On Large Deviations in Load Sharing Networks[1]

Murat Alanyali[2] and Bruce Hajek

Bell Laboratories and the University of Illinois at Urbana Champaign

Three policies, namely Optimal Repacking, Least Load Routing, and Bernoulli Splitting, are considered for dynamic resource allocation in load sharing networks with standard Erlang type statistics. Large deviations principles are established for the three policies in a simple network of three consumer types and two resource locations, and are used to identify the network overflow exponents. The overflow exponents for networks with arbitrary topologies are identified for Optimal Repacking and Bernoulli Splitting policies, and conjectured for the Least Load Routing policy.

[1]Research supported by the National Science Foundation under contract NSF NCR 93-14253.

[2]Completed while a student at the University of Illinois, with partial support provided by a TUBITAK NATO Fellowship

AMS 1991 subject classifications. Primary 60K30; secondary 60F10, 68M20, 90B15, 93E, 60J.

Key words and phrases. Large deviations, load balancing, loss networks, allocation, Erlang systems.

Abbreviated title. Large deviations in networks.

1 Introduction

The generic dynamic resource allocation problem involves a number of locations containing resources. The dynamic aspect of the problem is the arrival 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 network can get congested and may eventually overflow in the sense that consumers may be lost. The purpose of this work is to provide estimates on the network overflow time under certain resource allocation policies. In particular we concentrate on three allocation policies, namely *Optimal Repacking* (OR), *Least Load Routing* (LLR), and *Bernoulli Splitting* (BS). The OR policy continuously repacks the consumers in the network so as to minimize the maximum load over the locations. LLR and BS are *non-repacking* policies, under which the assignment decisions are irrevocable. The LLR policy assigns each arriving consumer to an admissible location with the least load, whereas the BS policy assigns each arriving consumer to one of the admissible locations randomly and independently, so as to minimize the maximum *mean* load over the locations.

The policies investigated in this paper have been considered by several authors and in various contexts. In particular, Gibbens, Kelly, and Turner (1993) considered the LLR policy for dynamic routing in circuit switched networks. Azar, Broder, and Karlin (1992) compared the LLR and the optimal nonrepacking policies based on a worst case analysis, whereas Alanyali and Hajek (1997) compared the three policies based on certain properties implied by fluid limit approximations.

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 Figure 1 for examples). A *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.

Provided a demand vector λ, we consider the following stochastic description of the network dynamics, indexed by a scalar $\gamma > 0$: 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 $\kappa \geq 0$, the *overflow time of a location* $v \in V$ is the first time that its load, $X_t(v)$, exceeds the designated capacity $\lfloor \gamma\kappa \rfloor$, and the *network overflow time* is the minimum over all v of

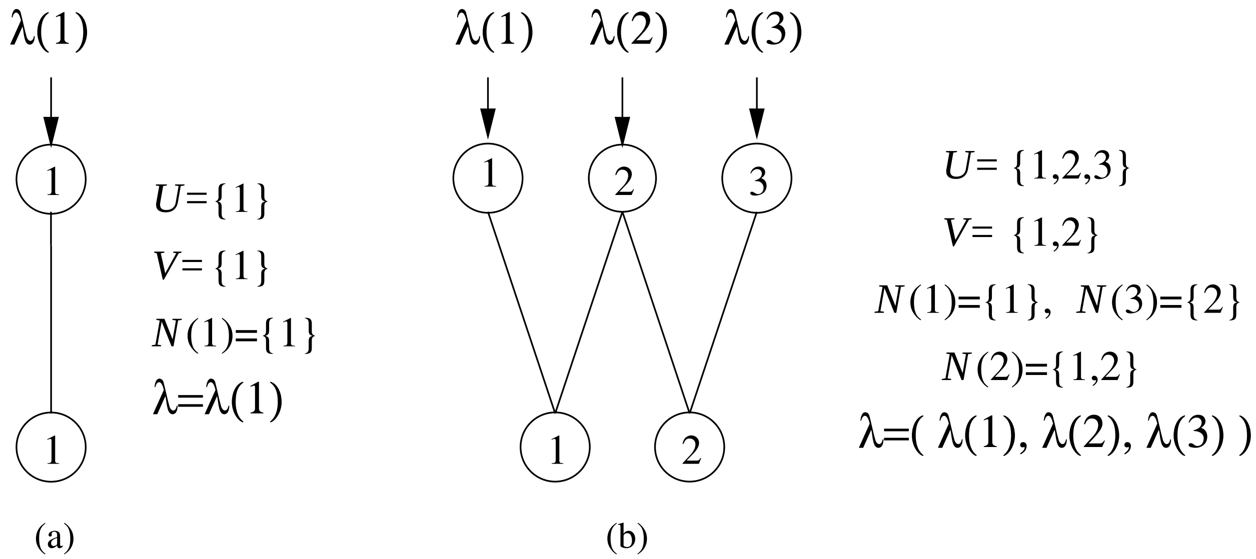

Figure 1: Two load sharing networks of interest: (a) The single-location network, and (b) the W network.

the overflow times of location v.

Under allocation policy π, the *overflow exponent of the network*, $F^\pi(\kappa)$, and the *overflow exponent of a location* v, $F^\pi(v,\kappa)$, are defined as

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

It is easy to see that $F^\pi(\kappa) = \min_{v\in V} F^\pi(v,\kappa)$ whenever the above quantities exist. A crude interpretation of the overflow exponent of the network is that for fixed but large T, $P($ Network overflow time $\le T|X_0 = 0) \approx \exp(-\gamma F^\pi(\kappa))$. Note that larger values of the overflow exponent indicate larger overflow times. The approach taken in this paper is to compare allocation policies based on the corresponding overflow exponents. The rest of this section states the main results. We start with the two basic networks of Figure 1.

The Single-Location Network. The *single-location network* of Figure 1.a has been studied extensively in the context of Erlang's model for circuit switched traffic. In particular, the following theorem can be obtained by applying the results in Section 12 of Shwartz and Weiss (1995). Details of the proof are given in Section 3, since the basic notation and concepts carry over to analysis of more general network topologies.

Theorem 1.1 *(Single Location) The overflow exponent of the single-location network exists and is given by* $H_{\lambda(1)}(0,\kappa)$, *where*

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

Intuitively, for $x < y$, $H_{\lambda(1)}(x,y)$ is a measure of how unlikely it is for the normalized load, starting at level x, to reach level y within a fixed, long time interval. Note that $H_{\lambda(1)}(x,y) = 0$ for $0 \leq x < y \leq \lambda(1)$, since such transition is not a rare event in this case. The reader is referred to Shwartz and Weiss (1995) for large deviations exponents for transitions within *fixed* time duration. Fairly explicit solutions for the extremization problem exist for the one dimensional systems. As described on page 267 of Shwartz and Weiss (1995), the difference in overflow exponent for T finite is larger than for unconstrained overflow time by a term asymptotically equivalent to $(1-\lambda(1))\exp(-T)$ (in case $\lambda(1) < \kappa = 1$) as $T \to \infty$. While the finite time calculations can probably be carried through for the W network considered here, for simplicity, we concentrate on the unconstrained hitting time case.

The W Network. In the *W network* of Figure 1.b it is assumed without loss of generality that $\lambda(1) \geq \lambda(3)$. We first discuss two upper bounds on the network overflow time which apply to *any* allocation policy, and then provide three theorems which identify the overflow exponents under the policies of interest. The proofs of the theorems are the subjects of subsequent sections.

Stochastic ordering arguments provide the two upper bounds on the overflow time of the W network 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 $\gamma\lambda(1)$. Hence 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 network. *2) Pooling Bound:* The network necessarily overflows if the total load exceeds $\lfloor 2\gamma\kappa \rfloor$. Thus the overflow time of the network is dominated by the overflow time of a single-location network with capacity $\lfloor 2\gamma\kappa \rfloor$ and demand $\gamma(\lambda(1)+\lambda(2)+\lambda(3))$.

We now discuss the three policies, starting with some essential definitions: For real x, a, b such that $a \leq b$, let $[x]_a^b$ denote the number in the interval $[a, b]$ that is closest to x. Let $q(1) = \lambda(1)+p\lambda(2)$ and $q(2) = \lambda(3) + (1-p)\lambda(2)$ where $p \in [0,1]$ is chosen to minimize $|q(1) - q(2)|$. More explicitly,

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

and $p = [(\lambda(3) - \lambda(1) + \lambda(2))/2\lambda(2)]_0^1$. The assumption $\lambda(1) \geq \lambda(3)$ implies that $q(1) \geq q(2)$.

Consider first the BS policy, under which each type 2 consumer is assigned to location 1 with probability p, or to location 2 with probability $(1-p)$. The load at each location v behaves as in a single-location network with demand $\gamma q(v)$, independent of the other location. In turn the overflow exponent of each location can be obtained by appealing to the single-location result, and the overflow exponent of the network is equal to that of the more heavily loaded location 1:

Theorem 1.2 *(BS) For $v = 1, 2$, the overflow exponent of location v under the BS policy exists and is given by $F^{BS}(v, \kappa) = H_{q(v)}(0, \kappa)$. In particular $F^{BS}(\kappa) = H_{q(1)}(0, \kappa)$.*

As for the BS policy, the network load under the OR policy can be represented in terms of single-location loads. Under the OR policy, network overflow occurs at the first time that one of the following three happens: the number of type 1 consumers exceeds $\lfloor \gamma\kappa \rfloor$, the number of type 3 consumers exceeds $\lfloor \gamma\kappa \rfloor$, or the total number of consumers exceeds $2\lfloor \gamma\kappa \rfloor$. The following theorem holds:

Theorem 1.3 *(OR) The network overflow exponent under the OR policy is given by $F^{OR}(\kappa) = \min\{H_{\lambda(1)}(0,\kappa), H_{\lambda(3)}(0,\kappa), H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa)\}$, or equivalently (using the assumption $\lambda(1) \geq \lambda(3)$),*

$$F^{OR}(\kappa) = F^{OR}(1,\kappa) = \begin{cases} H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) & \text{if } \kappa \leq \kappa_o \\ H_{\lambda(1)}(0,\kappa) & \text{if } \kappa > \kappa_o, \end{cases}$$

where κ_o is the larger root of $\kappa_o = \kappa_o \log(\kappa_o \lambda(1)/q(1)q(2)) + \lambda(2) + \lambda(3)$. Also,

$$F^{OR}(2,\kappa) = \min\{H_{\lambda(3)}(0,\kappa),\ \min\{H_{\lambda(1)}(0,a)+H_{\lambda(2)}(0,b)+H_{\lambda(3)}(0,c) :\ a,b,c \geq 0,\ b+c \geq \kappa,\ a+b+c \geq 2\kappa\}\}.$$

The overflow exponents under the LLR policy are identified by the following theorem, for which we provide somewhat detailed comments. For simplicity, it is assumed that if the locations are equally loaded, an arriving type 2 consumer is assigned to location 1.

Theorem 1.4 *(LLR) For $v = 1, 2$, the overflow exponent of location v under the LLR policy exists and is given by*

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

where $\kappa_*(v) = q(1)q(2)/\lambda(2v-1)$. In particular $F^{LLR}(\kappa) = F^{LLR}(1,\kappa)$, which can also be expressed as

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

Remark 1.1 *To see the equivalence of the two expressions for $F^{LLR}(\kappa)$, note that (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, and (ii) if $q(1) > q(2)$ then $\lambda(1) = q(1) > q(2) = \kappa_*(1)$, hence $H_{q(1)}(0,\kappa) + H_{q(2)}(0,\kappa) = H_{\lambda(1)+\lambda(2)+\lambda(3)}(0,2\kappa) = 0$ whenever $\kappa \le \kappa_*(1)$, so that $F^{LLR}(\kappa) = F^{LLR}(1,\kappa) = H_{\lambda(1)}(0,\kappa)$ for all κ.*

Here we give an intuitive explanation for the formulas appearing in Theorem 1.4. In the present and the next paragraphs, the "load" at a location is understood to be the normalized load for some suitably large value of γ. Focusing first on location $v = 1$, refer to Figure 2 which pictures the extremal trajectories of the limiting scaled load $(\phi(1), \phi(2))$ associated with Theorem 1.4 for $v = 1$. Consider first the case $q(1) = q(2)$. If $\kappa \le q(1) = q(2)$, then $F^{LLR}(1,\kappa) = 0$, which is expected since overflow of location 1 is not a rare event for such κ. If $q(1) = q(2) < \kappa \le \kappa_*(1)$, then overflow in location 1 typically occurs because the whole network becomes overloaded, and both locations maintain roughly equal loads. For larger values of κ, the most likely scenario is that first the loads at the two locations together build up to level $\kappa_*(1)$, and then the load at location 1 continues to grow to level κ. The given value of $\kappa_*(1)$ minimizes the expression for $F^{LLR}(1,\kappa)$. Finally, consider the case that $q(1) > q(2)$. Then $F^{LLR}(1,\kappa) = H_{\lambda(1)}(0,\kappa)$ as explained in Remark 1.1, and the typical scenario for overflow of location 1 is that the load at location 1 reaches κ, while the load at location 2 relaxes towards its mean $q(2)$.

Now focusing on location 2, let us give an intuitive explanation for the expression for the overflow exponent $F^{LLR}(2,\kappa)$. Consult Figure 3. 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 $F^{LLR}(2,\kappa) = 0$, 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 depend on κ as κ ranges over $\kappa > q(1) > q(2)$, as illustrated by the multiple trajectories in Figure 3.b.

As a numerical example to compare the three policies, consider the W network with demand $\lambda = (1-\alpha, 2\alpha, 1-\alpha)$ where $0 \leq \alpha \leq 1$. The network overflow exponents under the three policies are plotted in Figure 4, along with the single-location and pooling upper bounds, for the case $\alpha = 0.5$. The OR policy employs the tightest possible control, hence the network overflow time under OR dominates the network overflow time under *any* allocation policy. Furthermore, in the W network, $F^{OR}(\kappa)$ is equal to the smaller of the single-location and pooling bounds. ¿From a practical point of view the OR policy has drawbacks such as high computational complexity and the required repacking of consumers. For the values of $\kappa \leq \kappa_*(1)$ the simple, nonrepacking LLR policy performs as well as any other policy, in the sense that $F^{LLR}(\kappa) = F^{OR}(\kappa)$. For larger values of κ the nonrepacking nature of LLR reveals itself, and LLR is outperformed by OR. The BS policy only exerts open-loop control, and its performance is significantly worse than the LLR policy for the whole range of capacities as illustrated in Figure 4.

