u \in U)$ is a vector of independent Poisson random variables with mean vector $\gamma\rho$, where $\rho = -\lambda R^{-1}$, so that the normalized cost of any allocation policy is lower bounded by $\Psi(\rho)$. This section extends the analysis of Section 3 to the more general setting.

Let the *contribution* of type u to location v at time t, denoted by $C_t(u, v)$, be the number of type u consumers at location v at time t. Define $C_t^\gamma(u,v) = \gamma^{-1}C_t(u,v)$, and $L_t^\gamma(u) = \gamma^{-1}L_t(u)$. Note that $X_t^\gamma(v) = \sum_u C_t^\gamma(u,v)$, and $L_t^\gamma(u) = \sum_v C_t^\gamma(u,v)$. Under LLR, C is Markov on the state space $Z_+^{U\times V}$, and $C^\gamma(u,v)$ can be represented as

$$C_t^\gamma(u,v) = C_0^\gamma(u,v) + M_t^\gamma(u,v) + A_{u,v}^\gamma(t) + \int_0^t r_{u,u}C_s^\gamma(u,v)ds,$$

where

$$A_{u,v}^\gamma(t) = \int_0^t (\lambda(u) + \sum_{u'\neq u} L_s^\gamma(u')r_{u',u})\frac{I\{X_s(v) = m_u(X_s)\}}{\sum_{v'\in N(u)} I\{X_s(v') = m_u(X_s)\}}ds,$$

and $M^\gamma(u,v)$ is a local martingale with $M_0^\gamma(u,v) = 0$. Given that (C_0^γ) is tight, the methods of Section 5 can be applied directly to establish the tightness of (C^γ, A^γ) and characterize the limits of weakly convergent subsequences. Namely, the following lemma holds:

Lemma 9.1 *Suppose (C_0^γ) is tight. Then every subsequence $(C^{\gamma_n}, A^{\gamma_n})$ has a further subsequence $(C^{\gamma_{n_k}}, A^{\gamma_{n_k}})$ such that $(C^{\gamma_{n_k}}, A^{\gamma_{n_k}}) \Longrightarrow (c, A)$, where (c, A) satisfies the following fluid equations:*

$$c_t(u,v) = c_0(u,v) + A_{u,v}(t) + \int_0^t r_{u,u}c_s(u,v)ds \tag{9.1}$$

$$A_{u,v}(0) = 0, \quad A_{u,v}(t) \text{ nondecreasing}, \quad \sum_{v\in N(u)} A_{u,v}(t) = \lambda(u)t + \sum_{u'\neq u}\int_0^t l_s(u')r_{u',u}ds, \tag{9.2}$$

$$\int_0^t I\{x_s(v) > m_u(x_s)\}dA_{u,v}(s) = 0, \tag{9.3}$$

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

The following example shows, in contrast to Lemma 6.2, that the initial state c_0 and the fluid equations (9.1)-(9.3) do not necessarily determine a unique load trajectory x.

Example. (Type dependent departure rates, no routing) Consider the load sharing network and the routing matrix of Figure 3. Suppose $c_0(3,3) = .9495\times 10^7$, and $c_0(u,v) = 0$ otherwise. Let $\tau = \ln 10$, and consider the two assignment regimes (c, A) and $(\tilde{c}, \tilde{A})$ for $t \in [0,\tau]$ as listed in Table 1. It is straightforward to verify that both (c, A) and $(\tilde{c}, \tilde{A})$ satisfy (9.1)-(9.3), and $x_t(v) = \tilde{x}_t(v)$ for $v \in V$ and $t \in [0,\tau]$. Note that $\tau = \inf\{t : x_t(1) = x_t(2) = x_t(3)\}$, and $x_\tau(1) = .9495$. Under both regimes, at time τ, the instantaneous rate of decrease of load at location 3 is 7(.9495), whereas the

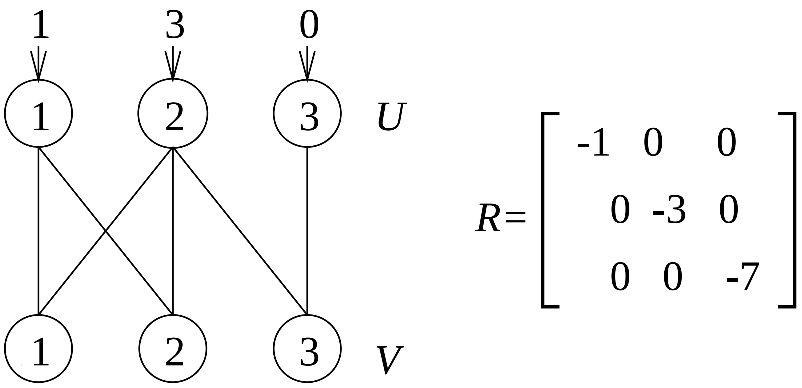

Figure 3: An example to illustrate the nonuniqueness of the solutions to the fluid equations (9.1)-(9.3).

Table 1: Two assignment regimes for the network of Figure 3.

