1
On Load Balancing in Erlang Networks

Murat Alanyali and Bruce Hajek

University of Illinois at Urbana Champaign[1]

Abstract

This chapter summarizes our recent work on the dynamic resource allocation problem. The question of interest is the performance of simple allocation strategies which can be implemented on-line. The chapter focuses on the least load routing policy. The analysis is based on fluid limit equations and the theory of large deviations for Markov processes with discontinuous statistics.

1 Introduction

The dynamic resource allocation problem arises in a variety of applications with load sharing features (see Ganger *et al.* (1993), and Willebeck-LeMair and Reeves (1993) for some examples). The generic resource allocation setting involves a number of locations containing resources. The dynamic aspect of the problem is the arrivals of consumers, each of which requires a certain amount of service from the resources, and the control variable of the problem is the "allocation policy", which specifies at which location each consumer is to be served. Oftentimes in applications, locations contain finitely many resources, hence the main objective of the allocation 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 allocation policy to have low complexity, require little information about the system state, and be robust to changes in the traffic parameters.

An instance of resource allocation arises in the wireless network pictured in Fig. 1. The network consists of a number of base stations and users. The users require communication channels that are available at the base stations, whereas each station may serve the users within its geographical range. The resource allocation problem in this setting concerns the question of station selection.

Our mathematical abstraction of a load sharing network is a triple (U, V, N), where 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 Fig. 2 for examples). A

[1]The work in this chapter was sponsored in part by the National Science Foundation under Contract NSF NCR 93 14253 and by a TUBITAK NATO Fellowship.

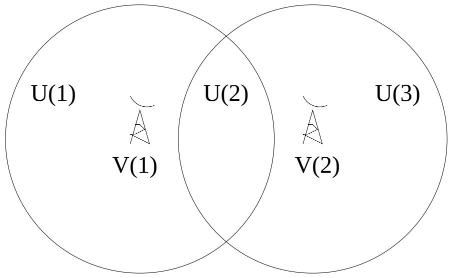

FIG. 1. A typical wireless network with overlapping neighborhoods.

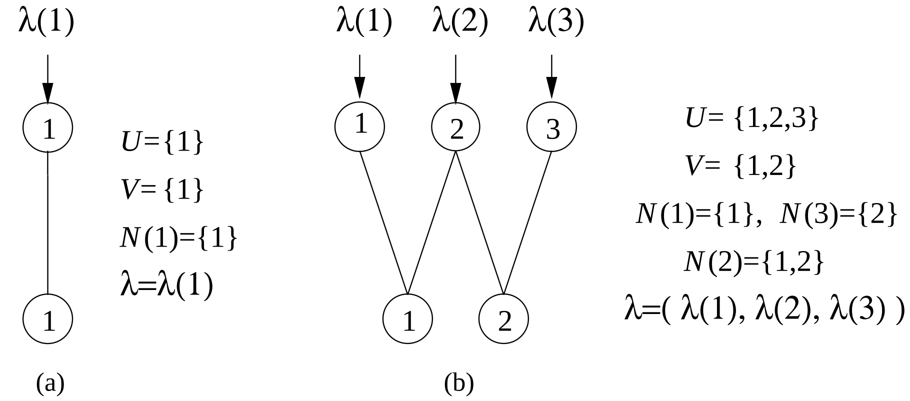

FIG. 2. Two simple load sharing networks: (a) The single-location network, and (b) the W network.

demand for this network is a vector $(\lambda(u) : u \in U)$ of positive numbers, where $\lambda(u)$ 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 which assigns consumers to locations within their respective neighborhoods. The *load* at location $v \in V$ at a given time t, $X_t(v)$, is the number of consumers at v at time t. The allocation policy, together with the consumer arrival and departure times and an initial condition, determine the load process $X = (X_t : t \geq 0)$.

Load balancing is a possible guiding principle for resource allocation, whereby the load is allocated across locations as evenly as possible. 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, $x(1)^2 + x(2)^2$ is minimized over probability vectors $(x(1), x(2))$ by $x(1) = x(2) = 1/2$. This is connected with the convexity of the function $f(x) = x^2$.

A reasonable allocation strategy for dynamic load balancing is the *Least Load*

Routing (LLR) policy, under which each arriving consumer is assigned to a location with the least load in the associated neighborhood. We focus on the analysis of the LLR policy under the following stochastic description of the network dynamics, indexed by a positive scalar γ: 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. The holding time of each consumer is exponentially distributed with unit mean, independent of the past history. Given a positive number κ, we say that the network *overflows* when the load of some location exceeds its designated *capacity* $\lfloor\gamma\kappa\rfloor$.

Under the LLR policy, the load process is Markov with an explicit generator, therefore in principle one can learn much about the load process by computing its equilibrium distribution. Nevertheless, computation of the equilibrium distribution appears intractable for arbitrary network topologies. An alternative approach is to approximate the typical behavior of the network load by *fluid limit approximations*, which are deterministic weak limits of the network load, suitably normalized, for large values of γ.

Certain events of interest, such as network overflow for large enough values of κ, correspond to large deviations of the network from its typical behavior, and hence are not described accurately by the fluid limit approximation. Although network overflow is a rare event, it has a strong impact on network performance; it is thus desirable to estimate its probability and mode accurately. An appropriate tool for this purpose is large deviations theory, which may be employed to study network overflow in terms of *overflow exponents*: The overflow exponent of the network, $F^{LLR}(\kappa)$, and the overflow exponent of a location $v \in V$, $F^{LLR}(v,\kappa)$, under policy LLR are defined as

$$\begin{aligned} F^{LLR}(\kappa) &= -\lim_{T\to\infty}\lim_{\gamma\to\infty}\gamma^{-1}\log P(\text{ Network overflow time} \leq T|X_0=0) \\ F^{LLR}(v,\kappa) &= -\lim_{T\to\infty}\lim_{\gamma\to\infty}\gamma^{-1}\log P(\text{ Overflow time of location } v \leq T|X_0=0). \end{aligned}$$

Note that $F^{LLR}(\kappa) = \min_{v\in V} F^{LLR}(v,\kappa)$ whenever the above quantities exist.