Consider also the dependence of $F^{LLR}(\kappa)$ on α, illustrated in Figure 5. Larger values of α correspond to increased load sharing capability of the network for the same total demand, so it is

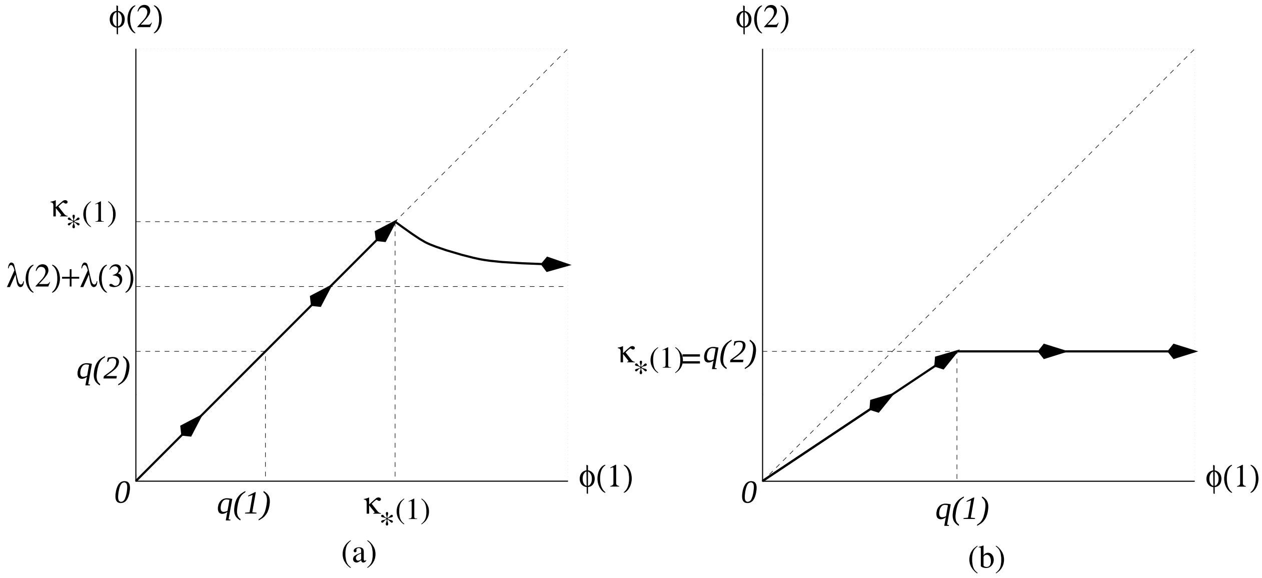

Figure 2: The most likely scenario for the overflow of location 1 of the W network under the LLR policy, for the cases (a) $q(1) = q(2)$, (b) $q(1) > q(2)$.

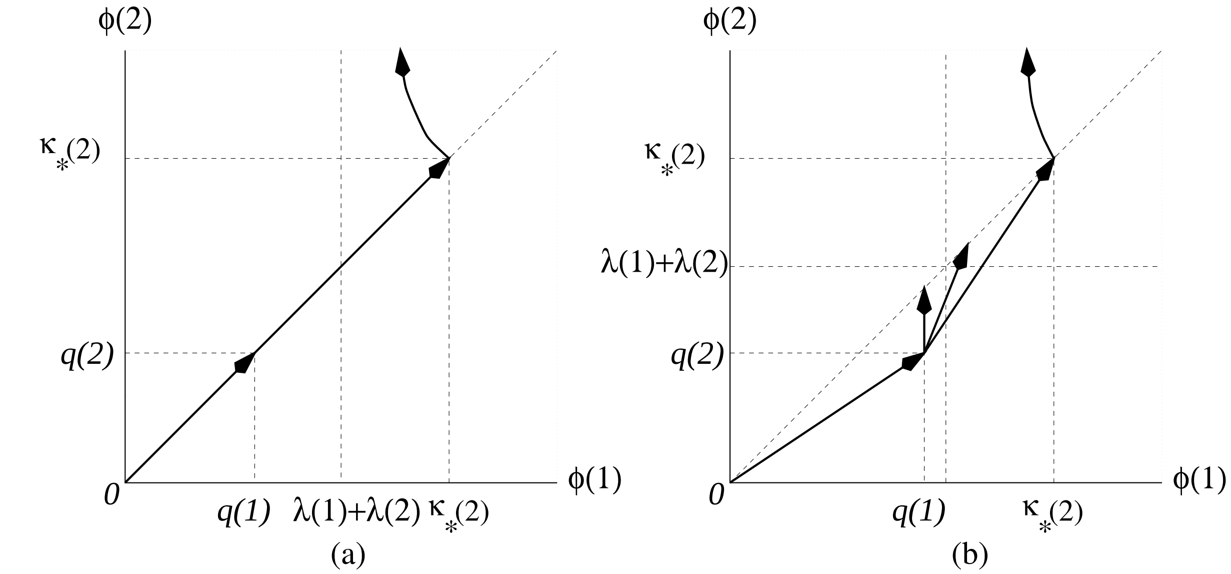

Figure 3: The most likely scenario for the overflow of location 2 of the W network under the LLR policy, for the cases (a) $q(1) = q(2)$, (b) $q(1) > q(2)$.

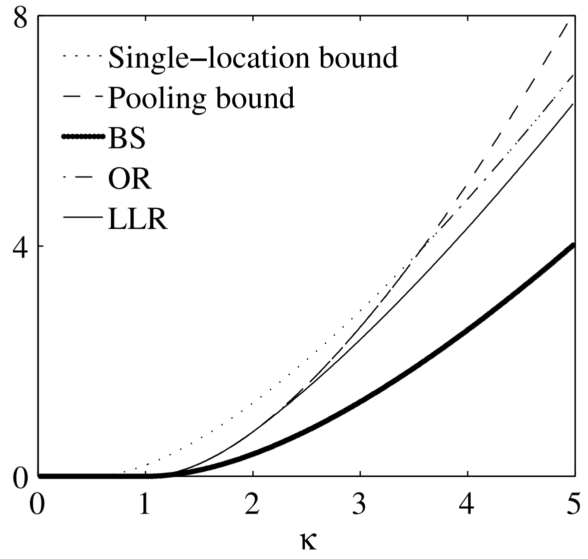

Figure 4: The network overflow exponents of the three policies, along with the single-location and pooling bounds, for $\alpha = 0.5$.

not surprising that $F^{LLR}(\kappa)$ is increasing in α. Note that when $\alpha = 0$ and $\alpha = 1$, $F^{LLR}(\kappa)$ achieves respectively the single-location and pooling bounds.

Networks with Arbitrary Topologies. We next consider the overflow exponents of networks with arbitrary topologies under the three policies. Some definitions are in order: Let a load sharing network (U, V, N) and a demand vector λ be given. An *assignment* a, given by $(a_{u,v} : u \in U, v \in V)$, is said to be *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*. By Lemma 2.1 of Alanyali and Hajek (1997), there exists an admissible assignment satisfying the demand λ which minimizes $\sum_{v \in V}(q(v))^2$, and all such assignments yield the same *balanced* load vector. By Corollary 3 of Hajek (1990) the balanced load vector minimizes $\max_v q(v)$ over admissible assignments satisfying λ. In what follows, let q denote the balanced load vector and a denote an assignment with load q.

The BS policy assigns each type u consumer to location $v \in N(u)$ with probability $a_{u,v}/\lambda(u)$, so that the load at each location v behaves as an independent single-location network load with

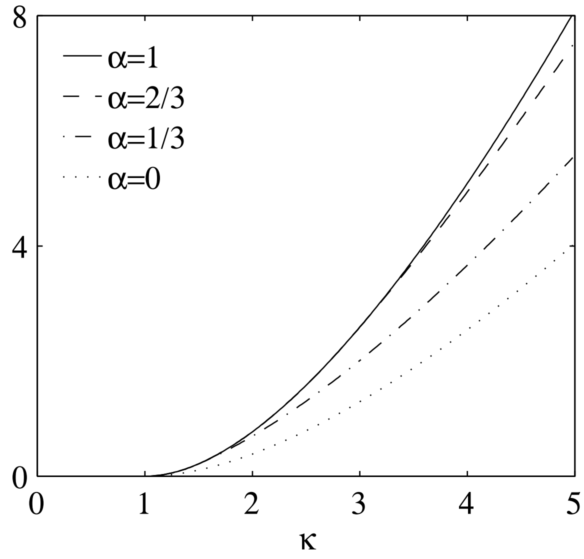

Figure 5: $F^{LLR}(\kappa)$ for several values of α.

demand $\gamma q(v)$. By changing notation to account for an arbitrary number of locations, the proof of Theorem 1.2 can be generalized to yield the following theorem.

Theorem 1.5 *(BS) For $v \in V$, the overflow exponent of location v under the BS policy exists and is given by $F^{BS}(v, \kappa) = H_{q(v)}(0, \kappa)$. In particular $F^{BS}(\kappa) = \min\{H_{q(v)}(0, \kappa) : v \in V\}$.*

To analyze the OR policy in general networks, let $L_t(u)$ denote the number of type u consumers in the network at time t, and define $L_t = (L_t(u) : u \in U)$. The network overflow time is the first time t such that there is a subset A of locations such that $\sum_{u:N(u)\subset A} L_t(u) > \lfloor \gamma\kappa \rfloor |A|$ (see Corollary 7 of Hajek (1990)). Therefore, just changing the notation in the proof of Theorem 1.2 to account for an arbitrary number of locations yields the following theorem.

Theorem 1.6 *(OR) The network overflow exponent under the OR policy is given by*

$$F^{OR}(\kappa) = \min_{A \subset V} H_{\lambda(A)}(0, \kappa|A|)$$

where $\lambda(A) = \sum_{u:N(u)\subset A} \lambda(u)$.

Except for simple network topologies such as the W network, the load process under the LLR policy has discontinuous statistics along complicated geometries. Due to this fact, establishing explicit large deviations principles for arbitrary load sharing networks appears difficult. Nevertheless, the form of the overflow exponents provided by Theorem 1.4, together with the extremal paths of Figures 2 and 3 suggest the following conjecture:

Conjecture 1.1 *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 $v \subset S(x) \subset V$ 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*

$$F^{LLR}(v, \kappa) = \inf_{S} \int_0^{\kappa} \sum_{v' \in S(x)} \max\left(0, \log(\frac{x}{q(v', x)})\right) dx.$$

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 and small numbers of departures in order to cause the load at locations in $S(x)$ to grow beyond x. 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 1.4.

The rest of the paper is organized as follows: Section 2 consists of the basic definitions regarding the techniques employed in the analysis, namely the theory of large deviations. Theorem 1.1 regarding the single-location network is proved in Section 3, and Theorems 1.2-1.4 regarding the W network are proved in Section 4. A large deviations principle for the W network under the LLR policy is stated as a proposition in Section 4, and is proved in Section 5.

2 Definitions

Given a positive integer d, let R^d denote the d dimensional Euclidean space. A collection $\nu = (\nu(x) : x \in R^d)$ is called a *rate-measure field* if for each x, $\nu(x) = \nu(x, \cdot)$ is a positive Borel measure on R^d, and $\sup_x \nu(x, R^d) < \infty$. For each positive scalar γ, a right continuous Markov jump process X^γ is said to be *generated* by the pair (γ, ν) if given its value at time t, the process X^γ jumps after a random time exponentially distributed with parameter $\gamma\nu(X_t^\gamma, R^d)$, and the jump size is a random variable Δ where $\gamma\Delta$ has distribution $\nu(X_t^\gamma)/\nu(X_t^\gamma, R^d)$, independent of the past history. The *polygonal interpolation* of the process X^γ, $\tilde{X}^\gamma$, is defined as

$$\tilde{X}_t^\gamma = \frac{t - \tau_k}{\tau_{k+1} - \tau_k} X_{\tau_{k+1}}^\gamma + \frac{\tau_{k+1} - t}{\tau_{k+1} - \tau_k} X_{\tau_k}^\gamma \qquad \tau_k \leq t \leq \tau_{k+1},$$

where τ_k is the k^{th} jump time of X^γ. Since X^γ has a finite number of jumps in bounded time intervals $\tilde{X}^\gamma$ has sample paths in $C_{[0,\infty)}(R^d)$, the space of continuous functions $\phi : [0, \infty) \to R^d$ with the topology of uniform convergence on compact sets.

The following are some standard definitions of large deviations theory. Let $\mathcal{X}$ be a topological space and let Z^γ denote a $\mathcal{X}$-valued random variable for each $\gamma > 0$. The sequence $(Z^\gamma : \gamma > 0)$ is said to satisfy the large deviations principle with rate function $\Gamma : \mathcal{X} \to R_+ \cup \{\infty\}$ if Γ is lower semicontinuous, and for any Borel measurable $S \subset \mathcal{X}$,

$$\begin{aligned} \limsup_{\gamma \to \infty} \gamma^{-1} \log P(Z^\gamma \in S) &\leq -\inf_{z \in \overline{S}} \Gamma(z) \\ \liminf_{\gamma \to \infty} \gamma^{-1} \log P(Z^\gamma \in S) &\geq -\inf_{z \in S^o} \Gamma(z), \end{aligned}$$

where $\overline{S}$ and S^o denote respectively the closure and the interior of S. The rate function Γ is called *good* if for each $l \geq 0$ the level set $\{z : \Gamma(z) \leq l\}$ is compact.

The large deviations principles stated in this paper are for $\mathcal{X}$ of the following form. It is a space of continuous functions on $[0, T]$ with values in some subset D of a finite dimensional Euclidean space, $\mathcal{X} = C_{[0,T]}(D)$, with topology given by the sup norm. The corresponding rate functions depend on T, but for brevity the letter "T" is suppressed from the notation.

We use the notation $a \vee b$ (respectively $a \wedge b$) to denote $\max(a, b)$ (respectively $\min(a, b)$), and I_A to denote the indicator function of a set A.

3 The Single-Location Network

This section presents the proof of Theorem 1.1. The essential ingredient of the proof is Lemma 3.1 which establishes a large deviations principle for the load process. In view of Lemma 3.1 the proof of Theorem 1.1 hinges on the solution of a variational optimization problem which is provided by Lemma 3.4.