	$A_t(u,v)$			$c_t(u,v)$		
$u \setminus v$	1	2	3	1	2	3
1	$t/2$	$t/2$	0	$(1-e^{-t})/2$	$(1-e^{-t})/2$	0
2	$3t/2$	$3t/2$	0	$(1-e^{-3t})/2$	$(1-e^{-3t})/2$	0
3	0	0	0	0	0	$(.9495 \times 10^7)e^{-7t}$

	$\tilde{A}_t(u,v)$			$\tilde{c}_t(u,v)$		
$u \setminus v$	1	2	3	1	2	3
1	t	0	0	$1-e^{-t}$	0	0
2	$1-e^{-t}$	$3t-1+e^{-t}$	0	$(e^{-t}-e^{-3t})/2$	$1-(e^{-t}+e^{-3t})/2$	0
3	0	0	0	0	0	$(.9495 \times 10^7)e^{-7t}$

instantaneous rates of decrease of load at locations 1 and 2, respectively, are each upper bounded by 3(.9495). The difference is larger than the flow rate of type 2 arrivals; therefore, there exists a $\delta > 0$ such that (9.1)-(9.3) are not violated only if all type 2 arrivals are directed to location 3 during $[\tau, \tau+\delta]$. Under (c, A), type 1 arrivals can split evenly between locations 1 and 2 and maintain $x_t(1) = x_t(2)$ for $t \in [\tau, \tau+\delta]$. However, under $(\tilde{c}, \tilde{A})$, at time τ location 1 discharges at an instantaneous rate of $.9 + 3(.0495) = 1.0485$, and location 2 discharges at an instantaneous rate of $3(.9495) = 2.8485$. The difference between the discharge rates is greater than the flow rate of type 1 arrivals; therefore, there exists a $\delta' > 0$ such that all type 1 arrivals are directed to location 2, and $\tilde{x}_t(1) > \tilde{x}_t(2)$ for $t \in (\tau, \tau+\delta']$. Thus, c_0 together with the fluid equations (9.1)-(9.3) does not determine x uniquely. □

We now concentrate on the properties of the load trajectories corresponding to the solutions of the fluid equations. By the definition of the demand vector, the components of ρ are strictly positive; however, the extension of the results to nonnegative ρ is trivial.

Lemma 9.2 *Given $\bar{l}_0 \in R_+^U$, $\lim_{t\to\infty} \|l_t - \rho\|_{sup} = 0$ uniformly for $0 \le l_0 \le \bar{l}_0$.*

Proof. By Equations (9.1)-(9.3), l_t satisfies

$$l_t = l_0 + \lambda t + \int_0^t R l_s ds,$$

which can be solved to yield

$$l_t = l_0 e^{Rt} + \rho(I - e^{Rt}),$$

where the exponential e^{Rt} of the matrix R can be defined by a power series. Since $e^{Rt} \to 0$ exponentially, l_t converges to ρ exponentially fast, uniformly for $l_0 \le \bar{l}_0$. This establishes the lemma. □

The following auxiliary lemma is proved in the Appendix.

Lemma 9.3 *Suppose that $a_i \le \bar{a}_i$ for $1 \le i \le J$, and $w_{min} = \min_{1\le i\le J} w_i$. Then*

$$\sum_i a_i w_i \le \sum_i \bar{a}_i w_i + (\sum_i a_i - \sum_i \bar{a}_i) w_{min}.$$

Lemma 9.4 *Let q denote the unique load vector corresponding to the solutions of $SLB(\rho, \Phi)$. Given $\bar{l}_0 \in R_+^U$ and $\epsilon > 0$, there exists $t_1(\bar{l}_0, \epsilon)$ such that for all $v \in V$,*

$$x_t(v) \geq q(v) - \epsilon$$

whenever $t \geq t_1(\bar{l}_0, \epsilon)$, and $0 \leq l_0 \leq \bar{l}_0$.

Proof. Let $\{V_1, V_2, \cdots, V_J\}$ and $\{U_1, U_2, \cdots, U_J\}$ be the unique partitions of V and U, respectively, defined by Lemma 2.2 adapted to $SLB(\rho, \Phi)$. It is enough to show that

$$\inf_{v \in \bigcup_{i=j}^J V_i} x_t(v) \geq q_j - \epsilon \tag{9.4}$$

for all $j \in \{1, 2, \cdots, J\}$, and $t \geq t_1(\bar{l}_0, \epsilon)$, where q_j is the value such that $q(v) = q_j$ for all $v \in V_j$. Towards this end, for each $j \in \{1, 2, \cdots, J\}$ define

$$\begin{aligned} m_t^j &= \inf_{v \in \bigcup_{i=j}^J V_i} x_t(v), \\ F_t^j &= \left\{ v \in \bigcup_{i=j}^J V_i \ : \ x_t(v) = m_t^j \right\}, \\ c_t(u, F_t^j) &= \sum_{v \in F_t^j} c_t(u, v), \\ N^*(F_t^j) &= \{u : N(u) \cap F_t^j \neq \emptyset \text{ and } N(u) \cap (\bigcup_{i=1}^{j-1} V_i) = \emptyset\}, \end{aligned}$$

with the understanding that $\bigcup_{i=1}^0 V_i = \emptyset$.

