Logarithmic asymptotics for a single-server processing distinguishable sources

Ken Duffy and David Malone,
Hamilton Institute,
National University of Ireland, Maynooth.
E-mail:{ken.duffy, david.malone}@nuim.ie

17rd April 2007; revised 5th October 2007

Abstract

We consider a single-server first-in-first-out queue fed by a finite number of distinct sources of jobs. For a large class of short-range dependent and light-tailed distributed job processes, using functional large deviation techniques we prove a large deviation principle and logarithmic asymptotics for the joint waiting time and queue lengths distribution. We identify the paths that are most likely to lead to the rare events of large waiting times and long queue lengths. A number of examples are presented to illustrate salient features of the results.

1 Introduction

Consider a single-server processing system with infinite waiting space that employs the first-in-first-out (FIFO) processing policy. The server processes at a fixed rate C and there are a finite number $d \in \mathbb{N}$ of, possibly dependent, distinguishable sources of jobs. Each source of jobs is described by their own stationary job size and inter-arrival time sequences. We assume that the server has been running for an infinite period of time and select an arbitrary instant, which for convenience we call time zero.

What concerns us in this paper is the likelihood that the period of time a job would have to wait before being processed is long and/or the number of jobs from each source awaiting processing is large. This is of practical interest when storage space for jobs awaiting processing is segregated based on job source.

We prove logarithmic asymptotics for the likelihood of this rare event and determine most likely paths. By comparison with simulation, we demonstrate that these asymptotic results can make good predictions, even in an arguably non-asymptotic regime.

*American Mathematical Society 1991 subject classifications: Primary 60K25; Secondary 60F10, 90B05.
†Keywords: Functional Large Deviations, Single Server FIFO, Waiting Time, Queue length.

1.1 Notation and quantities of interest

For each source $k \in \{1, \ldots, d\}$, let $\tau_1^k > 0$ denote the most recent time prior to 0 that a job from source k arrived. Label that job as -1 and those prior to it sequentially. For each $n \geq 2$, let $\tau_n^k > 0$ denote the time between the arrival of jobs $-n$ and $-n+1$ and let $\xi_n^k > 0$ denote the size of job $-n$; that is the processor will take time ξ_n^k/C to process job $-n$.

Define the sum of job sizes from $-n$ to -1 from source $k \in \{1, \ldots, d\}$ and the vector of the summed job sizes by:

$$B^k(n) := \sum_{i=1}^{n} \xi_i^k \text{ and } \vec{B}(n) := (B^1(n), \ldots, B^d(n)).$$

Define the total time that passes between the arrival of jobs $-n$ and time 0 for source $k \in \{1, \ldots, d\}$ and the vector of times by

$$T^k(n) := \sum_{i=1}^{n} \tau_i^k \text{ and } \vec{T}(n) := (T^1(n), \ldots, T^d(n)).$$

Define the number of jobs that have arrived from source k in the interval of length $t > 0$ prior to time 0, the corresponding vector, and the total number of jobs across all sources, by

$$N^k(t) := \sup\{n : T^k(n) \leq t\},\ \vec{N}(t) := (N^1(t), \ldots, N^d(t)) \text{ and } N(t) = \sum_{k=1}^{d} N^k(t).$$

The total amount of work that source k brings in the interval $[-t, 0)$ is

$$A^k(t) := B^k(N^k(t)), \text{ with } \vec{A}(t) := (A^1(t), \ldots, A^d(t)).$$

In the stochastic process nomenclature, $\{A^k(t)\}$ is a (pure, zero-delayed) cumulative process. The total amount of work brought by the sum of all d sources is

$$A(t) := \sum_{k=1}^{d} B^k(N^k(t)).$$

The waiting time of a job is the time between its arrival and the initiation of service upon it. The waiting time ω of a virtual job (a job of zero size) inserted into the queue at time 0 can be determined from Lindley's recursion (see Loynes [14]) to be

$$\omega = \sup_{t>0} \left(\frac{A(t)}{C} - t \right). \tag{1}$$

As we assume the server employees the FIFO queueing discipline, the waiting time is the same for a virtual job from any source.

We define source k's queue length to be the number of jobs from source k whose processing has not completed by the time the tagged job arrives. We are interested in the queue length, η^k, for a virtual job (a job of zero size) of type k that arrives at time 0. This is given by

$$\eta^k = N^k \left(\sup \{s : A(s) \leq C\omega\} \right). \tag{2}$$

This is, perhaps, easiest to see by first considering the total queue length, $\eta = \sum_{k=1}^{d} \eta^k$. It is given by the number of recently arrived jobs that would explain the waiting time:

$$\eta \quad = \quad N\left(\sup\left\{s : A(s) \leq C\omega\right\}\right).$$

In this paper we are interested in asymptotics for the likelihood of long waiting times and large queue lengths. We are also interested in the most likely paths that lead to these rare events.

1.2 A brief introduction to existing work

A clear introduction to the application of large deviation methods in queueing theory can be found in Ganesh, O'Connell and Wischik [10]. The asymptotic considered here is called the "large buffer" or "long waiting time" asymptotic.

For a single source, when job sizes and inter-arrival times are independent sequences of i.i.d. random variables, an example work that proves strong asymptotic results for the tail of the joint distribution of $\{(\omega, \eta^1)\}$ is Asmussen and Collamore [2], which strengthens results of Aspandiiarov and Pechersky [3]. In many practical problems, however, it is natural to assume there is some dependence in the job size and inter-arrival time processes. In the absence of heavy-tailed distributions and long range dependence, with a single source logarithmic asymptotics have been proved under varying degrees of generality:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\omega > n) = -\delta_\omega; \tag{3}$$

$$\text{and } \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\eta^1 > n) = -\delta_{\eta^1}, \tag{4}$$

so that we have the approximations $\mathbb{P}(\omega > n) \approx \exp(-n\delta_\omega)$ and $\mathbb{P}(\eta^1 > n) \approx \exp(-n\delta_{\eta^1})$ for large n.