This chapter concerns both the fluid limit approximations and the overflow exponents of load sharing networks operating under various allocation policies. Section 2 identifies the fluid limit approximations of the network load under the LLR policy as solutions of certain integral equations with boundary constraints. Section 3 provides the overflow exponents for the two simple networks of Fig. 2 under LLR, and conjectures the form of the overflow exponents for networks with arbitrary topologies. Finally, extensions of Section 2 to networks with either migration of load or finite capacity constraints, and extensions of Section 3 to two alternative allocation policies are discussed in Section 4.

2 The Fluid Limit Approximation

The fluid limit approximation is a description of the network behavior which is asymptotically exact in heavy traffic. To observe the typical behavior of the network load under the LLR policy, consider the *W network* of Fig. 2.b with the

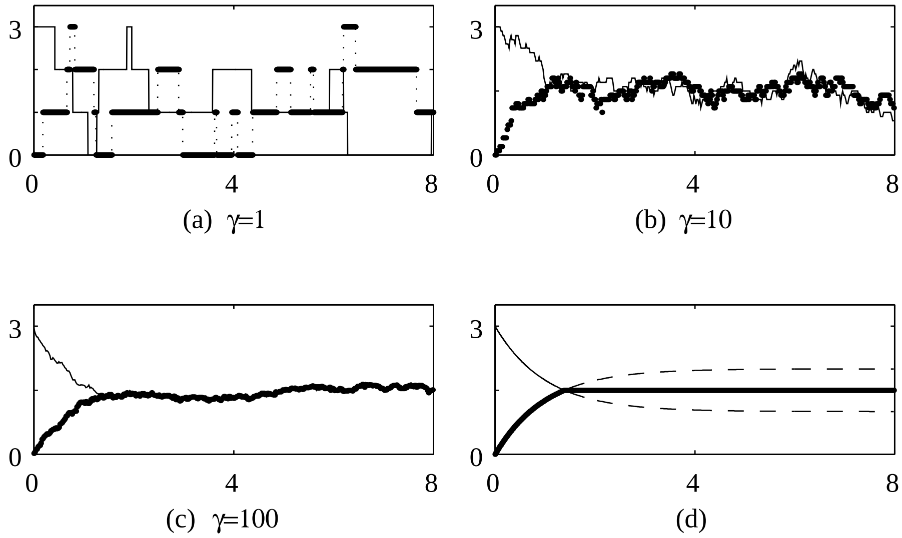

FIG. 3. The normalized load of the W Network under the LLR policy.

demand $\gamma\lambda$ where $\lambda = (1, 1, 1)$, so that consumers of each type arrive at rate γ. Suppose that initially location 1 has zero load, whereas location 2 has load 3γ. Figures 3.a-3.c depict typical sample paths of the *normalized load* (load divided by γ) $X_t^\gamma = (X_t^\gamma(1), X_t^\gamma(2))$ for $\gamma = 1, 10, 100$, during the time interval $[0, 8]$. As γ tends to infinity, the sample paths converge to the limiting trajectory depicted in Fig. 3.d.

Roughly speaking, the sequence of processes $(X^\gamma : \gamma > 0)$ is uniformly stochastically equicontinuous. It can thus be shown using methods of weak convergence (see Ethier and Kurtz (1986)) that if the sequence of initial conditions $(X_0^\gamma : \gamma > 0)$ is tight, then $(X^\gamma : \gamma > 0)$ is tight. Therefore, for any sequence $\gamma_n \to \infty$, there is a subsequence γ_{n_k} such that the distribution of $X^{\gamma_{n_k}}$ converges in Prohorov metric. This is equivalent to weak convergence of $X^{\gamma_{n_k}}$, and the weak limit x, together with some A, satisfies the following equations (Lemma 3.3 of Alanyali and Hajek (1995a)):

$$x_t(v) = x_0(v) + \sum_{u \in N^{-1}(v)} A_{u,v}(t) - \int_0^t x_s(v) ds \tag{2.1}$$

$$A_{u,v}(0) = 0, \quad A_{u,v}(t) \text{ nondecreasing}, \tag{2.2}$$

$$\sum_{v \in N(u)} A_{u,v}(t) = \lambda(u) t, \tag{2.3}$$

$$\int_0^t I\{x_s(v) > \min_{v' \in N(u)} x_s(v')\} dA_{u,v}(s) = 0. \tag{2.4}$$

In the fluid analogy, $A_{u,v}(t)$ corresponds to the amount of type u fluid received by location v up until time t. By (2.1), $x_t(v)$ is the load at location v at time t under assignment regime A and linear discharge rate. Condition (2.3) is a "conservation of fluid" equation, stating that the total type u fluid arrived by time t is equal to the total type u fluid that is assigned to locations by time t. Note that Conditions (2.1)–(2.3) would be satisfied under any allocation policy. Condition (2.4) is the impact of the LLR policy, which implies that location v does not receive any type u fluid unless the load at v is equal to the minimum load in the neighborhood of u.