Let $r_{min} = \min_u\{-r_{u,u}\}$ and $r_{max} = \max_u\{-r_{u,u}\}$. Given $\epsilon > 0$, let $\epsilon_0 < \epsilon r_{min}(2 \sum_u \sum_{u'} |r_{u,u'}|)^{-1}$. Appeal to Lemma 9.2 to fix $t_0(\bar{l}_0, \epsilon_0)$ such that $\sup_u |l_t(u) - \rho(u)| < \epsilon_0$ for all $t \geq t_0(\bar{l}_0, \epsilon_0)$ whenever $0 \leq l_0 \leq \bar{l}_0$. To prove the lemma by induction on j, let $j = 1$, and choose $t > t_0(\bar{l}_0, \epsilon_0)$. Suppose that t is a regular point of m^1 and x, and that

$$m_t^1 < q_1 - \epsilon. \tag{9.5}$$

Then, by Lemma 8.8 and the fact that $N^*(F_t^1) = N^{-1}(F_t^1)$,

$$|F_t^1|\dot{m}_t^1 = \sum_{v \in F_t^1} \dot{x}_t(v)$$

$$
\begin{aligned}
&= \sum_{u\in N^*(F_t^1)} \lambda(u) + \sum_{u\in U} l_t(u) \sum_{u'\in N^*(F_t^1)\setminus\{u\}} r_{u,u'} - \sum_{u\in U} c_t(u,F_t^1)(-r_{u,u}) \\
&= \sum_{u\in N^*(F_t^1)} \lambda(u) + \sum_{u\in U} l_t(u) \sum_{u'\in N^*(F_t^1)\setminus\{u\}} r_{u,u'} - \sum_{u\in N^*(F_t^1)} c_t(u,F_t^1)(-r_{u,u}). \qquad (9.6)
\end{aligned}
$$

By the choice of t, $c_t(u,F_t^1) \le \rho(u)+\epsilon_0$, so that Lemma 9.3 and (9.5) can be used to bound the third term on the right-hand side of (9.6) to obtain

$$
\begin{aligned}
|F_t^1|\dot{m}_t^1 \ \ge\ & \sum_{u\in N^*(F_t^1)} \lambda(u) + \sum_{u\in U} l_t(u) \sum_{u'\in N^*(F_t^1)\setminus\{u\}} r_{u,u'} - \sum_{u\in N^*(F_t^1)} (\rho(u)+\epsilon_0)(-r_{u,u}) \\
& - \left(|F_t^1|(q_1-\epsilon) - \sum_{u\in N^*(F_t^1)} (\rho(u)+\epsilon_0) \right) r_{min}.
\end{aligned}
$$

Note that $\sum_{u\in N^*(F_t^1)} \rho(u) \ge |F_t^1|q_1$, and $l_t(u) \ge \rho(u)-\epsilon_0$. This, together with the identity $\rho(u)(-r_{u,u}) = \lambda(u) + \sum_{u'\neq u} \rho(u')r_{u',u}$ and the choice of ϵ_0, yields that

$$
\begin{aligned}
|F_t^1|\dot{m}_t^1 &\ge -\epsilon_0 \left(\sum_{u\in U} \sum_{u'\in N^*(F_t^1)\setminus\{u\}} r_{u,u'} + \sum_{u\in N^*(F_t^1)} (-r_{u,u}) \right) + \epsilon |F_t^1| r_{min} \\
&\ge \frac{\epsilon r_{min} |F_t^1|}{2}.
\end{aligned}
$$

Thus, for almost all regular points t of m such that $t > t_0(\bar{l}_0,\epsilon_0)$, $\dot{m}_t^1 \ge \epsilon r_{min}/2$ whenever $m_t^1 < q_1-\epsilon$. Therefore, (9.4) holds for $j=1$ for all $t \ge T^1(\bar{l}_0,\epsilon) = t_0(\bar{l}_0,\epsilon_0) + 2\sum_u (\rho(u)+\epsilon_0)/\epsilon r_{min}$.

As the induction hypothesis, fix $j \in \{2,3,\cdots,J\}$, and suppose that given $\epsilon_0 > 0$, for each $i \in \{1,\cdots,j-1\}$ there exists a $T^i(\bar{l}_0,\epsilon_0)$ such that

$$
m_t^i \ge q_i - \epsilon_0
$$

whenever $t \ge T^i(\bar{l}_0,\epsilon_0)$.

Let $\epsilon_0 < \min\{\epsilon r_{min}(4(|U|+|V|)r_{max})^{-1}, \epsilon r_{min}(4\sum_u \sum_{u'} |r_{u,u'}|)^{-1}\}$, and choose $t > \max\{t_0(\bar{l}_0,\epsilon_0), T^1(\bar{l}_0,\epsilon_0),\cdots,T^{j-1}(\bar{l}_0,\epsilon_0)\}$. Suppose that t is a regular point of m^j and x, and $m_t^j < q_j-\epsilon$. Note that the methods used in the case $j=1$ imply that

$$
\begin{aligned}
|F_t^j|\dot{m}_t^j &= \sum_{v\in F_t^j} \dot{x}_t(v) \\
&\ge \sum_{u\in N^*(F_t^j)} \lambda(u) + \sum_{u} l_t(u) \sum_{u'\in N^*(F_t^j)\setminus\{u\}} r_{u,u'}
\end{aligned}
$$

$$
\begin{aligned}
- & \sum_{u\in N^*(F_t^j)} c_t(u,F_t^j)(-r_{u,u}) - \sum_{u\in N^*(F_t^j)^c} c_t(u,F_t^j)(-r_{u,u}) \\
\geq & -\epsilon_0 \left(\sum_u \sum_{u'\in N^*(F_t^1)\setminus\{u\}} r_{u,u'} + \sum_{u\in N^*(F_t^1)} (-r_{u,u}) \right) + \epsilon |F_t^1| r_{min} - r_{max} \sum_{u\in N^*(F_t^j)^c} c_t(u,F_t^j).
\end{aligned}
$$