The *normalized load process* X^γ, defined as $X^\gamma = \gamma^{-1}X$, is a Markov jump process which takes values in R_+ with initial value x^γ, such that γx^γ is a nonnegative integer. It is generated by the pair (γ, ν), where for each $x \in R_+$, $\nu(x, \{1\}) = \lambda(1)$, $\nu(x, \{-1\}) = x$, and $\nu(x, \{1, -1\}^c) = 0$. Note that $\gamma\nu(x, \{1\})$ and $\gamma\nu(x, \{-1\})$ are respectively the consumer arrival and departure rates when the normalized load is x. The polygonal interpolation of the normalized load process, $\tilde{X}^\gamma$, satisfies a large deviations principle as identified by the following lemma. The lemma is a slight variation of the results in Section 12 of Shwartz and Weiss (1995) which assume a bounded state space, and it follows by taking $\lambda(2) = 0$ in Lemma 4.2. Its proof is therefore omitted.

Lemma 3.1 *Suppose* $\lim_{\gamma\to\infty} x^\gamma = x_o$ *for some* $x_o \in R_+$. *Then the sequence* $(\tilde{X}^\gamma : \gamma > 0)$ *satisfies the large deviations principle in* $C_{[0,T]}(R_+)$ *with the good rate function* $\Gamma_{\lambda(1)}(\cdot, x_o)$, *where for each* $\phi \in C_{[0,T]}(R_+)$ *and* $x \in R_+$,

$$\Gamma_{\lambda(1)}(\phi, x) = \begin{cases} \int_0^T \Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t)dt & \text{if } \phi_0 = x \text{ and } \phi \text{ is absolutely continuous} \\ +\infty & \text{otherwise,} \end{cases}$$

and

$$\Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t) = \dot{\phi}_t \log\left(\frac{\dot{\phi}_t + \sqrt{\dot{\phi}_t^2 + 4\lambda(1)\phi_t}}{2\lambda(1)}\right) + \phi_t + \lambda(1) - \sqrt{\dot{\phi}_t^2 + 4\lambda(1)\phi_t}.$$

The parameter T is fixed in the above lemma. In solving the variational problem associated with Theorem 1.1, the parameter T is allowed to be arbitrarily large. Intuitively, the solution functions ϕ are not time constrained in the large T limit, and the slopes of the functions can be choosen to minimize the total cost of traversing from one point to another, when time is not constrained. The next two remarks discuss the optimal slopes when traveling towards, respectively away from, the stable point $\lambda(1)$.

Remark 3.1 *For fixed* $\lambda(1)$ *and* ϕ_t*, the function* $\Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t)$ *is a strictly convex, nonnegative function of* $\dot{\phi}_t$*. Furthermore* $\Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t) = 0$ *if and only if* $\dot{\phi}_t = \lambda(1) - \phi_t$*, in which case we say that* ϕ relaxes *under* $\lambda(1)$.

Remark 3.2 *If* $(\lambda(1)-\phi_t)\dot{\phi}_t < 0$ *then* $\Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t) \geq \inf_{\alpha>0} \Lambda_{\lambda(1)}(\phi_t, \alpha\dot{\phi}_t)/\alpha = \dot{\phi}_t \log(\phi_t/\lambda(1))$*. The equality holds if and only if* $\dot{\phi}_t = \phi_t - \lambda(1)$*, in which case we say that* ϕ relaxes in reverse time *under* $\lambda(1)$*. (Intuitively, this means that if* ϕ_t *is moving away from the point of attraction* $\lambda(1)$ *with no constraint on time, then the optimal slope is* $\phi_t - \lambda(1)$*. This is to be expected, as the terminology suggests, due to Remark 3.1 and the time-reversibility of* $\tilde{X}^{\gamma}$*.) Thus for absolutely continuous* ϕ *with* $\phi_0 < \phi_T$*,*

$$\begin{aligned}\int_0^T \Lambda_{\lambda(1)}(\phi_t, \dot{\phi}_t)dt &\geq \int_0^T \dot{\phi}_t \log(\frac{\phi_t}{\lambda(1)}) I_{\{\dot{\phi}_t>0,\ \phi_t>\lambda(1)\}} dt \\ &\geq \int_{\phi_0}^{\phi_T} \max(0, \log(\frac{x}{\lambda(1)}))dx,\end{aligned}$$

and therefore $\Gamma_{\lambda(1)}(\phi, \phi_0) \geq H_{\lambda(1)}(\phi_0, \phi_T)$*. If also* $\phi_0 > \lambda(1)$*, then equality holds if and only if* ϕ *relaxes in reverse time under* $\lambda(1)$*.*

Lemma 3.2 *For each* $x \geq 0$*,* $y \in R$*, and* $\epsilon > 0$*,* $\Lambda_{\lambda(1)}(x+\epsilon, y) \leq \Lambda_{\lambda(1)}(x, y) + \epsilon$*.*

Proof. The lemma follows by the fact that for all $x \geq 0$ and $y \in R$,

$$\frac{\partial \Lambda_{\lambda(1)}(x,y)}{\partial x} = 1 - \frac{2\lambda(1)}{y + \sqrt{y^2 + 4\lambda(1)x}} \leq 1.$$

□

Lemma 3.3 *For each* $0 \leq x \leq y$ *and* $T \geq 0$*,*

$$\inf\{\Gamma_{\lambda(1)}(\phi, x) : \phi_0 = x,\ \sup_{0\leq t\leq T} \phi_t > y\} = \inf\{\Gamma_{\lambda(1)}(\phi, x) : \phi_0 = x,\ \sup_{0\leq t\leq T} \phi_t \geq y\}.$$

Proof. To prove the lemma, it suffices to show that

$$\inf\{\Gamma_{\lambda(1)}(\phi,x) : \phi_0 = x,\ \sup_{0\leq t\leq T}\phi_t > y\} \leq \inf\{\Gamma_{\lambda(1)}(\phi,x) : \phi_0 = x,\ \sup_{0\leq t\leq T}\phi_t \geq y\}. \tag{3.1}$$

Fix $\epsilon > 0$. By the goodness of $\Gamma_{\lambda(1)}(\cdot,x)$ there exists a solution ϕ to the right hand side of inequality (3.1), and clearly $\Gamma_{\lambda(1)}(\phi,x)$ is finite. Set $M = \sup_{0\leq t\leq T}\phi_t < \infty$, and choose B large enough so that $\inf_{0\leq z\leq M}\Lambda_{\lambda(1)}(z,\dot\phi_t) > \Gamma_{\lambda(1)}(\phi,x)/T$ whenever $\dot\phi_t > B$, and such that the set $S = \{t\in[0,T] : \dot\phi_t \leq B\}$ has positive measure. (For example, if $\dot\phi$ is bounded, B can be taken so that $\dot\phi_t \leq B$ for $0\leq t\leq T$.) Then $\xi \in C_{[0,T]}(R_+)$ defined by $\xi_0 = \phi_0$ and $\dot\xi_t = \dot\phi_t + \epsilon I_{\{t\in S\}}$ satisfies $\sup_{0\leq t\leq T}\xi_t > y$. Lemma 3.2 and the fact that $\partial\Lambda_{\lambda(1)}(\phi_t,y)/\partial y$ is bounded on S imply that $\Gamma_{\lambda(1)}(\xi,x) \leq \Gamma_{\lambda(1)}(\phi,x) + f(\epsilon)$ for some $f(\epsilon)\to 0$ as $\epsilon\to 0$. The arbitrariness of $\epsilon > 0$ proves the lemma. □

Each of the previous lemmas concerned a fixed value of T, whereas the next lemma involves all values of T.

Lemma 3.4 *For each* $0\leq x\leq y$,

$$\inf\{\Gamma_{\lambda(1)}(\phi,x) : T\geq 0,\ \phi\in C_{[0,T]}(R_+),\ \phi_0 = x,\ \sup_{0\leq t\leq T}\phi_t \geq y\} = H_{\lambda(1)}(x,y). \tag{3.2}$$

Proof. By the nonnegativity of $\Lambda_{\lambda(1)}$, it suffices to show that $\inf\{\Gamma_{\lambda(1)}(\phi,x) : T\geq 0,\ \phi\in C_{[0,T]}(R_+),\ \phi_0 = x, \phi_T = y\} = H_{\lambda(1)}(x,y)$. Consider the following three cases:

Case 1: $x\leq y<\lambda(1)$. There exists a $T\geq 0$ and $\phi\in C_{[0,T]}(R_+)$ such that $\phi_0 = x$, $\phi_T = y$, and ϕ relaxes under $\lambda(1)$. By Remark 3.1, $\Gamma_{\lambda(1)}(\phi,x) = H_{\lambda(1)}(x,y) = 0$. The nonnegativity of $\Gamma_{\lambda(1)}$ implies equality (3.2).

Case 2: $\lambda(1) < x\leq y$. There exists $T\geq 0$ and a $\phi\in C_{[0,T]}(R_+)$ such that $\phi_0 = x$, $\phi_T = y$, and ϕ relaxes in reverse time under $\lambda(1)$. Remark 3.2 implies equality (3.2).

Case 3: $x\leq\lambda(1)\leq y$. Fix $\epsilon > 0$. Note that the nonnegativity of $\Lambda_{\lambda(1)}$ and Remark 3.2 imply that the left hand side of (3.2) is bounded below by $H_{\lambda(1)}(\lambda(1),y) = H_{\lambda(1)}(x,y)$. The lemma is established by constructing $T \geq 0$ and $\phi \in C_{[0,T]}(R_+)$ such that $\Gamma_{\lambda(1)}(\phi, x)$ is arbitrarily close to $H_{\lambda(1)}(x, y)$: Set $\phi_0 = x$, and let ϕ relax under $\lambda(1)$ until it reaches level $x \vee (\lambda(1) - \epsilon)$, then satisfy $\dot{\phi}_t = (y \wedge (\lambda(1) + \epsilon) - x \vee (\lambda(1) - \epsilon))/\epsilon$ until it reaches level $y \wedge (\lambda(1) + \epsilon)$, from then on relax in reverse time under $\lambda(1)$ until time T such that $\phi_T = y$. By Remarks 3.1 and 3.2, $\Gamma_{\lambda(1)}(\phi, x) = H_{\lambda(1)}(y \wedge (\lambda(1) + \epsilon), y) + f(\epsilon)$ for some $f(\epsilon) \to 0$ as $\epsilon \to 0$. The arbitrariness of $\epsilon > 0$ and continuity of $H_{\lambda(1)}$ imply inequality (3.2). $\square$

Proof of Theorem 1.1. The fact

$$\{ \sup_{0 \leq t \leq T} \tilde{X}_t^\gamma - \gamma^{-1} > \kappa \} \subset \{ \text{ Overflow time} \leq T \} \subset \{ \sup_{0 \leq t \leq T} \tilde{X}_t^\gamma + \gamma^{-1} \geq \kappa \},$$

together with Lemma 3.1 and the exponential equivalence of $(\tilde{X}^\gamma - \gamma^{-1} : \gamma > 0)$ and $(\tilde{X}^\gamma + \gamma^{-1} : \gamma > 0)$ imply that for fixed T,

$$\begin{aligned} \limsup_{\gamma \to \infty} \gamma^{-1} \log P(\text{ Overflow time} \leq T | X_0 = 0) &\leq -\inf\{\Gamma_{\lambda(1)}(\phi, 0) : \sup_{0 \leq t \leq T} \phi_t \geq \kappa\}, \\ \liminf_{\gamma \to \infty} \gamma^{-1} \log P(\text{ Overflow time} \leq T | X_0 = 0) &\geq -\inf\{\Gamma_{\lambda(1)}(\phi, 0) : \sup_{0 \leq t \leq T} \phi_t > \kappa\}. \end{aligned}$$

By Lemma 3.3 $\lim_{\gamma \to \infty} \gamma^{-1} \log P(\text{ Overflow time} \leq T | X_0 = 0)$ exists, in turn Lemma 3.4 implies that $\lim_{T \to \infty} \lim_{\gamma \to \infty} \gamma^{-1} \log P(\text{ Overflow time} \leq T | X_0 = 0) = -H_{\lambda(1)}(0, \kappa)$. This establishes the theorem. $\square$

We close this section with another result about $\Lambda_\lambda(x, y)$, which is used in Section 4.3.

Lemma 3.5 *For any positive integer d, $x \in R_+^d$, $y \in R^d$, and positive $\alpha \in R_+^d$*

$$\sum_{u=1}^d \Lambda_{\alpha(u)}(x(u), y(u)) \geq \Lambda_{\sum_{u=1}^d \alpha(u)}(\sum_{u=1}^d x(u), \sum_{u=1}^d y(u)).$$

Proof. Note that $\sigma \Lambda_\alpha(x, y) = \Lambda_{\sigma\alpha}(\sigma x, \sigma y)$ for $\sigma > 0$, and that $\Lambda_{\sum_{u=1}^d \alpha(u)}(\cdot, \cdot)$ is convex on $R_+ \times R$, as can be verified by checking that the Hessian matrix is positive semidefinite. Therefore

$$\begin{aligned} \sum_{u=1}^d \Lambda_{\alpha(u)}(x(u), y(u)) &= \sum_{u=1}^d \frac{\alpha(u)}{\sum_{w=1}^d \alpha(w)} \Lambda_{\sum_{w=1}^d \alpha(w)} \left(\frac{\sum_{w=1}^d \alpha(w)}{\alpha(u)} x(u), \frac{\sum_{w=1}^d \alpha(w)}{\alpha(u)} y(u) \right) \\ &\geq \Lambda_{\sum_{u=1}^d \alpha(u)} \left(\sum_{u=1}^d x(u), \sum_{u=1}^d y(u) \right), \end{aligned}$$

and the lemma follows. $\square$

4 The W Network

This section consists of the proofs of Theorems 1.2-1.4. A large deviations principle is established for each of the policies, and in the case of BS and LLR a variational optimization problem is solved to yield the desired conclusions.

4.1 Bernoulli Splitting

The Bernoulli Splitting (BS) policy is to assign each type 2 consumer to location 1 with probability $p = [(\lambda(3) - \lambda(1) + \lambda(2))/2\lambda(2)]_0^1$, or to location 2 with probability $1-p$, independently of the past history. Thus under the BS policy $X(1)$ and $X(2)$ are independent single-location network loads with respective demands $\gamma q(1)$ and $\gamma q(2)$.