For example, Glynn and Whitt [12] prove (3) under general large deviation assumptions. They also treat (4) when inter-arrival times form a stationary sequence that are independent of i.i.d. job sizes. Duffy and Sullivan [9] have since extended this result to the case where job sizes form an stationary sequence. Related work in the multiple server setting can be found in Sadowsky and Szpankowski [19] and Sadowsky [18]. Proving more than equation (3), the Large Deviation Principle (LDP) for $\{\omega/n\}$ has been established by Ganesh and O'Connell [11] using functional techniques.

In the presence of heavy-tailed distributions or long range dependence, the dominant behavior of the tails of the waiting time and queue length distributions is no longer exponential in n. In this case, an example work treating the equivalent of equation (3) is Duffy, Lewis and Sullivan [6]. That paper does not treat the case where the appropriate scale is $\log(n)$, where an example work is Mikosch and Nagaev [15].

1.3 This work's contribution

We treat a FIFO single-server queue fed by several (possibly dependent) sources of jobs. We consider the joint distribution of the overall waiting time and the queue length of each source. We make general large deviation assumptions on job-size and inter-arrival time processes. These assumptions hold, for example, for sources with Markovian dependencies. Our assumptions exclude heavy tailed distributions and long range dependence.

The main contributions of this paper are the following.

- Theorems 7 and 8 prove that $\{\omega/n, \eta^1/n, \ldots, \eta^d/n\}$ satisfies the large deviation principle for a large collection of sources of jobs.
- We deduce logarithmic asymptotics for the likelihood there is a large waiting time and a large number of jobs from each source queued at the server. In particular, for sources whose job-sizes and job inter-arrival times may consist of non-i.i.d. random variables and may be correlated processes, but with sources being independent of each other, Corollary 9 proves that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{\omega}{n} > w, \frac{\eta^1}{n} > q_1, \ldots, \frac{\eta^d}{n} > q_d\right) = - \inf_{w' \geq w, q'_1 \geq q_1, \ldots, q'_d \geq q_k} K(w', q'_1, \ldots, q'_d), \qquad (5)$$

  where K is a convex rate function that is linear on rays, which is written in terms of the rate functions for the sources. This is illustrated with an example of sources whose job size and inter-arrival times are correlated and are jointly driven by a Markov chain.
- Using a methodology based on functional large deviation techniques, we also determine the most likely paths to this large waiting time and these large queue lengths. The paths are more intricate than those conditioned solely on a large waiting time, with two different piece-wise linear behaviors leading to the rare event; see Corollary 10, and equations (15) and (16) that follow it.
- We demonstrate how our theorems recover known results under simplifying assumptions. We demonstrate their novelty by treating an example where sources have Markovian dependencies. We compare the theory's predictions, including most likely paths, with output from simulation.

1.4 A practical motivation

In modern packet-switched networks, such as the Internet, data packets (i.e. jobs) are of variable size, described by a number of bytes. Link speeds are given in bytes per second and, therefore, the time taken to transmit a packet across a link is a function of its size. As the devices routing these packets can have multiple input and output ports, they contain buffering to deal with temporary backlogs of packets. Packets that arrive when a buffer is full are lost. An interesting and fundamental peculiarity of engineering design is that these buffers are typically determined as a number of packets, not a number of bytes. Thus the remaining storage space at any instant is not determined by the total number of bytes that remain to be processed, but how many packets are awaiting processing. That is, the remaining storage space is determined by the number of back-logged jobs, not how long it will take to process the back-logged work.

For example, on standard Ethernet (such as the wired connection on most PCs) there is a Maximum Transmission Unit (MTU) [largest job size] of 1500 bytes, while on IEEE 802.11 (the wireless connection used on most laptops) the MTU is 2356 bytes. Buffers in switches and routers are determined by a number of "pigeon holes" each capable of holding a single packet of any size up to this MTU. When a packet arrives and is placed in a pigeon hole, it fully occupies the space, no matter how few bytes it contains.

For some of these buffered devices, particularly high end routers, the overall buffering capability is fixed. A network engineer can then decide how to divide buffer space between input ports (i.e. between

sources of jobs) to ensure that the likelihood losses occur (i.e. the buffers overflow) is minimized. A naive approach to this question would be based on the average mix of traffic that is transported through the device. However, assuming that network capacity is well provisioned so that large backlogs are rare events, this approach may be far from optimal. We use the model to consider this question in Section 4.3.

1.5 Organization

The rest of this paper is organized as follows. In section 2 we remind the reader of basic results from large deviation theory and introduce the function space we will use. In section 3 we introduce the functional quantities of interest and state our main results; all proofs are deferred to appendix A. In section 4 we show that the main theorem generalizes known results and present a number of examples that illustrate salient features.