Solutions of a static optimization problem play an important role in the characterization of the fluid limit approximations: Let an *assignment* a, given by $(a_{u,v} : u \in U, v \in V)$, be called *admissible* if $a \geq 0$, and $a_{u,v} = 0$ whenever $v \notin N(u)$. 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 $\mathcal{A}_\lambda$ denote the set of admissible assignments that satisfy demand λ, and 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

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

Lemma 2.1 *(Alanyali and Hajek (1995a)) 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 $v \in N(u)$*

$$a_{u,v} = 0 \quad whenever \quad q(v) > \min_{v' \in N(u)} q(v').$$

Furthermore, all such assignments yield the same load vector.

The unique load vector corresponding to the solutions of $SLB(\lambda,\Phi)$ is called the *balanced load vector*. One connection between the problem SLB and load balancing is the fact that the balanced load vector minimizes the maximum load $\max_{v \in V} q(v)$ over the admissible assignments satisfying the demand λ (Corollary 3 of Hajek (1990)).

Solutions of the fluid equations (2.1)–(2.4) have the following properties:

Monotonicity : If (x, A) and $(\tilde{x}, \tilde{A})$ are two solutions to the fluid equations with $x_0(v) \geq \tilde{x}_0(v)$ for all $v \in V$, then $x_t(v) \geq \tilde{x}_t(v)$ for all $v \in V$ and $t \geq 0$.

Uniqueness : If (x, A) and $(\tilde{x}, \tilde{A})$ are two solutions to the fluid equations with $x_0 = \tilde{x}_0$, then $x_t = \tilde{x}_t$ for all $t \geq 0$. This 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 denote the unique process such that for some A, (x, A) is a solution to the fluid equations with initial state x_0, where x_0 is arbitrary. Then, $\lim_{t\to\infty} x_t = q$, where q is the balanced load vector for $SLB(\lambda,\Phi)$.

By the uniqueness property, if X_0^γ converges weakly to x_0, then X^γ converges weakly to the process x that solves the fluid equations (2.1)–(2.4) with the initial state x_0. This implies the convergence of X^γ to x on finite time intervals, whereas x converges to the balanced load vector q as time tends to infinity. The following lemma establishes the weak limit of X^γ in steady state. 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 2.2 *The steady state distribution of the process X^γ converges weakly, as γ tends to infinity, to the deterministic distribution concentrated at the balanced load vector q.*

Given an allocation policy π, define for each γ,

$$J_\gamma^\pi = \liminf_{T\to\infty} E^\pi \left[\frac{1}{T} \int_0^T \sum_v X_t(v)^2 dt \,|X_0 = x_0 \right].$$

Note that the function $\Phi(x) = \sum_v x(v)^2$ is convex and symmetric in its coordinates, hence a smaller value of J_γ^π indicates a more balanced load, on average. Lemma 2.2 describes the limiting equilibrium distribution of X^γ, in sufficient detail to imply the following optimality property of the LLR policy:

Theorem 2.3 *For any allocation policy π,*

$$\liminf_{\gamma\to\infty} \gamma^{-2} J_\gamma^\pi \geq \lim_{\gamma\to\infty} \gamma^{-2} J_\gamma^{LLR} = \sum_v q(v)^2.$$

Theorem 2.3 shows that, to first order, the LLR policy achieves asymptotically optimal balancing. Other strategies, such as Bernoulli Splitting covered in Section 4, have the same property. The large deviations analysis described below can distinguish between the strategies.

3 Overflow Exponents

Identifying the overflow exponent of a load sharing network entails establishing a large deviations principle regarding the network load. Let $\tilde{X}^\gamma$ denote the piecewise linearization of the normalized load X^γ at the jump instants. The sequence $(\tilde{X}^\gamma : \gamma > 0)$ is said to *satisfy the large deviations principle* (LDP) in $C_{[0,T]}(R_+^V)$ with the rate function $\Gamma : C_{[0,T]}(R_+^V) \times R_+^V \to R_+ \cup \{+\infty\}$ if for each $x_0 \in R_+^V$ the function $\Gamma(\cdot, x_0)$ is lower semicontinuous, and for any sequence $(x^\gamma : \gamma > 0)$ such that $\lim_{\gamma\to\infty} x^\gamma = x_0$ and Borel measurable $S \subset C_{[0,T]}(R_+^V)$,

$$\begin{aligned}
\limsup_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma} \left((\tilde{X}_t^\gamma : 0 \leq t \leq T) \in S \right) &\leq -\inf_{\phi\in\overline{S}} \Gamma(\phi, x_0) \\
\liminf_{\gamma\to\infty} \gamma^{-1} \log P_{x^\gamma} \left((\tilde{X}_t^\gamma : 0 \leq t \leq T) \in S \right) &\geq -\inf_{\phi\in S^o} \Gamma(\phi, x_0),
\end{aligned}$$

where $\overline{S}$ and S^o denote respectively the closure and the interior of S. This definition of LDP is somewhat weaker than the uniform (in initial conditions) LDP stated in Freidlin and Wentzell (1984), page 92.

Given that $(\tilde{X}^\gamma : \gamma > 0)$ satisfies the large deviations principle in $C_{[0,T]}(R_+^V)$ with the rate function Γ, the overflow exponents are given by the solutions of the following variational optimization problems:

$$F^{LLR}(\kappa) = \inf\{\Gamma(\phi, 0) : \phi \in C_{[0,T]}(R_+^V),\ \phi_0 = 0,\ \max_{v \in V} \phi_T(v) = \kappa,\ T > 0\}$$
$$F^{LLR}(v, \kappa) = \inf\{\Gamma(\phi, 0) : \phi \in C_{[0,T]}(R_+^V),\ \phi_0 = 0,\ \phi_T(v) = \kappa,\ T > 0\}.$$

This section identifies the overflow exponents for the two basic networks of Fig. 2 under the LLR policy, and conjectures their general form for networks with arbitrary topologies.