Let $\tilde{X}^\gamma$ denote the polygonal interpolation of the normalized load process X^γ with $\tilde{X}_0^\gamma = x^\gamma$, where $\gamma x^\gamma \in Z_+^2$ and $\lim_{\gamma \to 0} x^\gamma = x_o$ for some x_o. Note that the processes $\tilde{X}^\gamma(1)$ and $\tilde{X}^\gamma(2)$ are independent and each satisfies a large deviations principle in the complete separable metric space $C_{[0,T]}(R_+)$ with a good rate function, therefore $\tilde{X}^\gamma = (\tilde{X}^\gamma(1), \tilde{X}^\gamma(2))$ satisfies a large deviations principle in the product space with the good rate function given by the sum of individual rate functions (see Theorems 4.1.18 and 4.1.11, Lemma 1.2.18, and Exercise 4.1.10 of Dembo and Zeitouni (1992)):

Lemma 4.1 *The sequence $(\tilde{X}^\gamma : \gamma > 0)$ satisfies a large deviations principle in $C_{[0,T]}(R_+^2)$ with the good rate function $\Gamma^{BS}(\cdot, x_o)$, where for each $\phi \in C_{[0,T]}(R_+^2)$ and $x \in R_+^2$, $\Gamma^{BS}(\phi, x) = \Gamma_{q(1)}(\phi(1), x(1)) + \Gamma_{q(2)}(\phi(2), x(2))$.*

The next lemma gives the solution of a variational optimization problem associated with the overflow of each location. For $\kappa \geq 0$ and $v = 1, 2$, define the set $\Omega(v, \kappa)$ by the disjoint union

$$\Omega(v, \kappa) = \bigcup_{T \geq 0} \{ \ \phi \in C_{[0,T]}(R_+^2) : \phi_0 = 0, \ \sup_{0 \leq t \leq T} \phi_t(v) \geq \kappa \ \}$$

The overflow exponents for the BS policy are each given by an infimum over a set of pairs (T, ϕ) such that $T \geq 0$ and ϕ is a function on $[0, T]$. (A similar situation arose in the proof of Theorem 1.1, in particular see Lemma 3.4). For brevity of notation, we write this as an infimum over $\Omega(v, \kappa)$, with the following understanding: Associated with each $\phi \in \Omega(v, \kappa)$, there is a value of T, and the quantity $\Gamma^{BS}(\phi, 0)$ is understood to mean: the rate function for BS with initial state 0 and time interval $[0, T]$, evaluated at ϕ.

Lemma 4.2 *For each $\kappa \geq 0$ and $v = 1, 2$,* $\inf\{\Gamma^{BS}(\phi, 0) : \phi \in \Omega(v, \kappa)\} = H_{q(v)}(0, \kappa)$.

Proof. The same proof applies for both locations, therefore only location $v = 1$ is considered. If $\phi \in \Omega(1, \kappa)$ then $\phi_\tau(1) = \kappa$ for some $\tau \geq 0$, so the definition of Γ^{BS} and Lemma 3.4 imply that $\Gamma^{BS}(\phi, 0) \geq H_{q(1)}(0, \kappa)$. Thus $\inf\{\Gamma^{BS}(\phi, 0) : \phi \in \Omega(1, \kappa)\} \geq H_{q(1)}(0, \kappa)$. The proof is completed by constructing a $\phi \in \Omega(1, \kappa)$ such that $\Gamma^{BS}(\phi, 0)$ is arbitrarily close to $H_{q(1)}(0, \kappa)$: Fix $\epsilon > 0$ and appeal to Lemma 3.4 to choose $T \geq 0$ and $\phi(1) \in C_{[0,T]}(R_+)$ such that $\phi_0(1) = 0$, $\phi_T(1) = \kappa$, and $\Gamma_{q(1)}(\phi(1), 0) \leq H_{q(1)}(0, \kappa) + \epsilon$. Let $\phi(2) \in C_{[0,T]}(R_+)$ be such that $\phi_0(2) = 0$ and $\phi(2)$ relaxes under $q(2)$. Note that $\phi = (\phi(1), \phi(2)) \in \Omega(1, \kappa)$ and $\Gamma^{BS}(\phi, 0) \leq H_{q(1)}(0, \kappa) + \epsilon$. The lemma follows by the arbitrariness of $\epsilon > 0$. □

Proof of Theorem 1.2. The proof of Theorem 1.1, with Lemmas 4.1 and 4.2 in place of Lemmas 3.1 and 3.4 respectively and an adaptation of Lemma 3.3, applied separately on each location v establishes the existence and the desired form of $F^{BS}(v, \kappa)$. The fact that $H_{q(1)}(0, \kappa) \leq H_{q(2)}(0, \kappa)$ implies $F^{BS}(\kappa) = F^{BS}(1, \kappa)$. □

4.2 Optimal Repacking

The Optimal Repacking (OR) policy is to continuously rearrange the consumers in the network so as to minimize the maximum load in the network subject to the neighborhood constraints. For each type $u \in U$, let $L_t(u)$ continue to denote the number of type u consumers in the network at time t. Theorem 1.3 is proved in this section, and then a large deviations principle for the network load under the OR policy is given.

Proof of Theorem 1.3. Note that $L(1)$, $L(3)$, and $L(1)+L(2)+L(3)$ represent the loads for three single-location networks, with respective demands $\gamma\lambda(1)$, $\gamma\lambda(3)$, and $\gamma(\lambda(1)+\lambda(2)+\lambda(3))$. If the designated capacities of these three single-location networks are $\lfloor\gamma\kappa\rfloor$, $\lfloor\gamma\kappa\rfloor$, and $2\lfloor\gamma\kappa\rfloor$, respectively, then, by Theorem 1.1, the network overflow exponents are $H_{\lambda(1)}(0,\kappa), H_{\lambda(3)}(0,\kappa)$, and $H_{\lambda(1)+\lambda(2)+\lambda(3)}(0,2\kappa)$, respectively. The overflow time for the original network under the OR strategy is the minimum of the overflow times for these three single location networks. Thus, for any $T>0$, the probability the overflow time of the original network is in $[0,T]$ is sandwiched between the maximum and the sum of the corresponding probabilities for the three single-location networks. Thus, the overflow exponent $F^{OR}(\kappa)=F^{OR}(1,\kappa)$ is given as in Theorem 1.3.

Under the OR policy, the value of the load at time t is nearly determined by $L_t=(L_t(1),L_t(2),L_t(3))$. In particular, $|X_t(v)-m(L_t,v)|<1$ for each v, where the mapping $m: R_+^3\to R_+^2$ is defined by the relations

$$\begin{aligned} m(L_t,1) &= [(L_t(1)+L_t(2)+L_t(3))/2]_{L_t(1)}^{L_t(1)+L_t(2)} \\ m(L_t,2) &= [(L_t(1)+L_t(2)+L_t(3))/2]_{L_t(3)}^{L_t(3)+L_t(2)}. \end{aligned}$$

The vector $m(L_t)=(m(L_t,1),m(L_t,2))$ is the load vector that would result at time t under the OR policy if the integer constraint on load is dropped (i.e. if the load of a consumer could be split between the two locations). Note also that $m(L_t,2)>\lfloor\kappa\gamma\rfloor$ if and only if $L_t(3)>\lfloor\kappa\gamma\rfloor$ or $(L_t(2)+L_t(3)>\lfloor\kappa\gamma\rfloor$ and $L_t(1)+L_t(2)+L_t(3)\geq 2\lfloor\kappa\gamma\rfloor)$. Assume that $\lim_{\gamma\to\infty}L_0/\gamma=l_o$ for some $l_o\in R_+^3$. Let $\tilde{L}^\gamma$ denote the polygonal interpolation of the scaled process $\gamma^{-1}L$. By Lemma 3.1, the independent load processes $\tilde{L}^\gamma(1)$, $\tilde{L}^\gamma(2)$, and $\tilde{L}^\gamma(3)$ satisfy large deviation principles in $C_{[0,T]}(R_+)$ with good rate functions $\Gamma_{\lambda(1)}(\cdot,l(1))$, $\Gamma_{\lambda(2)}(\cdot,l(2))$, and $\Gamma_{\lambda(3)}(\cdot,l(3))$ respectively. Therefore the sequence $(\tilde{L}^\gamma:\gamma>0)$ satisfies the large deviations principle in $C_{[0,T]}(R_+^3)$ with the good rate function $\hat{\Gamma}(\cdot,l)$, where for each $\xi\in C_{[0,T]}(R_+^3)$ and $l\in R_+^3$, $\hat{\Gamma}(\xi,l)=\Gamma_{\lambda(1)}(\xi(1),l(1))+\Gamma_{\lambda(2)}(\xi(2),l(2))+\Gamma_{\lambda(3)}(\xi(3),l(3))$. By the observation at the end of the previous paragraph, and an adaption of Lemma 3.3, $\lim_{\gamma\to\infty}\gamma^{-1}\log P($ Location 2 overflow time $\leq T|X_0=0)$ exists and is given by

$$\inf\{\hat{\Gamma}(\xi,0):\xi_t(3)\geq\kappa \text{ or } (\xi_t(2)+\xi_t(3)\geq\kappa \text{ and } \xi_t(1)+\xi_t(2)+\xi_t(3)\geq 2\kappa) \text{ for some } t\in[0,T]\}.$$

The expression for $F^{OR}(2,\kappa)$ given in Theorem 1.3 follows, so that Theorem 1.3 is proved. □

For completeness, we state a large deviations principle for the load process for the original network under the OR policy.

Proposition 4.1 *Define the mapping* $\mathcal{M} : C_{[0,T]}(R_+^3) \to C_{[0,T]}(R_+^2)$ *by* $\mathcal{M}(\xi)_t = m(\xi_t)$, $0 \le t \le T$, *and assume that* $\lim_{\gamma\to\infty} L_0 = l_o$ *for some* l_o. *The sequence* $(\tilde{X}^\gamma : \gamma > 0)$ *satisfies a large deviations principle in* $C_{[0,T]}(R_+^2)$ *with the good rate function* $\Gamma^{OR}(\cdot, l_o)$ *where for each* $\phi \in C_{[0,T]}(R_+^2)$ *and* $l \in R_+^3$,

$$\Gamma^{OR}(\phi, l) = \inf_{\substack{\xi \in C_{[0,T]}(R_+^3),\\ \mathcal{M}(\xi)=\phi,\ \xi_0 = l}} \{\Gamma_{\lambda(1)}(\xi(1), l(1)) + \Gamma_{\lambda(2)}(\xi(2), l(2)) + \Gamma_{\lambda(3)}(\xi(3), l(3))\}$$

with the understanding that the infimum over an empty set equals $+\infty$.

Proof. The sequences $(\tilde{X}^\gamma : \gamma > 0)$ and $(\mathcal{M}(\tilde{L}^\gamma) : \gamma > 0)$ are exponentially equivalent, so it suffices to establish the desired large deviations principle for $(\mathcal{M}(\tilde{L}^\gamma) : \gamma > 0)$ (see Theorem 4.2.13 of Dembo and Zeitouni (1992)). Continuity of the mapping $\mathcal{M}$, and the Contraction Principle (Theorem 4.2.1 of Dembo and Zeitouni (1992)) imply the proposition. □

4.3 Least Load Routing

The Least Load Routing (LLR) policy is to assign each new consumer to an admissible location that has the least load within its associated neighborhood. In the W network of Figure 1.b, we assume that when both locations have the same load the assignment decision is made in favor of location 1. The normalized load process, X^γ, is a Markov jump process which takes values in R_+^2. The process X^γ has jumps of magnitude γ^{-1} in the four directions $e_1^+ = (1,0)$, $e_2^+ = (0,1)$, $e_1^- = (-1,0)$, $e_2^- = (0,-1)$, and is generated by the pair (γ, ν) where for each $x \in R_+^2$,

$$\begin{aligned}
\nu(x, \{e_1^-\}) &= x(1), \\
\nu(x, \{e_2^-\}) &= x(2), \\
\nu(x, \{e_1^+\}) &= \begin{cases} \lambda(1) + \lambda(2) & \text{if } x(1) \le x(2) \\ \lambda(1) & \text{if } x(1) > x(2), \end{cases}
\end{aligned}$$

$$\nu(x, \{e_2^+\}) = \begin{cases} \lambda(3) & \text{if } x(1) \le x(2) \\ \lambda(3) + \lambda(2) & \text{if } x(1) > x(2). \end{cases}$$

Let $\tilde{X}^\gamma$ denote the polygonal interpolation of X^γ, with $\tilde{X}_0^\gamma = x^\gamma$, where $\gamma x^\gamma \in Z_+^2$. The following proposition establishes a large deviations principle for the network load under the LLR policy. The proof of the proposition can be found in Section 5.

Proposition 4.2 *Suppose* $\lim_{\gamma\to\infty} x^\gamma = x_o$ *for some* x_o*. The sequence* $(\tilde{X}^\gamma : \gamma > 0)$ *satisfies the large deviations principle in* $C_{[0,T]}(R_+^2)$ *with the good rate function* $\Gamma^{LLR}(\cdot, x_o)$*, where for each* $\phi \in C_{[0,T]}(R_+^2)$ *and* $x \in R_+^2$*,*

$$\Gamma^{LLR}(\phi, x) = \begin{cases} \int_0^T \Lambda(\phi_t, \dot{\phi}_t) dt & \textit{if } \phi_0 = x \textit{ and } \phi \textit{ is absolutely continuous} \\ +\infty & \textit{otherwise,} \end{cases}$$

and Λ *satisfies*

$$\Lambda(\phi_t, \dot{\phi}_t) = \begin{cases} \Lambda_{\lambda(1)}(\phi_t(1), \dot{\phi}_t(1)) + \Lambda_{\lambda(2)+\lambda(3)}(\phi_t(2), \dot{\phi}_t(2)) & \textit{if } \phi_t(1) > \phi_t(2) \\ \Lambda_{q(1)}(\phi_t(1), \dot{\phi}_t(1)) + \Lambda_{q(2)}(\phi_t(2), \dot{\phi}_t(2)) & \textit{if } \phi_t(1) = \phi_t(2) \\ \Lambda_{\lambda(1)+\lambda(2)}(\phi_t(1), \dot{\phi}_t(1)) + \Lambda_{\lambda(3)}(\phi_t(2), \dot{\phi}_t(2)) & \textit{if } \phi_t(1) < \phi_t(2). \end{cases}$$

The following three lemmas provide the solutions of the two variational optimization problems regarding the overflow of each location. In particular, Lemma 4.3 concerns location 1 and Lemma 4.5 concerns location 2. Lemmas 4.3 and 4.5 differ since we have assumed that $\lambda(1) \ge \lambda(3)$. Lemma 4.4 provides an auxiliary result that is used in the proof of Lemma 4.5.