2 The large deviation principle and our functional setup

For convenience we recall the basic facts of the Large Deviation Principle (LDP), which can be found in a standard text such as Dembo and Zeitouni [5], and introduce the function space we will use. Let $\mathcal{X}$ be a Hausdorff space with Borel σ-algebra $\mathcal{B}$ and let $\{\mu_n, n \in \mathbb{N}\}$ be a sequence of probability measures on $(\mathcal{X}, \mathcal{B})$. We say that $\{\mu_n, n \in \mathbb{N}\}$ satisfies the LDP with rate function $I : \mathcal{X} \to [0, +\infty]$ if I is lower semi-continuous,

$$-\inf_{x \in G} I(x) \leq \liminf_{n \to \infty} \frac{1}{n} \log \mu_n[G] \qquad \text{and} \qquad \limsup_{n \to \infty} \frac{1}{n} \log \mu_n[F] \leq -\inf_{x \in F} I(x)$$

for all open G and all closed F. Furthermore, as in Varadhan's original definition [20], we will assume that the level sets of I, $\{x : I(x) \leq \alpha\}$, are compact for all $\alpha \geq 0$. Rate functions with this property are sometimes referred to as being "good", but all our rate functions will have this property.

We say that a process $\{X_n\}$ satisfies the LDP if X_n is a realization of μ_n for each n. Two sequences of random elements $\{X_n\}$ and $\{Y_n\}$ taking values in a metric space with metric d are defined to be exponentially equivalent if $\limsup_{n \to \infty} n^{-1} \log \mathbb{P}(d(X_n, Y_n) > \delta) = -\infty$ for all $\delta > 0$. If two processes are exponentially equivalent, then one process satisfies the LDP with rate function I if and only if the other process also does.

The Contraction Principle (e.g. Theorem 4.2.16 of [5]) states that if $\{X_n\}$ satisfies the LDP in $\mathcal{X}$ with good rate function I and $f : \mathcal{X} \to \mathcal{Y}$ is continuous, where $\mathcal{X}$ and $\mathcal{Y}$ are Hausdorff, then $\{f(X_n)\}$ satisfies the LDP in $\mathcal{Y}$ with good rate function given by $J(y) := \inf\{I(x) : f(x) = y\}$. Moreover, if $\{X_n\}$ and $\{Y_n\}$ are exponentially equivalent, then $\{f(Y_n)\}$ also satisfies the LDP with rate function J.

Let $\mathcal{C}^d[0, \infty)$ denote the collection of $\mathbb{R}^d$-valued continuous functions on $[0, \infty)$. Let $\mathcal{A}^d[0, \infty)$ denote the subset of functions on $[0, \infty)$ that are absolutely continuous on $[0, x]$ for all $x < \infty$. Letting ϕ_j denote the j^{th} component of ϕ, define the space

$$\mathcal{Y}^d := \left\{ \phi \in \mathcal{C}^d[0, \infty) : \lim_{t \to \infty} \frac{\phi(t)}{1 + t} \text{ exists in } \mathbb{R}^d \text{ and } \phi(0) = \vec{0} \right\}$$

and equip it with the topology induced by the norm

$$||\phi|| = \max_{j=1,\ldots,d} \sup_{t\geq 0} \left| \frac{\phi_j(t)}{1+t} \right|.$$

Define

$$\mathcal{Y}^d_\uparrow := \left\{ \phi \in \mathcal{Y}^d : \phi \text{ component-wise strictly increasing and } \min_{j=1,\ldots,d} \lim \frac{\phi_j(t)}{1+t} > 0 \right\}$$

and, for each $\vec{\mu} = (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d$, define

$$\mathcal{Y}^d_{\vec{\mu}} := \left\{ \phi \in \mathcal{Y}^d : \lim \frac{\phi_j(t)}{1+t} = \mu_j \right\}.$$

Treat $\mathcal{Y}^d_\uparrow$ and $\mathcal{Y}^d_{\vec{\mu}}$ as metric subspaces of $\mathcal{Y}^d$. Products of these spaces are equipped with the product topology.

The motivation for the value of these spaces in the consideration of the LDP is as follows. Given an $\mathbb{R}^d$-valued stochastic process $\{X_n, n \in \mathbb{N}\}$ and defining $X_0 := 0$, the usual sample paths of its partial sums process $\{\sum_{i=1}^n X_i, n \in \mathbb{N}\}$ defined by

$$\hat{S}_n(t) := \frac{1}{n} \sum_{i=0}^{[nt]} X_i, \text{ for } t \in [0, \infty),$$

are not continuous functions. They are right-continuous with left-hand limits (CADLAG functions). However their polygonal approximations,

$$S_n(t) := \frac{1}{n} \sum_{i=0}^{[nt]} X_i + \left(t - \frac{[nt]}{n} \right) X_{[nt]+1}, \text{ for } t \in [0, \infty),$$

are continuous. We shall call S_n a sample path.

Restricting S_n to $[0,1]$, Dembo and Zajic [4] generalize Mogulskii's theorem [16] by providing broad conditions under which $\{S_n, n \in \mathbb{N}\}$ satisfies the LDP with good rate function in the space of continuous functions on $[0,1]$ equipped with the sup norm. Under their assumptions, the rate function is given by

$$I(\zeta) = \begin{cases} \int_0^1 \lambda^*(\dot{\zeta}(t))dt & \text{if } \zeta \in \mathcal{A}^d[0,1], \\ +\infty & \text{otherwise,} \end{cases}$$

where $\mathcal{A}^d[0,1]$ denotes the absolutely continuous functions on $[0,1]$ with $\phi(0) = 0$, λ^* is the Legendre transform of the scaled cumulant generating function

$$\lambda(\theta) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[e^{n\langle \theta, S_n(1) \rangle}]$$

and $\langle \cdot, \cdot \rangle$ is the usual inner product.