3.1 The Single-location Network

Large deviations of the single-location network of Fig. 2.a have been studied extensively. In particular Section 12 of Shwartz and Weiss (1995) establishes a large deviations principle for the network load, which can be employed to obtain the following theorem:

Theorem 3.1 *The overflow exponent of the single-location network exists and is given by* $H_{\lambda(1)}(0, \kappa)$, *where*

$$H_{\lambda(1)}(x, y) = \int_x^y \left(\log\left(\frac{z}{\lambda(1)}\right) \right)_+ dz, \qquad y \geq x \geq 0.$$

Intuitively, for $x < y$, $H_{\lambda(1)}(x, y)$ is a measure of how improbable it is for the normalized load process, starting at x, to make a transition to y within a fixed, long time interval. Note that $H_{\lambda(1)}(x, y) = 0$ for $0 \leq x < y \leq \lambda(1)$. The time-dependent extremal trajectories associated with the most likely mode of transition are not indicated directly in $H_{\lambda(1)}$, but they are as follows: If $\lambda(1) < x < y$, the extremal trajectory from x to y satisfies the simple differential equation $\phi' = \phi - \lambda(1)$. This trajectory is the time reversal of the fluid limit trajectory from y to x (see Shwartz and Weiss (1995) for details).

3.2 The W Network

Stochastic ordering arguments provide two upper bounds on the overflow exponent of the W network of Fig. 2.b under *any* allocation policy: *1) The Single Location Bound:* The load at location 1 is stochastically larger than the load of a single-location network with demand $\gamma\lambda(1)$. Therefore the overflow time of a single-location network with capacity $\lfloor \gamma\kappa \rfloor$ and demand $\gamma\lambda(1)$ dominates the overflow time of location 1, which in turn dominates the overflow time of the W network. *2) Pooling Bound:* The network necessarily overflows if the total load exceeds $\lfloor 2\gamma\kappa \rfloor$. Thus the overflow time of the W network is dominated by the overflow time of a single-location network with capacity $\lfloor 2\gamma\kappa \rfloor$ and demand $\gamma(\lambda(1) + \lambda(2) + \lambda(3))$.

The transition mechanism of the load process under LLR changes discontinuously along the boundary $(x \in R_+^2 : x(1) = x(2))$. Large deviations of Markov processes with discontinuous transition mechanisms have been studied by a number of authors: The pioneering paper of Dupuis and Ellis (1992) concerns a Markov process with constant transition mechanism in each of two half spaces, and gives an explicit representation of the rate function for the process observed at a fixed point in time. Subsequently Blinovskii and Dobrushin (1994), and Ignatyuk, Malyshev, and Scherbakov (1994) derived process-level large deviations principles for the case of constant transition mechanism in each halfspace, using different approaches. In their book Shwartz and Weiss (1995) considered processes on a halfspace with a flat boundary which cannot be crossed. The recent paper by Dupuis and Ellis (1995) establishes large deviations principles for Markov processes with transition mechanisms that are continuous over facets generated by a finite number of hyperplanes. While in general the paper does not identify the rate function explicitly, it does give an explicit integral representation for the case of a single hyperplane of discontinuity. The paper by Alanyali and Hajek (1995c) concentrates on processes with a single hyperplane of discontinuity, and contains the same explicit form of the rate function under somewhat relaxed technical conditions, using a different approach.

The large deviations principle satisfied by the load process in the W network is established in Alanyali and Hajek (1995b). The proof is based on Alanyali and Hajek (1995c) regarding the discontinuity of the transition mechanism, and on the techniques used in Section 12.6 of Shwartz and Weiss (1995) regarding the flat boundaries of the state space R^2_+, with departure rates tending to zero at the boundaries.

Let $q = (q(1), q(2))$ denote the balanced load vector for the W network with demand λ. More explicitly,

$$q(1) = [(\lambda(1)+\lambda(2)+\lambda(3))/2]_{\lambda(1)}^{\lambda(1)+\lambda(2)} \quad \text{and} \quad q(2) = [(\lambda(1)+\lambda(2)+\lambda(3))/2]_{\lambda(3)}^{\lambda(3)+\lambda(2)},$$

where for real x, a, b such that $a \leq b$, $[x]_a^b$ denotes the number in the interval $[a, b]$ that is closest to x. We shall assume without loss of generality that $\lambda(1) \geq \lambda(3)$, which in turn implies that $q(1) \geq q(2)$. The following theorems of Alanyali and Hajek (1995b) identify the overflow exponents of the W network. Stephen Turner of Cambridge University mentioned to us at the workshop that he also had found the result in Theorem 3.2, in independent, unpublished work. He outlined a proof that is similar to ours which relies on the the existence of an LDP for the W network, but which does not require an explicit representation of the LDP rate function.

Theorem 3.2 *The overflow exponent of the W network under the LLR policy exists and is given by*

$$F^{LLR}(\kappa) = \begin{cases} H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) & \text{if } \kappa \leq \kappa_* \\ H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa_*) + H_{\lambda(1)}(\kappa_*, \kappa) & \text{if } \kappa > \kappa_*, \end{cases} \tag{3.1}$$

where $\kappa_* = q(1)q(2)/\lambda(1)$.