Lemma 4.3 *For each* $\kappa \ge 0$*,*

$$\inf\{\Gamma^{LLR}(\phi, 0) : \phi \in \Omega(1, \kappa)\} = H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2(\kappa_*(1) \wedge \kappa)) + H_{\lambda(1)}(\kappa_*(1) \wedge \kappa, \kappa).$$

Proof. Given absolutely continuous $\phi \in \Omega(1, \kappa)$, let $\tau = \inf\{t \ge 0 : \phi_t(1) = \kappa\}$ and $\tau' = \sup\{t \le \tau : \phi_t(1) = \phi_t(2)\}$. By the nonnegativity of Λ,

$$\Gamma^{LLR}(\phi, 0) \ge \int_0^{\tau'} \Lambda(\phi_t, \dot{\phi}_t) dt + \int_{\tau'}^{\tau} \Lambda(\phi_t, \dot{\phi}_t) dt.$$

Lemmas 3.5 and 3.4 can be used to bound the terms on the right hand side as:

$$\begin{aligned}
\int_0^{\tau'} \Lambda(\phi_t, \dot{\phi}_t)dt &\geq \int_0^{\tau'} \Lambda_{\lambda(1)+\lambda(2)+\lambda(3)}(\phi_t(1)+\phi_t(2), \dot{\phi}_t(1)+\dot{\phi}_t(2))dt \\
&\geq H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, \phi_{\tau'}(1)+\phi_{\tau'}(2)), \\
\int_{\tau'}^{\tau} \Lambda(\phi_t, \dot{\phi}_t)dt &\geq \int_{\tau'}^{\tau} \Lambda_{\lambda(1)}(\phi_t(1), \dot{\phi}_t(1))dt \\
&\geq H_{\lambda(1)}(\phi_{\tau'}(1), \kappa).
\end{aligned}$$

This, along with the observation $\phi_{\tau'}(1) = \phi_{\tau'}(2)$ imply

$$\begin{aligned}
\inf\{\Gamma^{LLR}(\phi, 0) : \phi \in \Omega(1,\kappa)\} &\geq \inf_{0\leq s\leq\kappa} \left\{ H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2s) + H_{\lambda(1)}(s,\kappa)\right\} \\
&= H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2(\kappa_*(1)\wedge\kappa)) + H_{\lambda(1)}(\kappa_*(1)\wedge\kappa, \kappa). \quad (4.1)
\end{aligned}$$

The proof is completed by constructing a function $\phi \in \Omega(1,\kappa)$ such that $\Gamma^{LLR}(\phi, 0)$ is arbitrarily close to the right hand side of inequality (4.1). Fix $\epsilon > 0$ and consider the following two cases:

Case 1: $q(1) = q(2)$. Appeal to Lemma 3.4 to choose $T \geq 0$ and $\phi(1) \in C_{[0,T]}(R_+)$ such that $\phi_0(1) = 0$, $\phi_T(1) = \kappa_*(1)\wedge\kappa$, and $\Gamma_{q(1)}(\phi(1), 0) \leq H_{q(1)}(0, \kappa_*(1)\wedge\kappa) + \epsilon$. Set $\phi = (\phi(1), \phi(1))$. If $\kappa \leq \kappa_*(1)$, the construction is complete and $\Gamma^{LLR}(\phi, 0) \leq H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa) + 2\epsilon$. Else if $\kappa > \kappa_*(1)$, then for some small $\delta < \kappa - \kappa_*(1)$, extend ϕ further by setting $\dot{\phi}_t = (1,1)$ for $T \leq t \leq T+\delta$ (this insures that $\phi_{T+\delta} > q(1)$), and by letting $\phi(1)$ relax in reverse time under $\lambda(1)$ and $\phi(2)$ relax under $\lambda(2)+\lambda(3)$ for $T+\delta \leq t \leq T'$, where T' is such that $\phi_{T'}(1) = \kappa$. Note that $\phi_t(1) > \phi_t(2)$ for $T+\delta < t \leq T'$, hence δ can be chosen small enough so that $\Gamma^{LLR}(\phi, 0) \leq H_{\lambda(1)+\lambda(2)+\lambda(3)}(0, 2\kappa_*(1)) + H_{\lambda(1)}(\kappa_*(1), \kappa) + 3\epsilon$.

Case 2: $q(1) > q(2)$. Note that in this case Remark 1.1 implies $F^{LLR}(1,\kappa) = H_{\lambda(1)}(0,\kappa)$. Appeal to Lemma 3.4 to choose $T \geq 0$ and $\phi(1) \in C_{[0,T]}(R_+)$ so that $\phi_0(1) = 0$, $\phi_T(1) = \kappa$, and $\Gamma_{\lambda(1)}(\phi(1), 0) \leq H_{\lambda(1)}(0,\kappa) + \epsilon$. Let $\phi(2) \in C_{[0,T]}(R_+)$ be such that $\phi_0(2) = 0$ and $\phi(2)$ relaxes under $\lambda(2)+\lambda(3)$. Set $\phi = (\phi(1), \phi(2))$. Note that $\phi(1)$ can be constructed as in the proof of Lemma 3.4, so that $\phi_t(1) \geq \phi_t(2)$ for $0 \leq t \leq T$, and therefore $\Gamma^{LLR}(\phi, 0) \leq H_{\lambda(1)}(0,\kappa) + \epsilon$.

Figure 2 sketches the function ϕ constructed above. The lemma follows by the arbitrariness of $\epsilon > 0$. $\square$

Lemma 4.4 *For each $s \geq 0$ and absolutely continuous $\phi \in C_{[0,T]}(R_+^2)$ such that $\phi_0 = 0$ and $\phi_T = (s,s)$,*

$$\int_0^T \Lambda(\phi_t, \dot{\phi}_t)dt \geq H_{q(1)}(0,s) + H_{q(2)}(0,s).$$

Proof. It is convenient to use the representation

$$\int_0^T \Lambda(\phi_t, \dot{\phi}_t)dt = \int_0^T \left(\Lambda_{\rho(1,t)}(\phi_t(1), \dot{\phi}_t(1)) + \Lambda_{\rho(2,t)}(\phi_t(2), \dot{\phi}_t(2))\right) dt \tag{4.2}$$

where

$$(\rho(1,t), \rho(2,t)) = \begin{cases} (\lambda(1), \lambda(2)+\lambda(3)) & \text{if } \phi_t(1) > \phi_t(2) \\ (q(1), q(2)) & \text{if } \phi_t(1) = \phi_t(2) \\ (\lambda(1)+\lambda(2), \lambda(3)) & \text{if } \phi_t(1) < \phi_t(2). \end{cases}$$

For each $v = 1, 2$, define $\tau(v,x) = \inf\{t \geq 0 : \phi_t(v) = x\}$ for $0 \leq x \leq s$, and define $\sigma_t(v) = I_{\{\dot{\phi}_t(v)>0,\ \ \phi_t(v) \geq \phi_z(v),\ 0 \leq z \leq t\}}$ and $\phi_t^*(v) = \sup_{0 \leq z \leq t} \phi_z(v)$ for $0 \leq t \leq T$. Note that the function $\phi^*(v)$ is absolutely continuous, and

$$\left\{ \begin{array}{c} \phi_t^*(v) = \phi_t(v),\ \dot{\phi}_t^*(v) = \dot{\phi}_t(v) \\ \text{and} \ \ \tau(v, \phi_t^*(v)) = t \end{array} \right\} \text{ for almost all } t \text{ such that } \sigma_t(v) > 0. \tag{4.3}$$

Therefore if $s \geq q(v)$ then

$$\begin{aligned} \int_0^T \Lambda_{\rho(v,t)}(\phi_t(v), \dot{\phi}_t(v))dt &\geq \int_{\tau(v,q(v))}^{\tau(v,s)} \Lambda_{\rho(v,t)}(\phi_t(v), \dot{\phi}_t(v))dt \\ &\geq \int_{\tau(v,q(v))}^{\tau(v,s)} \Lambda_{\rho(v,t)}(\phi_t(v), \dot{\phi}_t(v))\sigma_t(v)dt \\ &= \int_{\tau(v,q(v))}^{\tau(v,s)} \Lambda_{\rho(v,\tau(v,\phi_t^*(v)))}(\phi_t^*(v), \dot{\phi}_t^*(v))\sigma_t(v)dt & (4.4) \\ &\geq \int_{\tau(v,q(v))}^{\tau(v,s)} \dot{\phi}_t^*(v) \log\left(\frac{\phi_t^*(v)}{\rho(v, \tau(v, \phi_t^*(v)))}\right) \sigma_t(v)dt & (4.5) \\ &= \int_{\tau(v,q(v))}^{\tau(v,s)} \dot{\phi}_t^*(v) \log\left(\frac{\phi_t^*(v)}{\rho(v, \tau(v, \phi_t^*(v)))}\right) dt & (4.6) \\ &= \int_{q(v)}^{s} \log\left(\frac{x}{\rho(v, \tau(v,x))}\right) dx. & (4.7) \end{aligned}$$

Here equality (4.4) follows by the observation (4.3), inequality (4.5) is a consequence of Remark 3.2, equality (4.6) is implied by the fact that $\dot{\phi}_t^*(v)\sigma_t(v) = \dot{\phi}_t^*(v)$ for almost all t, and equality

(4.7) follows by a change of variables. Inequality (4.7), together with representation (4.2) and the nonnegativity of $\Lambda_{\rho(1,t)}$ and $\Lambda_{\rho(2,t)}$ imply

$$\begin{aligned}\int_0^T \Lambda(\phi_t, \dot{\phi}_t)dt &\geq \int_{q(1)\wedge s}^{s} \log\left(\frac{x}{\rho(1,\tau(1,x))}\right)dx + \int_{q(2)\wedge s}^{s} \log\left(\frac{x}{\rho(2,\tau(2,x))}\right)dx \\ &= \int_{q(2)\wedge s}^{q(1)\wedge s} \log\left(\frac{x}{\rho(2,\tau(2,x))}\right)dx \\ &\qquad + \int_{q(1)\wedge s}^{s} \log\left(\frac{x}{\rho(1,\tau(1,x))}\right) + \log\left(\frac{x}{\rho(2,\tau(2,x))}\right)dx. \qquad (4.8)\end{aligned}$$

We complete the proof by obtaining appropriate lower bounds for each of the terms on the right hand side of inequality (4.8): Note that if $\tau(1,x) = \tau(2,x)$ then $(\rho(1,\tau(1,x)), \rho(2,\tau(2,x))) = (q(1), q(2))$, else if $\tau(1,x) < \tau(2,x)$ then $\rho(1,\tau(1,x)) = \lambda(1)$, and if $\tau(1,x) > \tau(2,x)$ then $\rho(2,\tau(2,x)) = \lambda(3)$. Therefore $(\rho(1,\tau(1,x)), \rho(2,\tau(2,x)))$ takes values in the set

$$\{\ (q(1),q(2)),\ (\lambda(1),q(2)),\ (\lambda(1),\lambda(2)+\lambda(3)),\ (\lambda(1),\lambda(3)),\ (q(1),\lambda(3)),\ (\lambda(1)+\lambda(2),\lambda(3))\ \},$$

and a simple calculation yields

$$\log\left(\frac{x}{\rho(1,\tau(1,x))}\right) + \log\left(\frac{x}{\rho(2,\tau(2,x))}\right) \geq \log\left(\frac{x}{q(1)}\right) + \log\left(\frac{x}{q(2)}\right) \qquad (4.9)$$

$$\log\left(\frac{x}{\rho(2,\tau(2,x))}\right) \geq \log\left(\frac{x}{\lambda(2)+\lambda(3)}\right). \qquad (4.10)$$

Since $q(1) > q(2)$ implies $\lambda(2) + \lambda(3) = q(2)$, inequalities (4.9) and (4.10), together with inequality (4.8) imply that $\int_0^T \Lambda(\phi_t, \dot{\phi}_t)dt \geq H_{q(1)}(0,s) + H_{q(2)}(0,s)$. This establishes the lemma. □

Lemma 4.5 *For each $\kappa \geq 0$,*

$$\inf\{\Gamma^{LLR}(\phi,0) : \phi \in \Omega(2,\kappa)\} = H_{q(1)}(0, \kappa_*(2)\wedge\kappa) + H_{q(2)}(0,\kappa_*(2)\wedge\kappa) + H_{\lambda(3)}(\kappa_*(2)\wedge\kappa, \kappa).$$

Proof. Given absolutely continuous $\phi \in \Omega(2,\kappa)$, let $\tau = \inf\{t \geq 0 : \phi_t(2) = \kappa\}$ and $\tau' = \sup\{t \leq \tau : \phi_t(1) = \phi_t(2)\}$. By the nonnegativity of Λ,

$$\Gamma^{LLR}(\phi,0) \geq \int_0^{\tau'} \Lambda(\phi_t,\dot{\phi}_t)dt + \int_{\tau'}^{\tau} \Lambda(\phi_t,\dot{\phi}_t)dt.$$

Lemmas 4.4 and 3.4 can be used to bound the terms on the right hand side as:

$$\begin{aligned}\int_0^{\tau'} \Lambda(\phi_t,\dot{\phi}_t)dt &\geq H_{q(1)}(0,\phi_{\tau'}(1)) + H_{q(2)}(0,\phi_{\tau'}(2)),\\ \int_{\tau'}^{\tau} \Lambda(\phi_t,\dot{\phi}_t)dt &\geq H_{\lambda(3)}(\phi_{\tau'}(2),\kappa) \wedge \Big(H_{\lambda(1)}(\phi_{\tau'}(1),\kappa) + H_{\lambda(2)+\lambda(3)}(\phi_{\tau'}(2),\kappa)\Big).\end{aligned}$$

This, along with the observation $\phi_{\tau'}(1) = \phi_{\tau'}(2)$ imply

$$\begin{aligned}\inf\{\Gamma^{LLR}(\phi,0) : \phi \in \Omega(2,\kappa)\} &\geq \inf_{0\leq s\leq\kappa} \Big\{ H_{q(1)}(0,s) + H_{q(2)}(0,s)\\ &\qquad + H_{\lambda(3)}(s,\kappa) \wedge \Big(H_{\lambda(1)}(s,\kappa) + H_{\lambda(2)+\lambda(3)}(s,\kappa)\Big)\Big\}\\ &= H_{q(1)}(0,\kappa_*(2)\wedge\kappa) + H_{q(2)}(0,\kappa_*(2)\wedge\kappa) + H_{\lambda(3)}(\kappa_*(2)\wedge\kappa,\kappa).\end{aligned}$$

The proof is completed by constructing a function $\phi \in \Omega(2,\kappa)$ such that $\Gamma^{LLR}(\phi,0)$ is arbitrarily close to the right hand side of the above inequality. Fix $0 < \epsilon < 1$ and consider the following three cases:

Case 1: $\kappa < q(2)$. Choose $T \geq 0$ and $\phi \in C_{[0,T]}(R_+^2)$ such that $\phi_0 = 0$, $\phi(1)$ and $\phi(2)$ relax respectively under $q(1)$ and $q(2)$ so that $\phi_T = \kappa(q(1)/q(2),1)$. Note that $\phi_t(1) = (q(1)/q(2))\phi_t(2)$ for $0 \leq t \leq T$, therefore $\Gamma^{LLR}(\phi,0) = 0$.