By (2.5), the choice of t, and the induction hypothesis,

$$
\begin{aligned}
\sum_{u\in N^*(F_t^j)^c} c_t(u,F_t^j) & < \sum_{u\in\bigcup_{i=1}^{j-1} U_i} (\rho(u)+\epsilon_0) - \sum_{v\in\bigcup_{i=1}^{j-1} V_i} (q_v - \epsilon_0) \\
& \leq \frac{\epsilon r_{min}}{4 r_{max}}.
\end{aligned}
$$

Hence by the choice of ϵ_0, $|F_t^j|\dot{m}_t^j \geq \epsilon r_{min}|F_t^j|/2$. Thus, there exists a $T^j(\bar{l}_0,\epsilon)$ such that $m_t^j \geq q_j - \epsilon$ for all $t \geq T^j(\bar{l}_0,\epsilon)$. This completes the induction step and the proof of the lemma with $t_1(\bar{l}_0,\epsilon) = T^J(\bar{l}_0,\epsilon)$. □

Lemma 9.5 *(Global Convergence) Let x be the load function corresponding to an arbitrary solution of the fluid equations (9.1)-(9.3), and q be the unique load vector corresponding to the solutions of $SLB(\rho,\Phi)$. Then* $\lim_{t\to\infty} \|x_t - q\|_{sup} = 0$ *uniformly for all l_0 in bounded subsets of R_+^U.*

Proof. Fix $\bar{l}_0 \in R_+^U$, and let $l_0 \leq \bar{l}_0$. Given $\epsilon > 0$, set $\epsilon_0 = \epsilon/(|V|+1)$, and appealing to Lemma 9.2, let $t_0(\bar{l}_0,\epsilon_0)$ be such that $|\sum_{v\in V}(x_t(v) - q(v))| < \epsilon_0$ for all $t \geq t_0(\bar{l}_0,\epsilon_0)$. If $t > t_0(\bar{l}_0,\epsilon_0) \vee t_1(\bar{l}_0,\epsilon_0)$, then Lemma 9.4 implies that

$$
\inf_{v\in V}(x_t(v) - q(v)) \geq -\epsilon_0 \geq -\epsilon,
$$

which in turn implies

$$
\begin{aligned}
\sup_{v\in V}(x_t(v) - q(v)) & \leq \sum_{v\in V}(x_t(v) - q(v) + \epsilon_0) \\
& = \sum_{v\in V}(x_t(v) - q(v)) + |V|\epsilon_0 \\
& \leq (|V|+1)\epsilon_0 = \epsilon.
\end{aligned}
$$

This proves the desired result. □

Let P_μ denote the distribution of the process C^γ when C_0^γ has distribution μ and μ_0 denote the deterministic distribution concentrated at the zero state. The proof of the following lemma is immediate:

Lemma 9.6 *Given $\epsilon > 0$, there exists a γ_ϵ such that whenever $\gamma > \gamma_\epsilon$,*

$$P_{\mu_0}\left(\sum_v X_t^\gamma(v) \le \epsilon + \sum_u \rho(u)\right) \ge 1 - \epsilon \quad \textit{for any } t \ge 0.$$

Lemma 9.7 *(Convergence of Equilibrium Distributions) Let q be the unique load vector corresponding to solutions of $SLB(\rho, \Phi)$ and ν be the distribution of the equilibrium load X_∞^γ. Then for all $\epsilon > 0$*

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

Proof. The proof of Lemma 7.2 applies directly by redefining μ_t^γ as the distribution of C_t^γ and by using Lemmas 9.1, 9.5, and 9.6 in place of Lemmas 5.3, 6.4, and 7.1, respectively. □

Theorem 9.1 *(Asymptotic Optimality of LLR) Given an allocation policy π, let J_γ^π denote the cost under π when the network demand is $\gamma\lambda$. Then*

$$\liminf_{\gamma\to\infty} \gamma^{-p} J_\gamma^\pi \ge \lim_{\gamma\to\infty} \gamma^{-p} J_\gamma^{LLR} = \Psi(\rho) > 0.$$

Proof. The proof of Lemma 7.1 applies directly by using Lemma 9.7 in place of Lemma 7.2 and ρ in place of λ. □