Here we shall give an intuitive explanation for the formulas appearing in Theorem 3.2. In the remainder of this section, the "load" at a location will be

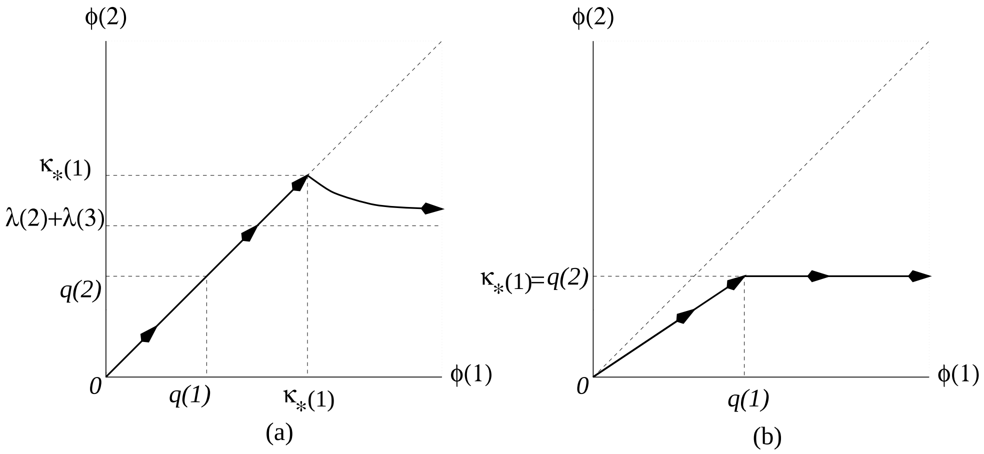

FIG. 4. The most likely scenario for the overflow of location 1 in the W network for the cases (a) $q(1) = q(2)$, (b) $q(1) > q(2)$.

understood to be the normalized load for some suitably large value of γ. Refer to Fig. 4, which depicts the extremal trajectories associated with Theorem 3.2. (The constant $\kappa_*(1)$ in the figure is the same as κ_*). Consider first the case $q(1) = q(2)$. If $\kappa \leq q(1) = q(2)$, $F^{LLR}(\kappa) = 0$, which is expected since network overflow is not a rare event for such κ. If $q(1) = q(2) < \kappa \leq \kappa_*$, then overflow typically occurs because the whole network becomes overloaded, and locations 1 and 2 maintain roughly equal loads and act as a single pooled location. For larger values of κ, the most likely scenario is that first the loads at the two locations together build up to level κ_*, and then the load at location 1 continues to grow to level κ. The given value of κ_* minimizes the right hand side of (3.1) for all κ. Finally, consider the case that $q(1) > q(2)$. Then $q(1) = \lambda(1) > \lambda(2) + \lambda(3) = q(2)$, and it follows that $F^{LLR}(\kappa) = H_{\lambda(1)}(0, \kappa)$ for all values of κ. In this case, the typical scenario for network overflow is that the load at location 1 reaches κ, while the load at location 2 remains near its mean $q(2)$.

Theorem 3.3 *For $v = 1, 2$, the overflow exponent of location v in the W network under LLR exists, and is given by*

$$F^{LLR}(v, \kappa) = \begin{cases} H_{q(1)}(0, \kappa) + H_{q(2)}(0, \kappa) & \text{if } \kappa \leq \kappa_*(v) \\ H_{q(1)}(0, \kappa_*(v)) + H_{q(2)}(0, \kappa_*(v)) + H_{\lambda(2v-1)}(\kappa_*(v), \kappa) & \text{if } \kappa > \kappa_*(v), \end{cases}$$

where $\kappa_*(v) = q(1)q(2)/\lambda(2v - 1)$.

Let us explain the intuition behind Theorem 3.3. First, the very definitions imply that when the overflow exponents exist, $F^{LLR}(1, \kappa) = F^{LLR}(\kappa)$. To see why the given expressions in fact satisfy this relation, note that $\kappa_*(1) = \kappa_*$, and (*i*) if $q(1) = q(2)$ then $H_{q(1)}(0, \kappa) + H_{q(2)}(0, \kappa) = H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa)$ for all κ, (*ii*) if $q(1) > q(2)$ then $\kappa_* = \kappa_*(1) = q(2)$, hence $H_{q(1)}(0, \kappa) + H_{q(2)}(0, \kappa) = H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) = 0$ whenever $\kappa \leq \kappa_*$. Thus, the expressions given for

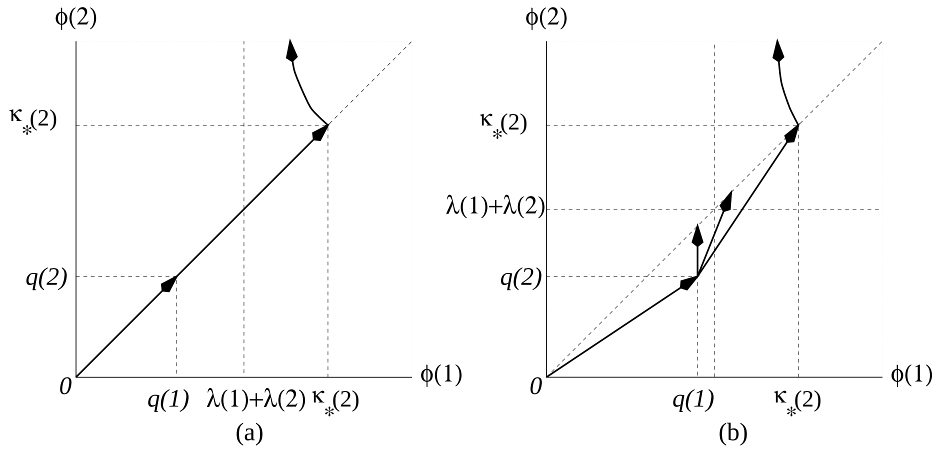

FIG. 5. The most likely scenario for the overflow of location 2 in the W network for the cases (a) $q(1) = q(2)$, (b) $q(1) > q(2)$.