Theorem 1 of Ganesh and O'Connell [11] establishes that if Dembo and Zajic's conditions are met and λ is differentiable at the origin, then $\{S_n, n \in \mathbb{N}\}$ also satisfies the LDP in $\mathcal{Y}$ with good rate function

$$I_\infty(\zeta) = \begin{cases} \int_0^\infty \lambda^*(\dot{\zeta}(t))dt & \text{if } \zeta \in \mathcal{A}^d[0,\infty) \cap \mathcal{Y}, \\ +\infty & \text{otherwise.} \end{cases} \tag{6}$$

Thus the sample paths of many partial sums processes satisfy the LDP in the function space $\mathcal{Y}$, which is indexed by the positive real line. Rate functions of the sort in equation (6) are particularly interesting. We refer to them as of *integral form* with integrand λ^*.

3 Functional quantities of interest and statement main results

In this section we state our main results; proofs are deferred to appendix A. We begin with the sample paths of job sizes and inter arrival times, $\{\vec{B}(n), \vec{T}(n)\}$. These are sample paths of partial sums processes, so it is known that a large class satisfy the functional LDP. We construct the sample paths for $\{\vec{A}(t), \vec{N}(t)\}$ and relate them to the sample paths of $\{\vec{B}(n), \vec{T}(n)\}$. These are no longer sample paths of partial sums processes, but their functional large deviation behavior can be deduced from that for $\{\vec{B}(n), \vec{T}(n)\}$. We then introduce a functional queue length map. When applied to sample paths it does not exactly match the system queue length, but we prove that the discrepancy is insignificant on the scale of large deviations. We prove this map is continuous, which enables us to deduce our main result. We then specialize this result to independent sources and give an explicit form for the paths leading to large queue lengths and long waiting times.

Define by $\{\vec{B}_n(\cdot) = (B_n^1(\cdot), \ldots, B_n^d(\cdot))\}$ and $\{\vec{T}_n(\cdot) = (T_n^1(\cdot), \ldots, T_n^d(\cdot))\}$ the (polygonal approximations to the usual) sample paths of $\{(\xi_n^1, \ldots, \xi_n^d)\}$ and $\{(\tau_n^1, \ldots, \tau_n^d)\}$. As $\tau_n^k > 0$ for all k and n, for each $k \in \{1, \ldots, d\}$ $T_n^k(\cdot)$ is strictly increasing and continuous. For each n we define the function $N_n^k(\cdot)$ to be the inverse of $T_n^k(\cdot)$: $N_n^k(T_n^k(t)) = T_n^k(N_n^k(t)) = t$. It can be explicitly determined to be the polygonal approximation to the function $t \mapsto N^k(nt)/n = \sup\{m : T^k(m) \leq nt\}/n$:

$$N_n^k(t) = \frac{1}{n} N^k(nt) + \left(t - \frac{1}{n} \sum_{i=0}^{N^k(nt)} \tau_i^k \right) \frac{1}{\tau_{N^k(nt)+1}}.$$

This is a polygonal approximation as

$$t - \frac{1}{n} \sum_{i=0}^{N^k(nt)} \tau_i^k \in \left[0, \frac{\tau_{N^k(nt)+1}}{n} \right)$$

and on this interval this difference is linear in t. For each $n > 0$, we define $\vec{N}_n(\cdot) = (N_n^1(\cdot), \ldots, N_n^d(\cdot))$.

We also define $\vec{A}_n(\cdot) = (B_n^1(N_n^1(\cdot)), \ldots, B_n^d(N_n^d(\cdot)))$, the sample paths of the total arrivals. Noting that $[nN_n^k(t)] = N^k(nt)$, it can be readily verified that

$$A_n^k(t) = B_n^k(N_n^k(t)) = \frac{1}{n} \sum_{i=0}^{N^k(nt)} \xi_i^k + \left(t - \frac{1}{n} \sum_{i=0}^{N^k(nt)} \tau_i^k \right) \frac{\xi_{N^k(nt)+1}^k}{\tau_{N^k(nt)+1}^k}.$$

The function $\vec{A}_n(\cdot)$ is the polygonal approximation to the sample paths of $t \mapsto \vec{A}(tn)/n$. Here the polygonal part is in $[0, \xi_{N^k(nt)+1}^k/n)$.

Assumption 1 (LDP) *The sample paths $\{\vec{B}_n(\cdot), \vec{T}_n(\cdot)\}$ satisfy the LDP in $\mathcal{Y}^d \times \mathcal{Y}_\uparrow^d$ with rate function I_∞ of integral form with integrand I, so that*

$$I_\infty(\phi_1, \ldots, \phi_{2d}) = \begin{cases} \int_0^\infty I(\dot{\phi}_1(s), \ldots, \dot{\phi}_{2d}(s)) ds & \text{if } \phi = (\phi_1, \ldots, \phi_{2d}) \in \mathcal{A}^{2d}[0, \infty) \\ +\infty & \text{otherwise.} \end{cases}$$

Assumption 1 is that the sample paths of the partial sums processes of job sizes and inter-arrival times satisfy the LDP. As they are partial sums, they fall within the remit of Ganesh and O'Connell's Theorem 1 in [11] and so the assumption holds for a large class of processes. The assumption enables us to deduce the LDP for the total arrivals up to a given time and the number of jobs from each source which have arrived by this time. As these are not partial sums processes, we could not have invoked Theorem 1 in [11] to justify the LDP for them directly.