Case 2: $q(2) \leq \kappa \leq \kappa_*(2)$. Let $T > 0$ and $(\phi_t : 0 \leq t \leq T)$ be constructed as in Case 1 with $\kappa = (1-\epsilon)q(2)$. Extend ϕ by setting $\dot{\phi} = (\kappa \vee q(1),\kappa)$ for $T \leq t \leq T+\epsilon$. Note that $\phi_{T+\epsilon}$ is on the line segment with end points $(q(1),q(2))$ and $(\kappa \vee q(1),\kappa)$. If $\kappa = q(2)$ this completes the construction. Else if $\kappa > q(2)$, extend ϕ further by letting $\phi(1)$ and $\phi(2)$ relax in reverse time respectively under $q(1)$ and $q(2)$ so that $\phi_{T'} = (\kappa \vee q(1),\kappa)$ for some time $T' > T+\epsilon$. Note that in this case $(\phi_t(1)-q(1))/(\phi_t(2)-q(2)) = (\kappa\vee q(1)-q(1))/(\kappa-q(2))$ for $T+\epsilon \leq t \leq T'$, so that ϕ traces out the line segment from $\phi_{T+\epsilon}$ to $\phi_{T'}$, and thus $\phi_t(1) > \phi_t(2)$ for $0 < t < T'$. Therefore $\Gamma^{LLR}(\phi,0) = H_{q(1)}(0,\kappa) + H_{q(2)}(0,\kappa) + f(\epsilon)$ for some $f(\epsilon) \to 0$ as $\epsilon \to 0$.

Case 3: $\kappa > \kappa_*(2)$. Let $T > 0$ and $(\phi_t : 0 \leq t \leq T)$ be constructed as in Case 2 with $\kappa = \kappa_*(2)$. Note that $\kappa_*(2) > q(1)$ thus $\phi_T(1) = \phi_T(2) = \kappa_*(2)$. Extend ϕ by letting $\phi(1)$ relax under $\lambda(1)+\lambda(2)$ and $\phi(2)$ relax in reverse time under $\lambda(3)$ so that $\phi_{T'}(2) = \kappa$ at some time $T' > T$. Note that $(\phi_t(1)-q(1))(\phi_t(2)-q(2)) = (\kappa_*(2)-q(1))(\kappa_*(2)-q(2))$ and $\phi_t(2) > \phi_t(1)$ for $T < t \leq T'$, therefore $\Gamma^{LLR}(\phi,0) = H_{q(1)}(0,\kappa_*(2)) + H_{q(2)}(0,\kappa_*(2)) + H_{\lambda(3)}(\kappa_*(2),\kappa) + f(\epsilon)$ for some $f(\epsilon) \to 0$ as $\epsilon \to 0$.

Figure 3 sketches the function ϕ constructed above. The arbitrariness of $\epsilon > 0$ establishes the lemma. $\square$

Proof of Theorem 1.4. The proof of Theorem 1.1, with Lemmas 4.2 and 4.3 (4.5) in place of Lemmas 3.1 and 3.4 respectively and an adaptation of Lemma 3.3, applied on location 1 (location 2) establishes the existence and the desired form of $F^{LLR}(1,\kappa)$ $(F^{LLR}(2,\kappa))$. Since $H_{\lambda(1)}(0,\kappa) \leq H_{\lambda(3)}(0,\kappa)$ and $\kappa_*(v)$ is minimizes $H_{q(1)}(0,\kappa_*(v)\wedge\kappa) + H_{q(2)}(0,\kappa_*(v)\wedge\kappa) + H_{\lambda(2v-1)}(\kappa_*(v)\wedge\kappa,\kappa)$ for each $v = 1,2$, it follows that $F^{LLR}(\kappa) = F^{LLR}(1,\kappa) \leq F^{LLR}(2,\kappa)$. $\square$

5 Large Deviations Principle for the W Network under the LLR Policy

This section proves Proposition 4.2, the large deviations principle satisfied by the normalized load process X^γ under the LLR policy. The proof entails an application of the theory of large deviations of Markov processes with discontinuous transition mechanisms (see Alanyali and Hajek (1996), Blinovskii and Dobrushin (1994), Dupuis and Ellis (1995), Shwartz and Weiss (1995)). The transition mechanism of X^γ changes smoothly in the two open halves of R^2_+ separated by the hyperplane $\{x \in R^2 : x(1) = x(2)\}$, so that the process X^γ nearly conforms to the conditions of Theorem 2.1 of Alanyali and Hajek (1996). The theorem does not apply directly, however, because for $v = 1,2$ the log rates $\log(\nu(x,e_v^-))$ are neither bounded above (since $\nu(x,e_v^-) \to \infty$ as $x(v) \to \infty$), nor continuous and finite over R^2_+ (since $\nu(x,e_v^-) \to 0$ as $x(v) \to 0$). We therefore establish Proposition 4.2 by approximating X^γ by a sequence of auxiliary processes each of which conforms to the conditions of Theorem 2.1 of Alanyali and Hajek (1996), and adapting the techniques used in Section 12.6 of Shwartz and Weiss (1995) for the one dimensional Erlang model.

The outline the proof is as follows: Lemma 5.2 identifies the large deviations principle satisfied by each auxiliary process. Lemma 5.6 establishes the goodness of the rate function $\Gamma^{LLR}(\cdot, x_o)$. Based on a coupling of the auxiliary processes with the load process, Lemmas 5.8 and 5.9 prove the large deviations upper and lower bounds. We start with the following lemma:

Lemma 5.1 *For $x_1, x_2 \geq 0$, $\sigma_1, \sigma_2 > 0$, $y \in R$, and $\beta \in (0,1)$,*

$$\inf_{\substack{y_1 \in R,\ y_2 \in R \\ \beta y_1 + (1-\beta) y_2 = y}} \{\ \beta \Lambda_{\sigma_1}(x_1, y_1) + (1-\beta)\Lambda_{\sigma_2}(x_2, y_2)\ \} = \Lambda_{\beta\sigma_1 + (1-\beta)\sigma_2}(\beta x_1 + (1-\beta)x_2, y). \tag{5.1}$$

There exists a unique solution to the left hand side of (5.1) which satisfies

$$\frac{y_1 + \sqrt{(y_1)^2 + 4\sigma_1 x_1}}{2\sigma_1} = \frac{y_2 + \sqrt{(y_2)^2 + 4\sigma_2 x_2}}{2\sigma_2}. \tag{5.2}$$

Proof. The function $\beta\Lambda_{\sigma_1}(x_1, y_1) + (1-\beta)\Lambda_{\sigma_2}(x_2, (y - \beta y_1)/(1-\beta)) \to \infty$ as $|y_1| \to \infty$, and is strictly convex in y_1, it therefore achieves its minimum at a unique stationary point, which satisfies equality (5.2) with y_2 defined by $\beta y_1 + (1-\beta)y_2 = y$. The quantity on both sides of (5.2) is the nonnegative root of the equation $\sigma_v z^2 - y_v z - x_v = 0$ for each $v = 1, 2$. This quantity is therefore equal to the nonnegative root of the equation $(\beta\sigma_1 + (1-\beta)\sigma_2)z^2 - (\beta y_1 + (1-\beta)y_2)z - (\beta x_1 + (1-\beta)x_2) = 0$, so that

$$\frac{y_1 + \sqrt{(y_1)^2 + 4\sigma_1 x_1}}{2\sigma_1} = \frac{y_2 + \sqrt{(y_2)^2 + 4\sigma_2 x_2}}{2\sigma_2} = \frac{y + \sqrt{y^2 + 4(\beta\sigma_1 + (1-\beta)\sigma_2)(\beta x_1 + (1-\beta)x_2)}}{2(\beta\sigma_1 + (1-\beta)\sigma_2)}.$$

Equality (5.1) follows by direct substitution. □

Given $0 < \epsilon \leq 1$, let $Y^{\gamma,\epsilon}$ denote the Markov process generated by the pair (γ, ν^ϵ) where for each $x \in R^2$,

$$\nu^\epsilon(x, \{e_1^-\}) = [x(1)]_\epsilon^{1/\epsilon}, \quad \nu^\epsilon(x, \{e_1^+\}) = \begin{cases} \lambda(1) + \lambda(2) & \text{if } x(1) \leq x(2) \\ \lambda(1) & \text{if } x(1) > x(2), \end{cases}$$

$$\nu^\epsilon(x, \{e_2^-\}) = [x(2)]_\epsilon^{1/\epsilon}, \quad \nu^\epsilon(x, \{e_2^+\}) = \begin{cases} \lambda(3) & \text{if } x(1) \leq x(2) \\ \lambda(3) + \lambda(2) & \text{if } x(1) > x(2), \end{cases}$$

and let $\tilde{Y}^{\gamma,\epsilon}$ denote the polygonal interpolation of $Y^{\gamma,\epsilon}$. Suppose $\tilde{Y}_0^{\gamma,\epsilon} = x^\gamma$, where $(x^\gamma : \gamma > 0)$ is the sequence with $\lim_{\gamma\to\infty} x^\gamma = x_o$ associated with Proposition 4.2.

Lemma 5.2 *For each $0 < \epsilon \leq 1$, the sequence $(\tilde{Y}^{\gamma,\epsilon} : \gamma > 0)$ satisfies the large deviations principle in $C_{[0,T]}(R^2)$ with the good rate function $\Gamma^\epsilon(\cdot, x_o)$, where for each $\phi \in C_{[0,T]}(R^2)$ and*

$x \in R^2$,

$$\Gamma^\epsilon(\phi, x) = \begin{cases} \int_0^T \Lambda^\epsilon(\phi_t, \dot{\phi}_t)dt & \text{if } \phi_0 = x \text{ and } \phi \text{ is absolutely continuous} \\ +\infty & \text{otherwise,} \end{cases}$$

and $\Lambda^\epsilon(\phi_t, \dot{\phi}_t) = \Lambda([\phi_t]_\epsilon^{1/\epsilon}, \dot{\phi}_t)$.

Proof. Let $A^o = \{x \in R^2 : x(1) = x(2)\}$, $A^+ = \{x \in R^2 : x(1) < x(2)\}$, $A^- = \{x \in R^2 : x(1) > x(2)\}$, and let the rate-measure fields $\nu^{+,\epsilon}$ and $\nu^{-,\epsilon}$ be defined as

$$\nu^{+,\epsilon}(x) = \begin{cases} \nu^\epsilon(x) & \text{if } x \in \overline{A^+} \\ \nu^\epsilon(x(1), x(1)) & \text{if } x \in A^- \end{cases} \qquad \nu^{-,\epsilon}(x) = \begin{cases} \nu^\epsilon(x) & \text{if } x \in A^- \\ \lim_{\delta \searrow 0} \nu^\epsilon(x(2) + \delta, x(2) - \delta) & \text{if } x \in \overline{A^+}. \end{cases}$$

Note that $\nu^{+,\epsilon}$, $\nu^{-,\epsilon}$, $Y^{\gamma,\epsilon}$ satisfy the conditions of Theorem 2.1 of Alanyali and Hajek (1996), therefore the sequence $(\tilde{Y}^{\gamma,\epsilon} : \gamma > 0)$ satisfies a large deviations principle with the good rate function $\Gamma^\epsilon(\cdot, x_o)$, where for $x, y \in R^2$

$$\begin{aligned} \Lambda^\epsilon(x,y) &= I_{\{\phi_t \in A^+\}} \Lambda^{+,\epsilon}(x,y) + I_{\{\phi_t \in A^o\}} \Lambda^{o,\epsilon}(x,y) + I_{\{\phi_t \in A^-\}} \Lambda^{-,\epsilon}(x,y), \\ \Lambda^{\pm,\epsilon}(x,y) &= \sum_{v=1}^{2} \sup_{\zeta \in R} \left\{ \zeta y(v) - \left((e^\zeta - 1)\nu^{\pm,\epsilon}(x, e_v^+) + (e^{-\zeta} - 1)\nu^{\pm,\epsilon}(x, e_v^-) \right) \right\}, \qquad (5.3) \\ \Lambda^{o,\epsilon}(x,y) &= \inf_{\substack{0 \le \beta \le 1,\ y^+ \in \overline{A^-}, y^- \in \overline{A^+} \\ \beta y^+ + (1-\beta) y^- = y}} \{ \beta \Lambda^{+,\epsilon}(x, y^+) + (1-\beta)\Lambda^{-,\epsilon}(x, y^-) \}. \end{aligned}$$

Applying Exercise 7.24 of Shwartz and Weiss (1995) on each term on the right hand side of equation (5.3) yields that

$$\begin{aligned} \Lambda^{+,\epsilon}(x,y) &= \Lambda_{\lambda(1)+\lambda(2)}([x(1)]_\epsilon^{1/\epsilon}, y(1)) + \Lambda_{\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y(2)) \qquad \text{for } x \in \overline{A^+}, \\ \Lambda^{-,\epsilon}(x,y) &= \Lambda_{\lambda(1)}([x(1)]_\epsilon^{1/\epsilon}, y(1)) + \Lambda_{\lambda(2)+\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y(2)) \qquad \text{for } x \in \overline{A^-}. \end{aligned}$$