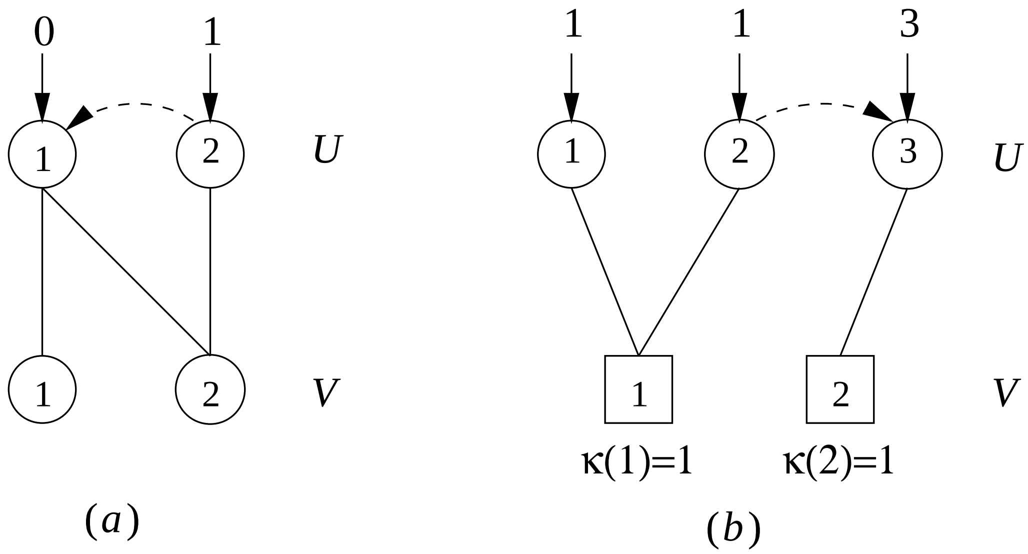

Figure 4: The counter examples for the optimality of (*a*) sticky LLR, (*b*) LRLR under finite capacities and migration.

10 Conclusions and Discussion

This paper concentrates on the dynamic load balancing problem and studies the performance of practical allocation policies, namely LLR and the class LRLR. When there are no capacity constraints on the resources, LLR is shown to achieve asymptotically the most balanced load in the sense of minimizing a wide class of long-term average costs. LLR is also robust to migration, provided that consumers are reassigned according to LLR whenever their types change. On the other hand, when the resources have finite capacities, LRLR policies asymptotically achieve the minimum possible loss probability. The desirable aspects of the considered policies are low computational complexity, decentralized implementation, and robustness to arrival and migration rates.

The reassignment of migrating consumers is important for the asymptotic optimality of LLR in the migration model. The network of Figure 4(a) is an example in which LLR is not optimal without reassignments. Let $\lambda = (0, 1)$ and the routing matrix be

$$R = \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}.$$

Hence all consumers arrive as type 2 and migrate to become type 1 before leaving the system. Suppose that consumers are assigned using LLR upon arrival; however, they maintain their original locations even though they migrate. Then at any time t, all of the load in the network is at location 2; hence the limiting normalized cost of this policy is 4. A simple calculation yields that the LLR policy splits the load equally between the two locations, thus having a limiting normalized cost of 2.

Optimality properties of the policies discussed in this paper do not necessarily persist in the case of finite capacities *and* migration. In particular, myopic policies, which accept a consumer whenever possible, may not be asymptotically optimal. As an example to illustrate this, consider the network of Figure 4(b) in heavy traffic. Let $\lambda = (1, 1, 3)$, $\kappa = (1, 1)$, and the routing matrix be

$$R = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}.$$

Hence type 2 arrivals first visit location 1 and then location 2 before exiting the system. Without loss of generality, assume that location 2 gives higher priority to exogenous arrivals, in the sense that an exogenous arrival blocks a migrated consumer that is already in location 2, provided that location 2 is full at the time of arrival. Since exogenous arrivals suffice to overload location 2, all type 2 arrivals are bound to be lost. Any myopic policy blocks half of type 1 arrivals and has a limiting consumer loss probability of 0.7. On the other hand, a policy that blocks type 2 arrivals regardless of the system state has a limiting consumer loss probability of 0.6. We therefore conclude that the optimal policies have considerably more complex structures under the more general case.

11 Appendix

This section contains the deferred proofs from previous sections. We start with the proof of Lemma 2.1 by first giving an auxiliary result.

Lemma 11.1 *Let $\Phi : R^d \to R$ be strictly convex and differentiable, and Φ_v denote the v^{th} partial derivative of Φ. If Φ is symmetric in its arguments (i.e., $\Phi(x(1), \cdots, x(d)) = \Phi(x(p(1)), \cdots, x(p(d)))$*

for any permutation p), then for all v, v'

$$x(v) > x(v') \Longrightarrow \Phi_v(x) > \Phi_{v'}(x).$$

Proof. Since the conclusion involves only two arguments, we can assume without loss of generality that $d = 2$. For $(a,b) \in R^2$, define $g_{a,b}(\alpha) = \Phi(\alpha(a,b) + (1-\alpha)(b,a))$. Then, $\dot{g}_{a,b}(\alpha) = (\Phi_1 - \Phi_2)(a-b)$, where the partial derivatives Φ_1 and Φ_2 are evaluated at $\alpha(a,b)+(1-\alpha)(b,a)$. Note that by the strict convexity of Φ, $g_{a,b}$ is strictly convex for $a \neq b$. Also by the symmetry of Φ, $\dot{g}_{a,b}(\alpha)|_{\alpha=\frac{1}{2}} = 0$; therefore, for $a \neq b$, $\dot{g}_{a,b}(\alpha)|_{\alpha=1} > 0$. This implies $(\Phi_1(a,b) - \Phi_2(a,b))(a-b) > 0$, which proves the claim. □

Proof of Lemma 2.1. The problem $SLB(\lambda, \Phi)$ is a convex optimization problem on a compact and convex set; thus, there exists a solution.