Theorem 1 (Cumulative and counting processes sample path LDP) *Under assumption 1, the sample paths* $\{\vec{A}_n(\cdot), \vec{N}_n(\cdot)\}$ *satisfy the LDP with rate function*

$$J_\infty(\phi_1, \ldots, \phi_d, \psi_1, \ldots, \psi_d) = I_\infty(\phi_1(\psi_1^{-1}), \ldots, \phi_d(\psi_d^{-1}), \psi_1^{-1}, \ldots, \psi_d^{-1}).$$

We will define functions that takes sample paths to the waiting time and to estimates of the queue lengths. They require knowledge of the sample path for the total arrival rate (summed across all sources) and the job number sample paths for each source. However, the sum of the polygonal sample paths of total arrivals

$$\hat{A}_n(t) := \sum_{k=1}^{d} A_n^k(t)$$

is not quite the same as the polygonal sample path, $A_n(\cdot)$, to the arrivals interlaced across all sources, which we now define. The interlaced arrival times are given by

$$T^*(0) := 0 \text{ and } T^*(n) := \min_{k=1,\ldots,d} \inf_{m \geq 1} \left\{ T^k(m) : T^k(m) \geq T^*(n-1) \right\} \text{ for each } n \geq 1.$$

Define $N^*(t) = \sup\{n : T^*(n) \leq t\}$. Let ξ_n^* denote the job size corresponding to the arrival at time $T^*(n)$ and $\tau_n^* = T^*(n) - T^*(n-1)$ the interlaced inter arrival times. If jobs from two or more sources arrive simultaneously, we combine the job sizes and treat them as a single arrival at that time; this causes no extra difficulty. We then define the (polygonal) sample paths to the interlaced arrivals by:

$$A_n(t) := \frac{A(nt)}{n} + \left(t - \frac{1}{n} \sum_{i=0}^{N^*(nt)} \tau_i^* \right) \frac{\xi_{N^*(nt)+1}^*}{\tau_{N^*(nt)+1}^*}.$$

The cumulative arrival rate to the processor is $A_n(\cdot)$ not $\hat{A}_n(\cdot)$, but the latter has a more convenient representation. We overcome the discrepancy between them with an assumption that ensures that the processes $\{\hat{A}_n(\cdot)\}$ and $\{A_n(\cdot)\}$ are exponentially equivalent.

Assumption 2 (Exponential equivalence) *For every* $k \in \{1, \ldots, d\}$

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left(\sup_m \frac{\xi_{m+1}^k}{1 + T^k(m)} > n \right) = -\infty.$$

If job sizes are i.i.d, then a sufficient condition for this assumption to hold (irrespective of the nature of job inter-arrival times), is that the tail of the job size distribution decays faster than exponentially (e.g. Weibull with parameter $\alpha > 1$). Note that regardless of correlation structure, if job sizes are bounded this assumption is trivially satisfied. If job sizes are i.i.d and independent of job inter-arrival times, then exponential job sizes will *not* satisfy this criterion.

Lemma 2 (Exp. equivalence of interlaced and summed sample paths) *Under assumption 2, the processes* $\{A_n(\cdot)\}$ *and* $\{\hat{A}_n(\cdot)\}$ *are exponentially equivalent.*

The functional representation for the waiting time is the well known map $W : \mathcal{Y}^1 \mapsto \mathbb{R}_+$ defined by $W(\phi) = \sup_{t\geq 0}\{\phi(t)/C - t\}$. The following lemma is known, but as we cannot find an explicit reference for it, we include it for completeness. It will also help to illustrate an additional difficulty that arises with the queue lengths.

Lemma 3 (Equivalence of waiting time and functional representation) *For every* $m > 0$, $W(A_m(\cdot)) = \omega/m$.

Ganesh and O'Connell prove that W is a continuous representation, so long as the queue is stable:

Lemma 4 (Ganesh and O'Connell [11]) *If* $\mu < C$, *then the map* W *is continuous on the space* $\mathcal{Y}^1_\mu$.

We wish to consider a similar construction for a function representation of the queue length. Define the map $Q^k : \mathcal{Y}^1 \times \mathcal{Y}^d_\uparrow \mapsto \mathbb{R}_+$ by:

$$Q^k(\phi, \vec{\psi}) := \psi_k\left(\sup\left\{s : \phi(s) \leq CW(\phi)\right\}\right). \tag{7}$$

When applied to $(A_n(\cdot), \vec{N}_n(\cdot))$, the map Q^k is a functional analogue of η^k in equation (2) but, unlike the waiting time, it is not exact when applied to sample paths. That is, $Q^k(A_n(\cdot), \vec{N}_n(\cdot)) \approx \eta^k/n$, but they are not necessarily equal. This discrepancy is caused by our use of polygonal approximation to the usual CADLAG sample paths. The following lemma proves that as far as the LDP is concerned the discrepancy is insignificant.

Lemma 5 (Exp. equivalence of queue lengths and the functional representation) *The sequence* $\{Q^k(A_n(\cdot), \vec{N}_n(\cdot)), n \in \mathbb{N}\}$ *is exponentially equivalent to the sequence* $\{\eta^k/n, n \in \mathbb{N}\}$.

In order to deduce the LDP for $\{\eta^k/n, n \in \mathbb{N}\}$ from the LDP for $\{A_n(\cdot), \vec{N}_n(\cdot), n \in \mathbb{N}\}$ it suffices to show that Q^k is continuous and apply the contraction principle. For the map Q^k to be continuous, the queueing system must be stable. That is, the long term average job inter-arrival time divided by the service rate must be greater than the average job size.

Lemma 6 (Continuity of the queue length functional) *Let* $\mu \in \mathbb{R}$ *be such that* $\mu < C$, *then* Q^k, *defined in equation* (7), *is continuous on* $(\mathcal{Y}^1_\mu \cap \mathcal{Y}^d_\uparrow) \times \mathcal{Y}^d$.