To complete the proof of the lemma, it remains to evaluate $\Lambda^{o,\epsilon}(x,y)$ for $x \in A^o$. Note that for absolutely continuous ϕ, $\dot{\phi}_t \in A^o$ for almost all t such that $\phi_t \in A^o$, therefore it suffices to consider the case when $y \in A^o$. Fix $x, y \in A^o$. For any $y^+, y^- \in R^2$ and $\beta \in [0,1]$ such that $\beta y^+ + (1-\beta)y^- = y$,

$$\beta \Lambda^{+,\epsilon}(x, y^+) \quad + \quad (1-\beta)\Lambda^{-,\epsilon}(x, y^-)$$

$$
\begin{aligned}
&= \beta\Lambda_{\lambda(1)+\lambda(2)}([x(1)]_\epsilon^{1/\epsilon}, y^+(1)) + \beta\Lambda_{\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y^+(2)) \\
&\quad + (1-\beta)\Lambda_{\lambda(1)}([x(1)]_\epsilon^{1/\epsilon}, y^-(1)) + (1-\beta)\Lambda_{\lambda(2)+\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y^-(2)) \quad (5.4)\\
&\geq \Lambda_{\lambda(1)+\beta\lambda(2)}([x(1)]_\epsilon^{1/\epsilon}, y(1)) + \Lambda_{(1-\beta)\lambda(2)+\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y(2)) \quad (5.5)\\
&\geq \inf_{0\leq\beta'\leq 1}\{\ \Lambda_{\lambda(1)+\beta'\lambda(2)}([x(1)]_\epsilon^{1/\epsilon}, y(1)) + \Lambda_{(1-\beta')\lambda(2)+\lambda(3)}([x(2)]_\epsilon^{1/\epsilon}, y(2))\ \} \quad (5.6)\\
&= \Lambda_{q(1)}([x(1)]_\epsilon^{1/\epsilon}, y(1)) + \Lambda_{q(2)}([x(2)]_\epsilon^{1/\epsilon}, y(2)). \quad (5.7)
\end{aligned}
$$

In the above argument inequality (5.5) follows by applying Lemma 5.1 separately on the first and third and the second and fourth terms on the right hand side of equality (5.4). Since $x(1) = x(2)$ and $y(1) = y(2)$, each of the terms of the right hand side of inequality (5.5) is the same convex function evaluated at $\lambda(1) + \beta\lambda(2)$ and $(1-\beta)\lambda(2) + \lambda(3)$ respectively. Equality (5.7) follows by straightforward minimization.

We next identify $\Lambda^{o,\epsilon}(x, y)$ with the right hand side of inequality (5.7) by establishing the existence of $\beta \in [0, 1]$, $y^+ \in \overline{A^-}$, and $y^- \in \overline{A^+}$ such that $\beta y^+ + (1-\beta)y^- = y$ and both inequalities (5.5) and (5.6) are satisfied with equality. Inequality (5.6) is satisfied with equality if $\beta = [(\lambda(3) - \lambda(1) + \lambda(2))/2\lambda(2)]_0^1$. If $\beta = 0$ ($\beta = 1$) then take $y^- = y$ ($y^+ = y$). Otherwise by Lemma 5.1 there exist $y^+, y^- \in R^2$ such that $\beta y^+ + (1-\beta)y^- = y$, inequality (5.5) is satisfied with equality, and the following two conditions hold:

$$
\frac{y^-(1) + \sqrt{(y^-(1))^2 + 4\lambda(1)[x(1)]_\epsilon^{1/\epsilon}}}{2\lambda(1)} = \frac{y^+(1) + \sqrt{(y^+(1))^2 + 4(\lambda(1)+\lambda(2))[x(1)]_\epsilon^{1/\epsilon}}}{2(\lambda(1)+\lambda(2))}, \quad (5.8)
$$

$$
\frac{y^+(2) + \sqrt{(y^+(2))^2 + 4\lambda(3)[x(2)]_\epsilon^{1/\epsilon}}}{2\lambda(3)} = \frac{y^-(2) + \sqrt{(y^-(2))^2 + 4(\lambda(2)+\lambda(3))[x(2)]_\epsilon^{1/\epsilon}}}{2(\lambda(2)+\lambda(3))}. \quad (5.9)
$$

The left hand side of equality (5.8) is increasing in $y^-(1)$ and decreasing in $\lambda(1)$, therefore (5.8) and (5.9) respectively imply that $y^+(1) \geq y^-(1)$ and $y^-(2) \geq y^+(2)$. This, together with the assumption that $\beta y^+ + (1-\beta)y^- \in A^o$ imply that $y^- \in \overline{A^+}$ and $y^+ \in \overline{A^-}$. The proof of the lemma is complete. □

For $x \in R_+$ and $\epsilon = 0$ set $[x]_\epsilon^{1/\epsilon} = x$, so that $\Gamma = \Gamma^\epsilon|_{\epsilon=0}$.

Lemma 5.3 *Let S be a finite set of positive numbers. For each $l \geq 0$ there exists an $M > 0$*

such that for any absolutely continuous $\phi \in C_{[0,T]}(R_+)$ *and* $0 \leq \epsilon \leq 1$,

$$\int_0^T \inf_{\sigma \in S} \Lambda_\sigma([\phi_t]_\epsilon^{1/\epsilon}, \dot{\phi}_t) dt \leq l \Longrightarrow \sup_{0 \leq t \leq T} (\phi_t - \phi_0) \leq M.$$

Proof. Examination of $\partial \Lambda_\sigma(x,y)/\partial x$ yields that $\Lambda_\sigma(x,y)$ is increasing in x whenever $y > \sigma$, thus $\lim_{y \to \infty} \Lambda_\sigma(x,y)/y = \infty$ uniformly in $x \in R_+$ and $\sigma \in S$. Given $l \geq 0$, choose a constant $B(l)$ large enough so that $\inf_{\sigma \in S, x \in R_+} \Lambda_\sigma(x,y)/y \geq l$ whenever $y \geq B(l)$. For absolutely continuous $\phi \in C_{[0,T]}(R_+)$ and $0 \leq \tau \leq T$,

$$\begin{aligned}
\int_0^T \inf_{\sigma \in S} \Lambda_\sigma([\phi_t]_\epsilon^{1/\epsilon}, \dot{\phi}_t) dt &\geq \int_{\{t \in [0,\tau] : \dot{\phi}_t \geq B(l)\}} \inf_{\sigma \in S} \frac{\Lambda_\sigma([\phi_t]_\epsilon^{1/\epsilon}, \dot{\phi}_t)}{\dot{\phi}_t} \dot{\phi}_t dt \\
&\geq \int_{\{t \in [0,\tau] : \dot{\phi}_t \geq B(l)\}} l \dot{\phi}_t dt \\
&\geq \int_0^\tau l(\dot{\phi}_t - B(l)) dt \\
&\geq l(\phi_\tau - \phi_0 - B(l)T).
\end{aligned}$$

Choosing $M = B(l)T + 1$ establishes the lemma. □

Lemma 5.4 *(Relative Compactness) For each* $l \geq 0$, *the collection* $C(l) = \cup_{0 \leq \epsilon \leq 1} \{\phi : \Gamma^\epsilon(\phi, x_0) \leq l\}$ *is relatively compact in* $C_{[0,T]}(R_+^2)$.

Proof. If $\phi \in C(l)$ then ϕ is absolutely continuous, $\phi_0 = x_0$, and for some $0 \leq \epsilon \leq 1$

$$\begin{aligned}
l &\geq \int_0^T \Lambda^\epsilon(\phi_t, \dot{\phi}_t) dt \\
&\geq \int_0^T \inf_{\sigma \in S} \Lambda_\sigma([\phi_t(1)]_\epsilon^{1/\epsilon}, \dot{\phi}_t(1)) dt + \int_0^T \inf_{\sigma \in S} \Lambda_\sigma([\phi_t(2)]_\epsilon^{1/\epsilon}, \dot{\phi}_t(2)) dt, \qquad (5.10)
\end{aligned}$$

where $S = \{\lambda(1), q(1), \lambda(1) + \lambda(2), \lambda(3), q(2), \lambda(2) + \lambda(3)\}$. Lemma 5.3, applied separately on the terms of the right hand side of (5.10), implies the existence of a finite $M > 1$ such that $\sup_{0 \leq t \leq T} |\phi_t| \leq M$ for all $\phi \in C(l)$.

Fix $\delta > 0$. Choose a constant $B(\delta)$ large enough so that $\inf_{0 \leq x \leq M, \sigma \in S} \Lambda_\sigma(x,y)/|y| > 2l/\delta$ whenever $|y| > B(\delta)$. Let $((s_j, t_j) : j = 1, \cdots, J)$ be a finite collection of nonoverlapping intervals in $[0,T]$, and set $D = \cup_j (s_j, t_j)$. Given $\phi \in C(l)$, let $0 \le \epsilon \le 1$ be such that $\Gamma^\epsilon(\phi, x_0) \le l$. Then

$$
\begin{aligned}
\sum_{j=1}^{J} |\phi_{t_j} - \phi_{s_j}| &\le \int_D |\dot{\phi}_t| dt \\
&= \int_{D \cap \{t : |\dot{\phi}_t| > B(\delta)\}} \frac{|\dot{\phi}_t|}{\Lambda^\epsilon(\phi_t, \dot{\phi}_t)} \Lambda^\epsilon(\phi_t, \dot{\phi}_t) dt + \int_{D \cap \{t : |\dot{\phi}_t| \le B(\delta)\}} |\dot{\phi}_t| dt \\
&\le \frac{\delta}{2l} \int_{D \cap \{t : |\dot{\phi}_t| > B(\delta)\}} \Lambda^\epsilon(\phi_t, \dot{\phi}_t) dt + B(\delta) \sum_{j=1}^{J} |t_j - s_j| \\
&\le \frac{\delta}{2} + B(\delta) \sum_{j=1}^{J} |t_j - s_j|.
\end{aligned}
$$

Thus $\sum_{j=1}^{J} |\phi_{t_j} - \phi_{s_j}| \le \delta$ whenever $\sum_{j=1}^{J} |t_j - s_j| \le \delta / 2B(\delta)$, uniformly for all $\phi \in C(l)$. The Arzela-Ascoli Theorem implies the relative compactness of $C(l)$. □

Lemma 5.5 *(Lower Semicontinuity) The function $\Gamma^{LLR}(\cdot, x_0)$ is lower semicontinuous.*

Proof. Let $(\phi^m : m \ge 1)$ be a sequence such that $\phi^m \to \phi$ in $C_{[0,T]}(R_+^2)$. To prove the lemma, it suffices to show that $\Gamma^{LLR}(\phi, x_0) \le \liminf_{m \to \infty} \Gamma^{LLR}(\phi^m, x_0)$. Assume, without loss of generality, the existence of $l, k \ge 0$ such that $\Gamma^{LLR}(\phi^m, x_0) \le l$ for all $m \ge k$. The proof of Lemma 5.4 shows that the sequence $(\phi^m : m \ge k)$ is uniformly absolutely continuous, therefore ϕ is absolutely continuous, $\phi_0 = x_0$, and by the explanations indicated in parentheses,

$$
\begin{aligned}
\Gamma^{LLR}(\phi, x_0) &= \textstyle\int_0^T \Lambda(\phi_t, \dot{\phi}_t) dt && \text{(Definition of } \Gamma^{LLR}) \\
&\le \liminf_{\epsilon \searrow 0} \textstyle\int_0^T \Lambda^\epsilon(\phi_t, \dot{\phi}_t) dt && \text{(Fatou's Lemma)} \\
&\le \liminf_{\epsilon \searrow 0} \liminf_{m \to \infty} \textstyle\int_0^T \Lambda^\epsilon(\phi_t^m, \dot{\phi}_t^m) dt && \text{(L.s.c. property of } \Gamma^\epsilon) \\
&\le \liminf_{\epsilon \searrow 0} \liminf_{m \to \infty} \left(\textstyle\int_0^T \Lambda(\phi_t^m \wedge \frac{1}{\epsilon}, \dot{\phi}_t^m) dt + 2\epsilon T \right) && \text{(Lemma 3.2)} \\
&= \liminf_{m \to \infty} \Gamma^{LLR}(\phi^m, x_0). && (\textstyle\sup_{0 \le t \le T} |\phi_t^m| \text{ is bounded})
\end{aligned}
$$

This establishes the lemma. □

Lemma 5.6 *(Goodness) $\Gamma^{LLR}(\cdot, x_o)$ is a good rate function.*

Proof. Note that for each $l \geq 0$ the level set $\{\phi : \Gamma^{LLR}(\phi, x_0) \leq l\}$ is contained in $C(l)$. Lemmas 5.4 and 5.5 imply respectively the relative compactness and closedness, and therefore the compactness, of the level set. $\square$

Lemma 5.7 *For closed* $F \subset C_{[0,T]}(R_+^2)$,

$$\inf_{\phi \in F} \Gamma^{LLR}(\phi, x_0) \leq \limsup_{\epsilon \searrow 0} \inf_{\phi \in F} \Gamma^{\epsilon}(\phi, x_0).$$

Proof. Without loss of generality, we may assume the existence of an $l \geq 0$ such that $\inf_{\phi \in F} \Gamma^{\epsilon}(\phi, x_0) \leq l$ for each $0 < \epsilon \leq 1$. For each such ϵ, appeal to the goodness of the rate function $\Gamma^{\epsilon}(\cdot, x_0)$ to choose a $\phi^{\epsilon} \in F$ such that $\Gamma^{\epsilon}(\phi^{\epsilon}, x_0) = \inf_{\phi \in F} \Gamma^{\epsilon}(\phi, x_0)$. By Lemma 5.4 the collection $(\phi^{\epsilon} : 0 < \epsilon \leq 1)$ is relatively compact, hence there exists a sequence $\epsilon_n \to 0$ and $\tilde{\phi} \in F$ such that $\phi^{\epsilon_n} \to \tilde{\phi}$. Define

$$\xi_t^{\epsilon_n} = \begin{cases} x_0 + (t, t) & 0 \leq t \leq \epsilon_n \\ (\epsilon_n, \epsilon_n) + \phi_{t-\epsilon_n}^{\epsilon_n} & \epsilon_n \leq t \leq T. \end{cases}$$