To establish the second statement of the lemma, we argue by contradiction in each direction. In what follows, $\psi(a)$ denotes the value $\Phi(q)$ induced by the assignment a. First, let a satisfy (2.1), and suppose that a is not a solution to $SLB(\lambda, \Phi)$. Then there exists an admissible perturbation vector h such that

$$\sum_{v \in N(u)} h_{u,v} = 0 \quad \text{for all } u \in U \tag{11.1}$$

$$h_{u,v} \geq 0 \ \text{ whenever } a_{u,v} = 0, \quad h_{u,v} = 0 \ \text{ whenever } v \in N(u)^c \tag{11.2}$$

$$\lim_{\epsilon \searrow 0} \frac{\psi(a+\epsilon h) - \psi(a)}{\epsilon} \ = \sum_u \sum_{v \in N(u)} h_{u,v} \Phi_v(q) \ < \ 0. \tag{11.3}$$

By (11.2) we have that for all u and $v \in N(u)$

$$\begin{aligned} h_{u,v} < 0 \quad &\Longrightarrow \quad a_{u,v} > 0 \\ &\Longrightarrow \quad q(v) \leq q(v') \quad \text{for all } v' \in N(u) \\ &\Longrightarrow \quad \Phi_v(q) \leq \Phi_{v'}(q) \quad \text{for all } v' \in N(u) \end{aligned} \tag{11.4}$$

by using Lemma 11.1 in the last step. Now for $u \in U$ define

$$\Phi_u^-(q) = \max(\Phi_v(q) : h_{u,v} < 0) \quad , \quad h_u^- = \textstyle\sum_{v: h_{u,v} < 0} h_{u,v},$$
$$\Phi_u^+(q) = \min(\Phi_v(q) : h_{u,v} > 0) \quad , \quad h_u^+ = \textstyle\sum_{v: h_{u,v} > 0} h_{u,v}.$$

Then by (11.1) $h_u^+ = -h_u^-$, and by (11.4) $\Phi_u^+(q) \geq \Phi_u^-(q)$, and therefore for all $u \in U$,

$$\sum_u \sum_{v \in N(u)} h_{u,v} \Phi_v(q) \geq \sum_u h_u^+ (\Phi_u^+(q) - \Phi_u^-(q)) \geq 0,$$

which contradicts (11.3).

To show that the converse also holds, suppose that a does not satisfy the condition (2.1). In particular, let u be such that for some $v, v' \in N(u)$,

$$a_{u,v} > 0 \quad \text{and} \quad q(v) > q(v').$$

Then by Lemma 11.1, $\Phi_v(q) - \Phi_{v'}(q) > 0$; thus, there exists a δ small enough such that it is possible to decrease $a_{u,v}$ and increase $a_{u,v'}$ by an amount δ without violating the constraints of $SLB(\lambda, \Phi)$ and obtain a smaller value for Φ. Therefore, a cannot be a solution.

Finally, by the strict convexity of Φ, there is a unique load vector corresponding to the solutions of $SLB(\lambda, \Phi)$. □

Proof of Lemma 2.2. It is straightforward to form the partition $\{V_1, V_2, \cdots, V_J\}$ that satisfies (2.2) and (2.3). This partition is unique since q is the same for all assignments a satisfying (2.1). Let a be an arbitrary assignment satisfying (2.1), and define the set of subsets $\{U_1, U_2, \cdots, U_J\}$ of U as

$$U_i = \{u \in U : a(u, v) > 0 \text{ for some } v \text{ in } V_i\}.$$

By (2.1), (2.4) and (2.5) hold; therefore, $\{U_1, U_2, \cdots, U_J\}$ is a partition of U.

It remains to show that *any* assignment satisfying (2.1) yields the same $\{U_1, U_2, \cdots, U_J\}$. Suppose $\tilde{a}$ is another such assignment which yields $\{\tilde{U}_1, \tilde{U}_2, \cdots, \tilde{U}_J\}$. To prove the claim by contradiction, assume that there exists a $u \in U$ such that $u \in U_j \cap \tilde{U}_i$, where $i < j$. Since $u \in U_j$, by (2.5) $N(u) \cap V_i = \phi$. Therefore, $\tilde{a}(u, v) = 0$ for all $v \in V_i$. This contradicts the assumption that $u \in \tilde{U}_i$ and establishes the uniqueness of $\{U_1, U_2, \cdots, U_J\}$. □

Proof of Lemma 2.3. For $i = 1, 2$, let $\lambda_i \in R_+^U$, and a_i solve $SLB(\lambda_i, \Phi)$ with the corresponding load vector q_i. Then for any $\alpha \in [0, 1]$,

$$\alpha\Psi(\lambda_1) + (1-\alpha)\Psi(\lambda_2) \quad = \quad \alpha\Phi(q_1) + (1-\alpha)\Phi(q_2)$$

$$
\begin{aligned}
&\geq \Phi(\alpha q_1 + (1-\alpha) q_2) \\
&\geq \Psi(\alpha\lambda_1 + (1-\alpha)\lambda_2).
\end{aligned}
$$