The assumption in Lemma 6 is a restriction on the long run average of the cumulative arrivals process. The following assumption translates this into one on job sizes and inter-arrival times.

Assumption 3 (LDP plus stability) *The sample paths* $\{\vec{B}_n(\cdot), \vec{T}_n(\cdot)\}$ *satisfy the LDP with rate function* I_∞ *of integral form with integrand* I *in* $(\mathcal{Y}^d_{\vec{\mu}} \cap \mathcal{Y}^d_\uparrow) \times (\mathcal{Y}^d_{\vec{\nu}} \cap \mathcal{Y}^d_\uparrow)$, *where* $\vec{\mu} = (\mu_1, \ldots, \mu_d)$, $\vec{\nu} = (\nu_1, \ldots, \nu_d)$ *and*

$$\rho := \sum_{k=1}^{d} \frac{\mu_k}{\nu_k} < C.$$

Define the map $Q : \mathcal{Y}^1 \times \mathcal{Y}^d_\uparrow \mapsto \mathbb{R}_+$ by:

$$Q(\phi, \vec{\psi}) := (Q^1(\phi, \vec{\psi}), \ldots, Q^d(\phi, \vec{\psi})), \tag{8}$$

where Q^k is defined in equation (7). The map Q describes the queue length for every source at time 0. The following is our main general result.

Theorem 7 (Main result for dependent sources) *If assumptions 2 and 3 are satisfied, then the process $\{(\omega/n, \eta^1/n, \ldots, \eta^d/n)\}$ satisfies the LDP with rate function*

$$K(w, \vec{q}) = \inf \left\{ J_\infty(\vec{\phi}, \vec{\psi}) : W(\phi) = w, Q(\phi, \vec{\psi}) = \vec{q},\ \vec{\phi} \in \left(\mathcal{Y}^d_{\vec{\kappa}} \cap \mathcal{Y}^d_\uparrow\right),\ \vec{\psi} \in \left(\mathcal{Y}^d_{\vec{\nu}^{-1}} \cap \mathcal{Y}^d_\uparrow\right) \right\}, \tag{9}$$

where $\vec{q} \in \mathbb{R}^d_+$, $\phi = \sum_{k=1}^d \phi_k$, $\vec{\kappa} = (\mu_1/\nu_1, \ldots, \mu_d/\nu_d)$ and $\vec{\nu}^{-1} := (1/\nu_1, \ldots, 1/\nu_d)$.

The formula (9) holds in broad generality and is qualitatively informative. However, to get a more tractable quantitative handle we assume that sources are independent of each other. Note that this does not require us to assume that jobs sizes are independent of inter-arrival times within a given source. This enables us to obtain a refined form as each process can be time-changed independently.

Assumption 4 (Independent sources) *For each pair $j, k \in \{1, \ldots, d\}$ with $j \neq k$, the processes $\{(\xi^j_i, \tau^j_i)\}$ and $\{(\xi^k_i, \tau^k_i)\}$ are independent.*

Note that under assumptions 1 and 4, we have

$$I_\infty(\phi_1, \ldots, \phi_{2d}) = \begin{cases} \displaystyle\int_0^\infty \left[\sum_{k=1}^d I_k \left(\dot{\phi}_k(s), \dot{\phi}_{d+k}(s) \right) \right] ds & \text{if } \phi = (\phi_1, \ldots, \phi_{2d}) \in \mathcal{A}^{2d}[0, \infty) \\ +\infty & \text{otherwise,} \end{cases} \tag{10}$$

where I_k is the integrand in the rate function for $\{(B^k_n(\cdot), T^k_n(\cdot))\}$.

Theorem 8 (Main result for independent sources) *In addition to assumptions 2 and 3, assume that assumption 4 holds and each I_k in equation (10) is convex. Define $J_k(x, y) := yI_k(x/y, 1/y)$ for each $k \in \{1, \ldots, d\}$. Then:*

- *$\{\omega/n\}$ satisfies the LDP with the linear rate function*

$$K_\omega(w) = w\delta_\omega := wC \inf_{z>0} zL(C + 1/z), \tag{11}$$

where

$$L(x) := \inf_{(x_1, \ldots, x_d) : \sum_{k=1}^d x_k = x} \left\{ \sum_{k=1}^d \left(\inf_{y_k \geq 0} J_k(x_k, y_k) \right) \right\} \tag{12}$$

is the rate function for $\{A_n(1)\}$;

- *and, with $\vec{q} = (q_1, \ldots, q_d)$, $\{(\omega/n, \eta^1/n, \ldots, \eta^d/n)\}$ satisfies the LDP with convex rate function*

$$K(w, \vec{q}) = \inf_{x \geq 0} \inf_{\{\vec{y} : \sum_{k=1}^d y_k = Cw\}} \left\{ \sum_{k=1}^d xJ_k \left(\frac{y_k}{x}, \frac{q_k}{x} \right) + x\delta_\omega \right\}, \tag{13}$$

where K is linear on rays; that is, for any $\alpha > 0$, $K(\alpha w, \alpha \vec{q}) = \alpha K(w, \vec{q})$.

Comment (K linear on rays): note that K has to be linear on rays as for any $\alpha \in (0,\infty)$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left((\omega, \eta^1, \ldots, \eta^d) > (\alpha n w, \alpha n q_1, \ldots, \alpha n q_d)\right)$$
$$= \alpha \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left((\omega, \eta^1, \ldots, \eta^d) > (n w, n q_1, \ldots, n q_d)\right),$$

by change of variable.