Note that $\xi^{\epsilon_n} \to \tilde{\phi}$, and

$$\begin{aligned} \int_0^T \Lambda(\xi_t^{\epsilon_n}, \dot{\xi}_t^{\epsilon_n}) dt &= \int_0^{\epsilon_n} \Lambda(\xi_t^{\epsilon_n}, (1,1)) dt + \int_{\epsilon_n}^T \Lambda^{\epsilon_n}((\epsilon_n, \epsilon_n) + \phi_{t-\epsilon_n}^{\epsilon_n}, \dot{\phi}_{t-\epsilon_n}^{\epsilon_n}) dt \\ &\leq \int_0^{\epsilon_n} (\Lambda(x_0, (1,1)) + 2\epsilon_n) dt + \int_0^T (\Lambda^{\epsilon_n}(\phi_t^{\epsilon_n}, \dot{\phi}_t^{\epsilon_n}) + 2\epsilon_n) dt, \end{aligned} \tag{5.11}$$

where the first step follows by the definition of Λ^{ϵ_n} and the construction of ξ^{ϵ_n}, and the second step follows by Lemma 3.2 and the nonnegativity of Λ^{ϵ_n}. Therefore by the explanations indicated in parentheses,

$$\begin{aligned} \inf_{\phi \in F} \Gamma^{LLR}(\phi, x_0) &\leq \Gamma^{LLR}(\tilde{\phi}, x_0) && (\tilde{\phi} \in F) \\ &\leq \liminf_{n \to \infty} \Gamma^{LLR}(\xi^{\epsilon_n}, x_0) && \text{(Lemma 5.5)} \\ &\leq \liminf_{n \to \infty} \Gamma^{\epsilon_n}(\phi^{\epsilon_n}, x_0) && \text{(Inequality (5.11))} \\ &\leq \limsup_{\epsilon \searrow 0} \inf_{\phi \in F} \Gamma^{\epsilon}(\phi, x_0) && (\text{Definition of } \phi^{\epsilon}), \end{aligned}$$

and the lemma is established. $\square$

For each $\gamma > 0$ construct X^γ on an appropriately extended probability space, and for each $0 < \epsilon \leq 1$ construct the process $Y^{\gamma,\epsilon}$ on the same space as follows: Let $Z(1)$ and $Z(2)$ be mutually independent Poisson processes which are also independent of X^γ, and each having rate $\gamma\epsilon$. Set $Y_0^{\gamma,\epsilon} \equiv X_0^\gamma$. Let $\tau = \inf\{t \geq 0 : X_t^\gamma \notin [0, 1/\epsilon]^2 \text{ or } Z_t \neq 0\}$. By the independence of Z and X^γ, with probability one either X^γ or Z jumps at time τ, but not both. At every time $t \leq \tau$ such that X^γ jumps, $Y^{\gamma,\epsilon}$ takes the same jump. In addition, for $v = 1, 2$, if $Z(v)$ jumps at time τ and $X_\tau^\gamma(v) \leq \epsilon$, then $Y^{\gamma,\epsilon}(v)$ jumps down at time τ by γ^{-1} with probability $(\epsilon - X_t^\gamma(v))/\epsilon$. After time τ the construction is done so that $Y^{\gamma,\epsilon}$ is generated by the specified pair (γ, ν^ϵ). (Note that Z is not used in the construction after it first jumps.)

Let $\tilde{X}^\gamma$ and $\tilde{Y}^{\gamma,\epsilon}$ denote the polygonal interpolations of X^γ and $Y^{\gamma,\epsilon}$ respectively.

Lemma 5.8 *(Upper Bound) For any closed $F \subset C_{[0,T]}(R_+^2)$,*

$$\limsup_{\gamma\to\infty} \gamma^{-1} \log P\left((\tilde{X}_t^\gamma : 0 \leq t \leq T) \in F\right) \leq -\inf_{\phi\in F} \Gamma^{LLR}(\phi, x_0).$$

Proof. Note that for each $\gamma > 0$ and $0 < \epsilon \leq 1$,

$$\begin{aligned}
P\left((\tilde{Y}_t^{\gamma,\epsilon} : 0 \leq t \leq T) \in F\right) &\geq P\left((\tilde{Y}_t^{\gamma,\epsilon} : 0 \leq t \leq T) \in F,\ \sup_{0\leq t\leq T} |\tilde{Y}_t^{\gamma,\epsilon}| < \frac{1}{\epsilon},\ Z_T = 0\right) \\
&= P\left((\tilde{X}_t^\gamma : 0 \leq t \leq T) \in F,\ \sup_{0\leq t\leq T} |\tilde{X}_t^\gamma| < \frac{1}{\epsilon},\ Z_T = 0\right) \\
&= P\left((\tilde{X}_t^\gamma : 0 \leq t \leq T) \in F,\ \sup_{0\leq t\leq T} |\tilde{X}_t^\gamma| < \frac{1}{\epsilon}\right) P(Z_T = 0) \\
&\geq \left(P\left((\tilde{X}_t^\gamma : 0 \leq t \leq T) \in F\right) - P(\sup_{0\leq t\leq T} |\tilde{X}_t^\gamma| \geq \frac{1}{\epsilon})\right) e^{-2\gamma\epsilon T},
\end{aligned}$$

where the second step follows by the construction of the processes X^γ and $Y^{\gamma,\epsilon}$, and the third step follows by the independence of X^γ and Z. In view of the above inequality,

$$\begin{aligned}
&\limsup_{\gamma\to\infty} \gamma^{-1} \log P\left((\tilde{X}_t^\gamma : 0 \leq t \leq T) \in F\right) \leq \\
&\left(\limsup_{\gamma\to\infty} \gamma^{-1} \log P\left((\tilde{Y}_t^{\gamma,\epsilon} : 0 \leq t \leq T) \in F\right) + 2\epsilon T\right) \bigvee \limsup_{\gamma\to\infty} \gamma^{-1} \log P(\sup_{0\leq t\leq T} |\tilde{X}_t^\gamma| \geq \frac{1}{\epsilon}) \qquad (5.12)
\end{aligned}$$

for each $0 < \epsilon \leq 1$. Note that $\sup_{0\leq t\leq T}|X_t|$ is stochastically dominated by X_0 plus a Poisson random variable of mean $\gamma(\lambda(1)+\lambda(2)+\lambda(3))T$, hence the second term in the right hand side of inequality (5.12) is less than any given negative number for small enough ϵ. Therefore

$$\begin{aligned}\limsup_{\gamma\to\infty}\gamma^{-1}\log P\left((\tilde{X}_t^\gamma : 0\leq t\leq T)\in F\right) &\leq \liminf_{\epsilon\searrow 0}\limsup_{\gamma\to\infty}\gamma^{-1}\log P\left((\tilde{Y}_t^{\gamma,\epsilon} : 0\leq t\leq T)\in F\right)\\ &\leq -\limsup_{\epsilon\searrow 0}\inf_{\phi\in F}\Gamma^\epsilon(\phi,x_0)\\ &\leq -\inf_{\phi\in F}\Gamma^{LLR}(\phi,x_0),\end{aligned}$$

where the second step is a consequence of Lemma 5.2, and the third step follows by Lemma 5.7. This establishes the lemma. □

Lemma 5.9 *(Lower Bound) For any open $G\subset C_{[0,T]}(R_+^2)$,*

$$\liminf_{\gamma\to\infty}\gamma^{-1}\log P\left((\tilde{X}_t^\gamma : 0\leq t\leq T)\in G\right)\geq -\inf_{\phi\in G}\Gamma^{LLR}(\phi,x_0).$$

Proof. Fix $\epsilon > 0$ and $\phi\in G$. Without loss of generality, we may assume that ϕ is absolutely continuous and $\phi_0 = x_0$. Let $\delta > 0$ be small enough such that the open ball of radius 6δ around ϕ, $B(\phi, 6\delta)$, is contained in G. By Lemma 3.3 of Alanyali and Hajek (1997), as $\gamma\to\infty$, the process $(X_t^\gamma : 0\leq t\leq T)$ converges weakly (relative to the Skorohod topology) to a Lipschitz continuous function $(x_t : 0\leq t\leq T)$ that satisfies $x_t(1)\wedge x_t(2) > 0$ for $t > 0$. Since the limit is continuous and deterministic, it follows that $\sup\{|\tilde{X}^\gamma(t) - x_t| : 0\leq t\leq T\}$ converges to zero in probability (see Theorem 10.2 of Ethier and Kurtz (1986)).

Let $d = \sup_{0\leq t\leq T}|\dot{x}_t|$, and choose a positive $\sigma < \delta/d$ such that

$$|\phi_t - \phi_s| < \delta \quad\text{and}\quad |x_t - x_s| < \delta \qquad \text{whenever } |t-s|\leq\sigma.$$

Construct ξ as

$$\xi_0 = x_0, \qquad \dot{\xi}_t = \begin{cases}\dot{x}_t & 0\leq t\leq\sigma\\ (2d,2d) & \sigma\leq t\leq 2\sigma\\ \dot{\phi}_{t-2\sigma} & 2\sigma\leq t\leq T.\end{cases}$$

It can be verified easily that $|\xi_t - \phi_t| < 5\delta$ for $0 \le t \le T$, and $\xi_t(1) \wedge \xi_t(2) \ge x_\sigma(1) \wedge x_\sigma(2) > 0$ for $\sigma \le t \le T$. Choose positive $\eta < (x_\sigma(1) \wedge x_\sigma(2) \wedge \delta)/2$ small enough, and choose δ and σ smaller if necessary, so that

$$\begin{aligned}\int_\sigma^T \Lambda^\eta(\xi_t, \dot{\xi}_t)dt &\le \int_\sigma^{2\sigma} (\Lambda(x_\sigma, (2s, 2s)) + 4\delta + 2\eta)dt + \int_{2\sigma}^T (\Lambda(\phi_{t-2\sigma}, \dot{\phi}_{t-2\sigma}) + 6\delta + 2\eta)dt \\ &\le \int_0^T \Lambda(\phi_t, \dot{\phi}_t)dt + \frac{\epsilon}{2},\end{aligned}$$

where the first inequality follows by Lemma 3.2 and the fact that for $v = 1, 2,\ x_\sigma(v) \le \xi_t(v) \le x_\sigma(v) + 2\delta$ for $\sigma \le t \le 2\sigma$ and $\phi_{t-2\sigma}(v) \le \xi_t(v) \le \phi_{t-2\sigma}(v) + 3\delta$ for $2\sigma \le t \le T$. Finally, appeal to the time-homogeneous Markov property of $Y^{\gamma,\eta}$ and Lemma 5.2 together with Remark 2.1 of Alanyali and Hajek (1996) to choose $\rho < \eta$ small enough so that for large enough γ,

$$\inf_{|x-\xi_\sigma|<\rho} P(\sup_{\sigma \le t \le T} |\tilde{Y}_t^{\gamma,\eta} - \xi_t| < \eta \mid \tilde{Y}_\sigma^{\gamma,\eta} = x) \ge \exp\left(-\gamma(\int_\sigma^T \Lambda^\eta(\xi_t, \dot{\xi}_t)dt + \frac{\epsilon}{2})\right).$$

For large enough γ,

$$\begin{aligned}P\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in G\right) &\ge P\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in B(\phi, 6\delta)\right) \\ &\ge P\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in B(\xi, \eta)\right) \qquad (5.13) \\ &\ge P(\sup_{0 \le t \le \sigma} |\tilde{X}_t^\gamma - \xi_t| < \rho) \\ &\qquad \inf_{|x-\xi_\sigma|<\rho} P(\sup_{\sigma \le t \le T} |\tilde{X}_t^\gamma - \xi_t| < \eta \mid \tilde{X}_\sigma^\gamma = x) \qquad (5.14) \\ &= P(\sup_{0 \le t \le \sigma} |\tilde{X}_t^\gamma - x_t| < \rho) \\ &\qquad \inf_{|x-\xi_\sigma|<\rho} P(\sup_{\sigma \le t \le T} |\tilde{Y}_t^{\gamma,\eta} - \xi_t| < \eta \mid \tilde{Y}_\sigma^{\gamma,\eta} = x) \qquad (5.15) \\ &\ge \frac{1}{2}\exp\left(-\gamma(\int_\sigma^T \Lambda^\eta(\xi_t, \dot{\xi}_t)dt + \frac{\epsilon}{2})\right) \\ &\ge \frac{1}{2}\exp\left(-\gamma(\int_0^T \Lambda(\phi_t, \dot{\phi}_t)dt + \epsilon)\right).\end{aligned}$$

In the above argument inequality (5.13) follows by the fact that $B(\xi, \eta) \subset B(\xi, \delta) \subset B(\phi, 6\delta)$, inequality (5.14) is a consequence of the choice of ρ and the Markov property of X^γ, and equality (5.15) is implied by the choice of η and the construction of $Y^{\gamma,\eta}$. The arbitrariness of $\epsilon > 0$ implies that

$$\liminf_{\gamma\to\infty} \gamma^{-1} \log P\left((\tilde{X}_t^\gamma : 0 \le t \le T) \in G\right) \ge -\Gamma^{LLR}(\phi, x_0),$$

and the arbitrariness of $\phi \in G$ establishes the lemma. $\square$

References

Alanyali, M., and Hajek, B. (1997). Analysis of simple algorithms for dynamic load balancing. *Mathematics of Operations Research.* To appear.

Alanyali, M., and Hajek, B. (1996). On large deviations of Markov processes with discontinuous statistics. Submitted to *Annals of Applied Probability.*

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

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.

Dembo, A., and Zeitouni, O. (1992). Large deviations techniques and applications. Jones and Bartlett, Boston.

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**(8) 2689–2751.

Ethier, S.N. and Kurtz, T.G. (1986). Markov Processes. Wiley, New York.

Gibbens, R., Kelly, F.P., and Turner S. (1993). Dynamic routing in multiparented networks. *IEEE/ACM Transactions on Networking.* **2** 261-270.

Hajek, B. (1990). Performance of global load balancing by local adjustment. *IEEE Transactions on Information Theory.* **36**(6) 1398–1414.

Shwartz, A., and Weiss, A. (1995). Large deviations for performance analysis, queues, communication and computing. Chapman & Hall, London.

MURAT ALANYALI
BELL LABORATORIES
LUCENT TECHNOLOGIES
101 CRAWFORDS CORNER ROAD
HOLMDEL, NJ 07733-3030
Email: murat@dnrc.bell-labs.com

BRUCE HAJEK
DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING AND THE
COORDINATED SCIENCE LABORATORY
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS 61801
Email: b-hajek@uiuc.edu