The second step follows by the convexity of Φ. The third step follows by the definition of Ψ and the fact that $\alpha a_1 + (1-\alpha)a_2$ is an admissible assignment that satisfies $\alpha\lambda_1 + (1-\alpha)\lambda_2$ with the load vector $\alpha q_1 + (1-\alpha)q_2$. □

Proof of Lemma 8.1. The proof of Lemma 2.1 applies by replacing $q(v)$ with $f(q,v)$ and by noting that (11.4) follows by the definition of Φ. □

Proof of Lemma 8.2. It is straightforward to see that $a \in \mathcal{B}_{\lambda,\kappa}$ and that q is the load vector corresponding to a. To show that a satisfies Condition 8.1, suppose that $f(q,v) > m_u(f(q))$ for some u, v with $v \in N(u)$. Then by Lemma 8.1 there exists a $v' \in N(u)$ such that $f(q,v') < 0$. Since $q(v') = \tilde{q}(v')$ for all v' with $f(\tilde{q},v') < 0$, $m_u(f(q)) = m_u(f(\tilde{q}))$, and therefore,

$$
f(\tilde{q},v) \geq f(q,v) > m_u(f(q)) = m_u(f(\tilde{q})). \tag{11.5}
$$

Since $\tilde{a}_{u,v}$ satisfies Condition 8.1, (11.5) implies that $\tilde{a}_{u,v} = 0$; thus, $a_{u,v} = 0$, and Condition 8.1 is satisfied.

To show that Condition 8.2 is satisfied, note that if $\sum_v a_{u,v} < \lambda(u)$, then there exists a $v \in N(u)$ such that $\tilde{a}_{u,v} > 0$ and $\tilde{q}(v) > \kappa(v)$. This implies that $f(\tilde{q},v') > 0$ for all $v' \in N(u)$ and hence that $f(q,v') = 0$ for all $v' \in N(u)$, establishing Condition 8.2. □

Proof of Lemma 8.3. Let f be a normalization function and a be an assignment satisfying Conditions 8.1 and 8.2 with f and the corresponding load vector q. To prove the optimality of a, argue by contradiction. If a is not a solution to $SLP(\lambda,\kappa)$, then there exists a perturbation vector h such that

$$
\begin{aligned}
h_{u,v} &= 0 \quad v \in N(u)^c, \\
h_{u,v} &\geq 0 \quad \text{if} \;\; a_{u,v} = 0, \qquad (11.6)\\
\sum_u h_{u,v} &\leq \kappa(v) - q(v), \qquad (11.7)
\end{aligned}
$$

$$\sum_v h_{u,v} \leq \lambda(u) - \sum_v a_{u,v}, \tag{11.8}$$

$$\sum_u \sum_v h_{u,v} > 0. \tag{11.9}$$

We use induction to arrive at the desired contradiction. Let

$$\begin{aligned} U_0 &= \{u : \sum_v h_{u,v} > 0\}, \\ U_{j+1} &= U_j \cup \{u \notin U_j : h_{u,v} < 0 \text{ for some } v \in N(U_j)\}, \ \ j \geq 0. \end{aligned}$$

By inequality (11.9), U_0 is nonempty. Inequality (11.8), Condition 8.2, and the Definition 8.1 of the normalization function imply that

$$q(v) = \kappa(v) \text{ for all } v \in N(U_0). \tag{11.10}$$

By (11.7), $\sum_{v \in N(U_0)} \sum_u h_{u,v} \leq 0$, and by the definition of U_0, $\sum_{u \in U_0} \sum_{v \in N(U_0)} h_{u,v} > 0$; therefore, $U_1 \neq U_0$. If $u \in U_1 \setminus U_0$, then inequality (11.6) implies that $a_{u,v} > 0$ for some $v \in N(U_0)$. By Condition 8.1, $f(q,v) = 0$ for all $v \in N(u)$. This, along with (11.10), implies that

$$q(v) = \kappa(v) \text{ for all } v \in N(U_1).$$

As the induction hypothesis, assume that $q(v) = \kappa(v)$ for all $v \in N(U_k)$. By the definition of $(U_j : j \geq 0)$, $\sum_{u \in U_{k+1}} \sum_{v \in N(U_k)} h_{u,v} > 0$; therefore, the argument of the base case yields that

$$U_{k+1} \neq U_k \quad \text{and} \quad q(v) = \kappa(v) \quad \text{for all } v \in N(U_{k+1}).$$

This contradicts the finiteness of the network, and hence the existence of h, proving the optimality of a.

Lemma 8.2 establishes the existence of an assignment $a \in \mathcal{B}_{\lambda,\kappa}$ that satisfies Conditions 8.1 and 8.2. To prove the uniqueness of the load vector by contradiction, let a and $\tilde{a}$ be two such assignments with the corresponding load vectors q and $\tilde{q}$ such that $q \neq \tilde{q}$. Let

$$F = \{v : q(v) > \tilde{q}(v)\}.$$

If $v \in F$, then for any u,

$$\begin{aligned} a_{u,v} > 0 &\Rightarrow f(q,v) \leq f(q,v') \quad \text{for all } v' \in F^c \cap N(u) \\ &\Rightarrow f(\tilde{q},v) < f(\tilde{q},v') \quad \text{for all } v' \in F^c \cap N(u) \\ &\Rightarrow \sum_{v \in F} \tilde{a}_{u,v} = \lambda(u), \end{aligned}$$

where the second step follows by the strict monotonicity of f, and the third step follows by Condition 8.1. However, this implies that $\sum_{v\in F}(\tilde{a}_{u,v} - a_{u,v}) \geq 0$ for any $u \in U$, which contradicts the definition of F and proves the desired result. $\square$