A natural corollary to this LDP is the following logarithmic asymptotic for the tail of joint waiting time and queue lengths distribution.

Corollary 9 (Logarithmic asymptotics for independent sources) *If*

$$\inf_{w'>w, q_1'>q_1, \ldots, q_d'>q_d} K(w', q_1', \ldots, q_d') = \inf_{w'\geq w, q_1'\geq q_1, \ldots, q_d'\geq q_d} K(w', q_1', \ldots, q_d'), \tag{14}$$

(a sufficient condition for which is that $(w, q_1, \ldots, q_d)$ is in the effective domain of K) then

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left((\omega, \eta^1, \ldots, \eta^d) > (n w, n q_1, \ldots, n q_d)\right) = - \inf_{w'>w, q_1'>q_1, \ldots, q_d'>q_d} K(w', q_1', \ldots, q_d').$$

Comment (finite waiting space): Taking an infimum over w in equation (14) results in the logarithmic asymptotic for the likelihood of dropped jobs when there is waiting space $q_k n$ for each source k: the lower bound is automatic, as the queue lengths with infinite waiting space path-wise dominate those with finite waiting space; the upper bounds follow as paths in a small neighborhood around those identified in Theorem 8 would cause overflow when there is finite waiting space and they have the correct asymptotic likelihood.

The following corollary on the most likely paths that lead to this rare event is in the spirit of Anantharam [1], who treats paths to a large waiting time over a finite time horizon for the queue with i.i.d. job size process that is independent of an i.i.d. inter-arrival time process. Note that, when constraining only on waiting time, the most likely paths are simpler.

Corollary 10 (Most likely paths for independent sources) *Assume the that infima for $K(\vec{q})$ in equations* (13) *and* (12) *are obtained at unique $(x^*, y_1^*, \ldots, y_d^*, z^*) \in (0,\infty)^{d+2}$ and $(a_1^*, \ldots, a_d^*, b_1^*, \ldots, b_d^*) \in (0,\infty)^{2d}$, where $\sum_{k=1}^d a_k^* = C(z^*+x^*)/z^*$, so we have that*

$$K(\vec{q}) = \sum_{k=1}^d \left(x^* J_k\left(\frac{y_k^*}{x^*}, \frac{q_k}{x^*}\right)\right) + z^* \left(\sum_{k=1}^d J_k(a_k^*, b_k^*)\right).$$

Define the following piecewise linear functions for each $k \in \{1, \ldots, d\}$:

$$\hat{\phi}_k(t) = \begin{cases} y_k^* t/x^* & \text{if } t < x^* \\ y_k^* + (t-x^*)a_k^* & \text{if } x^* \leq t < x^*+z^* \\ y_k^* + z^* a_k^* + (t-x^*-x^*)\mu_k & \text{if } t \geq x^*+z^* \end{cases}$$

and

$$\hat{\psi}_k(t) = \begin{cases} q_k t/x^* & \text{if } t < x^* \\ q_k + (t-x^*)b_k^* & \text{if } x^* \leq t < x^*+z^* \\ q_k + b_k^* z^* + (t-x^*-z^*)\nu_k & \text{if } t \geq x^*+z^* \end{cases}$$

where (μ_k, ν_k) is the unique pair such that $J_k(\mu_k, \nu_k) = 0$. Then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left((\vec{A}_n, \vec{N}_n) \in B_\epsilon\left(\hat{\phi}_1, \ldots, \hat{\phi}_d, \hat{\psi}_1, \ldots, \hat{\psi}_d\right) | \right.$$
$$\left. W(A_n) \geq w, Q^1(A_n, \vec{N}_n) \geq q_1, \ldots, Q^d(A_n, \vec{N}_n) \geq q_d\right) = 1,$$

where $B_\epsilon(\hat{\phi}_1, \ldots, \hat{\phi}_d, \hat{\psi}_1, \ldots, \hat{\psi}_d)$ is the ball of radius ϵ around $(\hat{\phi}_1, \ldots, \hat{\phi}_d, \hat{\psi}_1, \ldots, \hat{\psi}_d)$ in $\mathcal{Y}^{2d}$. That is, these $(\hat{\phi}_1, \ldots, \hat{\phi}_d, \hat{\psi}_1, \ldots, \hat{\psi}_d)$ are the most likely paths that lead to the rare event.

This corollary implies that given we observe a long waiting time and large queue lengths [i.e. $(\omega > wn, \eta^1 > q_1 n, \ldots, \eta^d > q_d n)$], that when n is large, with $\rho := \sum \mu_k/\nu_k$, the most likely path to this event for the waiting time satisfies:

$$\omega(t) \approx \begin{cases} 0 & \text{if } t \leq -n(x^* + z^*), \\ n(x^* + (x^* + t/n)x^*/z^*) & \text{if } -n(x^* + z^*) \leq t \leq -nx^*, \\ wn + t(w/x^* - 1) & \text{if } -nx^* \leq t \leq 0, \\ wn - t(C - \rho) & \text{if } 0 \leq t \leq wn/(C - \rho), \\ 0 & \text{if } t \geq wn/(C - \rho), \end{cases} \tag{15}$$

while for each queue, the path will satisfy

$$\eta^k(t) \approx \begin{cases} 0 & \text{if } t \leq -n(x^* + z^*), \\ nx^* b_k^* + tx^* b_k^*/(x^* + z^*) & \text{if } -n(x^* + z^*) \leq t \leq -nx^*, \\ nq_k + t(q_k/x^* - z^* b_k^*/(x^* + z^*)) & \text{if } -nx^* \leq t \leq 0 \\ q_k n - t(C - \rho)q_k/\omega & \text{if } 0 \leq t \leq wn/(C - \rho), \\ 0 & \text{if } t \geq wn/(C - \rho), \end{cases} \tag{16}$$

These paths have five piecewise linear parts, with the waiting time and queue lengths all switching modes at the same time. In the first there part is no queue build up until $t = -n(x^* - z^*)$. From $t = -n(x^* + z^*)$ to $t = -nx^*$ the sources combine to generate sufficient volume of arrivals so that at $t = -nx^*$ the workload will take until $t = 0$ for the server to process, allowing the queue lengths and waiting time to build up from $t = -nx^*$ to $t = 0$. Between $t = 0$ to $t = wn/(C - \rho)$, the server processes the backlog of work and from then on the queues are empty. We will demonstrate these distinct regimes in the Section 4.