$F^{LLR}(1,\kappa)$ and $F^{LLR}(\kappa)$ are indeed equivalent. The intuitive explanation given for Theorem 3.2 thus also applies to explain the expression for $F^{LLR}(1,\kappa)$.

Finally, let us give an intuitive explanation for the expression for the overflow exponent $F^{LLR}(2,\kappa)$. Consult Fig. 5. If $q(1) = q(2)$, the explanation is similar to that for $F^{LLR}(1,\kappa)$, so assume that $q(1) > q(2)$. If $\kappa \leq q(2)$ then $F^{LLR}(2,\kappa) = 0$, which makes sense since for such κ network overflow is not a rare event. If $q(2) < \kappa \leq q(1)$, then the load at location 2 can grow to level κ, while, even without any large deviation occurring at location 1, all type 2 arrivals are assigned to location 2. Thus, it makes sense that $F^{LLR}(2,\kappa) = H_{q(2)}(0,\kappa)$ for such values of κ. Finally, if $\kappa > q(1) > q(2)$, as the load at location 2 begins to build beyond $q(2)$, the load at location 1 begins to build beyond $q(1)$, even though the two loads are not equal. In that way, all type 2 consumers are assigned to location 2, even after the load at location 2 exceeds $q(1)$. Eventually the loads at the two locations simultaneously become approximately equal to $\kappa \wedge \kappa_*(2)$. If $\kappa > \kappa_*(2)$, then the load at location 2 unilaterally continues to increase to level κ. It is interesting to note that the initial segments of the most likely trajectories depends on κ as κ ranges over $\kappa > q(1) > q(2)$, as illustrated by the multiple trajectories in Fig. 5.b.

To close this section we compare the network overflow exponent of LLR to the single location and pooling upper bounds for a numerical example. Specifically, consider the W network with demand $\lambda = (1-\alpha, 2\alpha, 1-\alpha)$, where $0 \leq \alpha \leq 1$. The network overflow exponent under LLR, $F^{LLR}(\kappa)$, and the two bounds are plotted in Fig. 6 for $\alpha = 0.5$. Also depicted are the overflow exponents of two other policies mentioned in Section 4.3. For the values of $\kappa \leq \kappa_*$ the simple LLR policy performs as well as any other policy, in the sense that $F^{LLR}(\kappa)$ achieves the pooling bound. For larger values of κ however, the suboptimality of LLR reveals itself.

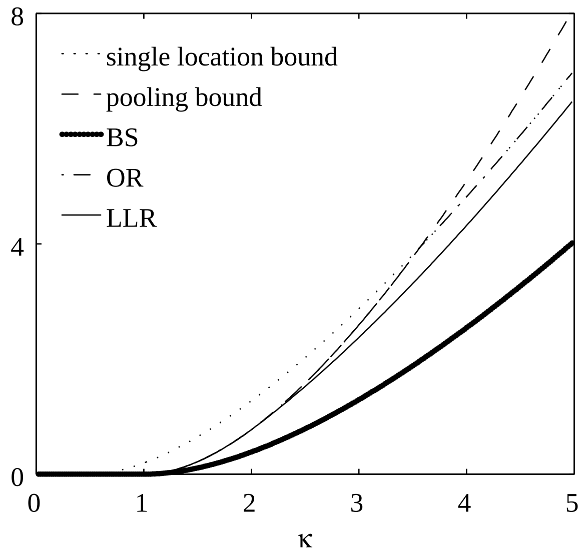

FIG. 6. The overflow exponents of some allocation policies, along with the single-location and pooling bounds, for $\alpha = 0.5$.

Larger values of α correspond to increased load sharing capability of the network for the same total demand. Figure 7 plots $F^{LLR}(\kappa)$ for different values of α. Note that when $\alpha = 0$ and $\alpha = 1$, $F^{LLR}(\kappa)$ achieves the single-location and pooling bounds respectively.

3.3 Arbitrary Topologies

Establishing explicit large deviations principles for arbitrary load sharing networks operating under the LLR policy appears difficult. Nevertheless, the forms of the overflow exponents given in Theorems 3.2 and 3.3 and the corresponding most likely scenarios for network overflow, suggest the following conjecture on the overflow exponents for networks with arbitrary topologies:

Conjecture For each $v \in V$ and $\kappa \geq 0$, $F^{LLR}(v, \kappa)$ can be identified as follows: Let S range over the set of set-valued functions of the form $S = (S(x) : 0 \leq x \leq \kappa)$, where $v \in S(x) \subset V$ for $0 \leq x \leq \kappa$, and $S(x) \subset S(x')$ for $x \geq x'$. Associated with each such S and $0 \leq x \leq \kappa$, let $R(x) = \{u \in U : N(u) \subset S(x) \cup \{v' : q(v') > x\}\}$, and let $(q(v', x) : v' \in S(x))$ denote the balanced load vector in the subnetwork $(R(x), S(x), N(x))$ with demand $(\lambda(u) : u \in R(x))$, where $N(x, u) = N(u) \cap S(x)$ for $u \in R(x)$. Then

$$F^{LLR}(v, \kappa) = \inf_S \int_0^\kappa \sum_{v' \in S(x)} \left(\log\left(\frac{x}{q(v', x)}\right) \right)_+ dx.$$