Proof of Lemma 8.6. The set N_α can be taken to be the union of $\{t : \dot{g}_t \text{ does not exist}\}$ and $\{t : \dot{g}_t \text{ exists}, \dot{g}_t \neq 0, g_t = \alpha\}$. The first of these sets has Lebesgue measure zero (Royden (1988), Corollary 5.12). All the points in the second set are isolated, hence it is finite or countably infinite, and therefore it also has measure zero. Thus N_α has measure zero. $\square$

Proof of Lemma 8.8. Since $g_t(i)$ $i = 1, 2, \cdots, I$ are absolutely continuous, so is m. Therefore, $g(1), \cdots, g(I), m$ are almost everywhere differentiable. Let t be a regular point of $g(1), \cdots, g(I), m$ and $\{i_1, \cdots, i_r\}$ be such that $g_t(i_1) = \cdots = g_t(i_r) = m_t$. Note that

$$\begin{aligned}
\max_{1\leq k\leq r} \dot{g}_t(i_k) &= \max_{1\leq k\leq r} \lim_{\epsilon\searrow 0} \frac{g_t(i_k) - g_{t-\epsilon}(i_k)}{\epsilon} \\
&\leq \liminf_{\epsilon\searrow 0} \max_{1\leq k\leq r} \frac{g_t(i_k) - g_{t-\epsilon}(i_k)}{\epsilon} \\
&\leq \liminf_{\epsilon\searrow 0} \frac{m_t - m_{t-\epsilon}}{\epsilon} \\
&= \dot{m}_t.
\end{aligned}$$

Similarly,

$$\min_{1\leq k\leq r} \dot{g}_t(i_r) \geq \limsup_{\epsilon\searrow 0} \frac{m_{t+\epsilon} - m_t}{\epsilon} = \dot{m}_t,$$

and the proof of the lemma is complete. $\square$

Proof of Lemma 9.3.

$$\begin{aligned}
\sum_i a_i w_i &= \sum_i \bar{a}_i w_i + \sum_i (a_i - \bar{a}_i) w_i \\
&\leq \sum_i \bar{a}_i w_i + \sum_i (a_i - \bar{a}_i) w_{min},
\end{aligned}$$

since $a_i \leq \bar{a}_i$ for all i. $\square$

References

[1] Alanyali, M., and Hajek, B. (1996). On Large Deviations in Load Sharing Networks. Submitted to The Annals of Applied Probability.

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

[3] Bertsekas, D.P., and Tsitsiklis, J.N. (1989). *Parallel and Distributed Computing — Numerical Methods.* Prentice-Hall.

[4] Chiu, G. M., Raghavendra, C. S., and Ng, S. M. (1989). Resource Allocation with Load Balancing Consideration in Distributed Computing Systems. *Proceedings of IEEE Infocom 89.* 758-765.

[5] Dai, J.G., and Williams, R.J. (1995). Existence and Uniqueness of Semimartingale Reflecting Brownian Motion in Convex Polyhedrons. *Theory of Probability and its Applications.* **50** 3-53.

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

[7] Ganger, G. R., Worthington, B. L., Hou, R. Y., and Patt, Y. N. (1993). Disk Subsystem Load Balancing: Disk Striping vs. Conventional Data Placement. *Proceedings of* 26^{th} *Hawaii International Conference on System Sciences.* **1** 40-49.

[8] Hajek, B. (1990). Performance of Global Load Balancing by Local Adjustment. *IEEE Trans. on Information Theory.* **36** 1398-1414.

[9] Hunt, P. J., and Kurtz, T. (1994). Large Loss Networks. *Stochastic Processes and Their Applications.* **53** 363-378.

[10] Hunt, P.J., and Laws, C.N. (1993). Asymptotically Optimal Loss Network Control. *Mathematics of Operations Research.* **18** 880-900.

[11] Hunt, P.J., and Laws, C.N. (1995). Optimization via Trunk Reservation in Single Resource Loss Systems under Heavy Traffic. Preprint.

[12] Ibaraki, I., and Katoh, N. (1989). *Resource Allocation Problems—Algorithmic Approaches.* MIT Press.

[13] Liu, H. T., and Silvester, J. (1988). An Approximate Performance Model for Load-dependent Interactive Queues with Application to Load Balancing in Distributed Systems. *Proceedings of IEEE Infocom 88.* 956-965.

[14] Royden, H. L. (1988). *Real Analysis.* Macmillan.

[15] Willebeck-LeMair, M. H., and Reeves, A. P. (1993). Strategies for Dynamic Load Balancing on Highly Parallel Computers. *IEEE Transactions on Parallel and Distributed Systems.* **9** 979-993.

[16] Winston, W. (1977). Optimality of the Shortest-Processing-Time Discipline. J. Applied Probability. **14** 181-189.

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING AND THE COORDINATED SCIENCE LABORATORY, UNIVERSITY OF ILLINOIS, 1308 W. MAIN STREET, URBANA, ILLINOIS 61801