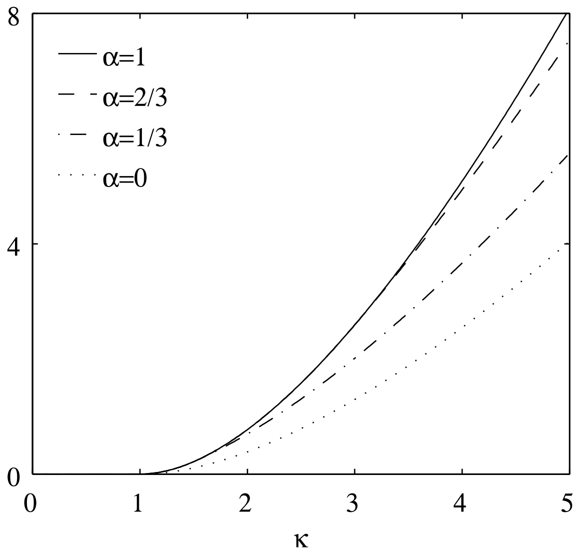

FIG. 7. $F^{LLR}(\kappa)$ for several values of α.

An intuitive justification of the conjecture is as follows. The set $S(x)$ denotes the set of locations with load that would increase to at least level x. The set $R(x)$ is the set of consumer types which would have exceptionally large numbers of arrivals in order to cause the large deviation of the load. The quantity $q(v', x)$ is the nominal arrival rate at location v' at the time the load at v' crosses level x. The conjecture is consistent with Theorem 3.3.

4 Extensions

The first two extensions below deal with the fluid limit analysis, and the third deals with the overflow exponents based on the theory of large deviations.

4.1 The Migration Model

Consider the wireless network of Section 1, where the type of a user 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. In our load sharing network abstraction, this corresponds to migration of consumers among different *types*, and is not covered by the basic dynamic model of Section 1. The migration model is a generalization of the basic model, and incorporates consumer migrations as well as type dependent departure rates from the network.

The analytical description of the migration model involves a *routing matrix* $R = [r_{u,u'}]_{U \times U}$, where $r_{u,u'} \geq 0$ for $u \neq u'$, and $\sum_{u' \in U} r_{u,u'} \leq 0$ for all $u \in U$. For $u, u' \in U$, 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 $-\sum_{u' \in U} r_{u,u'}$.We assume that R is nonsingular, so that every consumer eventually departs from the system. Define the *effective demand*, ρ, as $\rho = -\lambda R^{-1}$ so that $\gamma\rho(u)$ is the mean number of type u consumers in the network in equilibrium. We assume that each new consumer 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.

Let $C_t(u,v)$ denote 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 note that $X_t^\gamma(v) = \sum_u C_t^\gamma(u,v)$. If c is a weak limit of a subsequence of of C^γ, then (c, A) for some process A satisfies the following fluid equations:

$$c_t(u,v) = c_0(u,v) + A_{u,v}(t) + \int_0^t r_{u,u} c_s(u,v)ds \tag{4.1}$$

$$A_{u,v}(0) = 0, \quad A_{u,v}(t) \text{ nondecreasing}, \tag{4.2}$$

$$\sum_{v\in N(u)} A_{u,v}(t) = \lambda(u)t + \sum_{u'\neq u}\int_0^t l_s(u')r_{u',u}ds, \tag{4.3}$$

$$\int_0^t I\{x_s(v) > \min_{v'\in N(u)} x_s(v')\}dA_{u,v}(s) = 0, \tag{4.4}$$

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

In this more general setting, the initial state c_0 and the fluid equations (4.1)-(4.4) do not necessarily determine the load trajectory x uniquely, hence do not necessarily identify a unique weak limit for the normalized load process X^γ. However as t tends to infinity, the load trajectories determined by (4.1)-(4.4) converge to the balanced load vector for problem $SLB(\rho,\Phi)$. This suffices to establish the convergence of the equilibrium distribution of X^γ, hence Theorem 2.3 continues to hold under the migration model (Section 5 of Alanyali and Hajek (1995a)). The overflow exponents for the migration model have not been investigated.

4.2 Finite Capacities

Another variation of the basic model of Section 1 involves incorporation of finite capacity constraints on the locations. Namely, given a nonnegative capacity vector $\kappa = (\kappa(v) : v \in V)$, we consider the case in which for each location v, the number of consumers assigned to the location cannot exceed its capacity $\lfloor\gamma\kappa(v)\rfloor$. A consumer is *lost* if upon its arrival all the locations in its neighborhood are already loaded to capacity. The goal of the allocation policy is to minimize the fraction of consumers lost in the system.

Define $\mathcal{B}_{\lambda,\kappa}$ as the set of admissible assignments a such that $\sum_v a_{u,v} \le \lambda(u)$ for all $u \in U$, and $q(v) \le \kappa(v)$ for all $v \in V$, where q denotes the load vector determined by a. The *Static Load Packing Problem* (SLP) is defined as

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

and provides the following lower bound for the consumer loss probability under *any* allocation policy:

Lemma 4.1 *For any assignment policy π and $\gamma > 0$, the probability of consumer loss satisfies*

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

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

The *Least Ratio Routing* (LRR) policy is a variation of the LLR policy whereby an arriving consumer is assigned to an admissible location with the least *relative load*, $f(x,v)$, defined by $f(x,v) = x(v)/\kappa(v)$. The LRR policy can be analyzed within the framework of lossless networks by introducing a *loss location* v_L of infinite capacity, to which consumers are assigned if they would be lost otherwise. Under the LRR policy, weak limits of the normalized load process satisfy the following fluid equations together with some A:

$$x_t(v) = x_0(v) + \sum_{u\in N^{-1}(v)} A_{u,v}(t) - \int_0^t x_s(v)ds, \tag{4.5}$$

$$A_{u,v}(0) = 0, \ \ A_{u,v}(t) \text{ nondecreasing}, \tag{4.6}$$

$$0 \leq x_t(v) \leq \kappa(v), \qquad \sum_{v\in N(u)\cup v_L} A_{u,v}(t) = \lambda(u)t, \tag{4.7}$$

$$\int_0^t I\{f(x_s,v) > \min_{v'\in N(u)} f(x_s,v')\}dA_{u,v}(s) = 0 \qquad v \neq v_L, \tag{4.8}$$

$$\int_0^t I\{\min_{v\in N(u)} f(x_s,v) < 1\}dA_{u,v_L}(s) = 0. \tag{4.9}$$

The solutions (x,A) of equations (4.5)–(4.9) enjoy the monotonicity and thus the uniqueness properties of x. Furthermore $\lim_{t\to\infty} x_t = q$ where q is the load vector corresponding to a certain solution of $SLP(\lambda,\kappa)$, so that the mechanism of Section 2 can be applied to establish the following theorem (Section 4 of Alanyali and Hajek (1995a)):

Theorem 4.2 *For any allocation policy π,*

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

4.3 Optimal Repacking and Bernoulli Splitting

Optimal Repacking (OR) and *Bernoulli Splitting* (BS) are alternative allocation policies. 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 $SLB(L_t, \Phi)$, where for each $u \in U$, $L_t(u)$ denotes the number of type u consumers in the network at time t. The BS policy 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 $SLB(\lambda, \Phi)$. It is easy to see that both OR and BS policies enjoy the same optimality properties as the LLR policy that are implied by the fluid limit approximations. However, the network overflow exponents under the three policies differ rather drastically.

The process L is a vector of independent single-location network loads with demand vector $\gamma\lambda$. The large deviations principle for the single-location network, together with the Contraction Principle (Theorem 4.2.1 of Dembo and Zeitouni (1992)), yield the large deviations principle for the network load under the OR policy. Solution of the relevant variational optimization problem yields that the network overflow exponent under the OR policy can be expressed as $\min_{F\subset V} H_{\lambda(F)}(0, |F|\kappa)$ where $\lambda(F) = \sum_{u:N(u)\subset F} \lambda(u)$.

Under the BS policy the load process is a vector of independent single-location network loads with the demand vector γq, where q is the balanced load vector for problem $SLB(\lambda, \Phi)$. The techniques mentioned in the above paragraph yield that the network overflow exponent under the BS policy is given by $\min_{v\in V} H_{q(v)}(0, \kappa)$.

Figure 6 plots the network overflow exponents of the OR and BS policies for the W network with demand $\lambda = (0.5, 1, 0.5)$ on the same scale as the LLR policy. The OR policy minimizes the maximum load in the network at all times, hence its overflow exponent dominates the overflow exponent of any allocation policy. Furthermore the OR policy achieves the smaller of the single-location and pooling bounds. The BS policy only exerts open-loop control, and its performance is significantly worse than that of LLR for the whole range of capacities as illustrated in Fig. 6.

Bibliography

1. Alanyali, M. and Hajek, B. (1995a). Analysis of simple algorithms for dynamic load balancing. Submitted to *Mathematics of Operations Research.*
2. Alanyali, M. and Hajek, B. (1995b). On large deviations in load sharing networks. Submitted to *The Annals of Applied Probability.*
3. Alanyali, M. and Hajek, B. (1995c). On large deviations for Markov processes with discontinuous statistics. Submitted to *The Annals of Applied Probability.*
4. Blinovskii, V.M. and Dobrushin, R.L. (1994). Process level large deviations for a class of piecewise homogeneous random walks in *The Dynkin Festschrift: Markov Processes and their Applications*, 1–59, Birkhauser, Boston.
5. Dembo, A. and Zeitouni, O. (1992). *Large deviations techniques and applica-*

tions. Jones and Bartlett, Boston.

6. Dupuis, P., Ellis, R.S. and Weiss, A. (1991). Large deviations for Markov processes with discontinuous statistics, I: General upper bounds. *Annals of Probability*, **19**, 1280–1297.
7. Dupuis, P. and Ellis, R.S. (1992). Large deviations for Markov processes with discontinuous statistics, II: Random walks. *Probability Theory and Related Fields*, **91**, 153–194.
8. Dupuis, P. and Ellis, R.S. (1995). The large deviation principle for a general class of queueing systems I. *Transactions of the American Mathematical Society*, **347** 2689–2751.
9. Ethier, S. and Kurtz, T. (1986). *Markov Processes : Characterization and convergence.* Wiley, New York.
10. Freidlin, M.I. and Wentzell, A.D. (1984). *Random Perturbations of Dynamical Systems.* Springer-Verlag, New York.
11. 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.
12. Hajek, B. (1990). Performance of global load balancing by local adjustment. *IEEE Transactions on Information Theory*, **36**, 1398–1414.
13. Ignatyuk, I.A., Malyshev, V. and Scherbakov, V.V. (1994). Boundary effects in large deviation problems. *Russian Mathematical Surveys*, **49**, 41–99.
14. Shwartz, A. and Weiss, A. (1995). *Large deviations for performance analysis, queues, communication and computing.* Chapman & Hall, London.
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.