4 Examples

4.1 Simplification of the rate function, $K(w, \vec{q})$, under additional assumptions

We consider the form of the rate function $K(w, \vec{q})$ in equation (13) under a series of simplifying assumptions to demonstrate that Theorems 7 and 8 generalize existing results.

1. If we only concern ourselves with the waiting time, then

$$\begin{aligned}\inf_{\vec{q}} K(w,\vec{q}) &= \inf_{x\geq 0} \inf_{\{\vec{y}:\sum_{k=1}^d y_k = Cw\}} \left\{ \sum_{k=1}^d x \inf_{q_k} J_k\left(\frac{y_k}{x}, q_k\right) + x\delta_\omega \right\} \\ &= \inf_{x\geq 0} \left\{ xL\left(\frac{Cw}{x}\right) + x\delta_\omega \right\} \\ &= w\delta_\omega.\end{aligned}$$

   For example, this formula could be deduced from results in [12].

2. If we restrict our interest to considering the tail of the queue length distribution of a single source $k \in \{1,\ldots,d\}$, then, with $\vec{q} = (q_1,\ldots,q_d)$, rate function for the $\{\eta^k/n\}$ is:

$$\begin{aligned}K_k(q) &:= \inf\{K(w,\vec{q}) :\ q_k = q\} \\ &= q \inf_{x\geq 0} \left\{ \inf_y I_k(y,x) + x\delta_\omega \right\} \\ &=: q\delta_k, \end{aligned} \tag{17}$$

   where δ_ω defined in equation (11) is the exponent in the tail of the waiting time distribution. Thus the large deviation lower and upper bounds tell us that the tail of the queue length distribution of any individual source satisfies:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\eta^k > n) = -\delta_k.$$

   That is, for large n, $\mathbb{P}(\eta^k > n) \approx \exp(-n\delta_k)$. For example, related result appear in [19][18].

3. Finally, if there is only one source of jobs present, the service rate C is set to one and job sizes for this source are independent of inter-arrival times, so that $I_1(x,y) = I_\xi(x) + I_\tau(y)$ (with I_ξ, I_τ convex), then

$$\begin{aligned}K_1(q) &= q\delta_1 \\ &= q \inf_{y\geq 0} \inf_{a\leq y} \inf_{z\geq 0} \left\{ zI_\xi\left(\frac{y}{z}\right) + zI_\tau\left(\frac{y-a}{z}\right) + I_\tau(a) \right\} \\ &= q \inf_{z\geq 0} \inf_{y\geq 0} \left\{ zI_\xi\left(\frac{y}{z}\right) + (z+1)I_\tau\left(\frac{y}{z+1}\right) \right\},\end{aligned}$$

   where the last equality comes at $a = y/(z+1)$ as I_τ is convex. Thus, with a single source formed of a job size process that is independent of a job inter-arrival time process,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\eta^1 > n) = -\inf_{z\geq 0} \inf_{y\geq 0} \left\{ zI_\xi\left(\frac{y}{z}\right) + (z+1)I_\tau\left(\frac{y}{z+1}\right) \right\}.$$

   This recovers a result of Duffy and Sullivan [9], which was proved using arguments based directly on distributions. That result is itself a generalization of a theorem of Glynn and Whitt [12].

4.2 Specific processes

Even for partial sums processes whose rate functions are known, one cannot expect to get explicit solutions for equations (13) and (12). However, because of convexity, simple and efficient numerical techniques can be employed to determine the infima.

1. *Bernoulli job sizes and independent exponential service times.* To calculate the rate function, for many stochastic processes $\{X_n\}$ whose averages $\{\sum_{i=1}^{n} n^{-1}X_i\}$ satisfy the LDP it can be easiest to first calculate their scaled cumulant generating function (sCGF):

$$\lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left(\exp\left(\theta \sum_{i=1}^{n} X_i\right)\right).$$

Their rate function I is then given by the Legendre-Fenchel transform of λ:

$$I(x) = \sup_{\theta}(\theta x - \lambda(\theta)). \tag{18}$$

For example, if X_n is i.i.d., clearly $\lambda(\theta) = \log \mathbb{E}(\exp(\theta X_1))$. We use this with Bernoulli and exponentially distributed random variables.

Assume that $\{\xi_n^k\}$ forms a Bernoulli sequence such that $\mathbb{P}(\xi_n^k = A_k) = 1 - p_k$ and $\mathbb{P}(\xi_n^k = B_k) = p_k$, where $A_k < B_k$. Then $\mu_k = (1-p_k)A_k + p_k B_k$ and

$$\lambda(\theta) = \log \mathbb{E}(e^{\theta \xi_1^k}) = \log\left((1-p_k)\exp(\theta A_k) + p_k \exp(\theta B_k)\right),$$

using equation (18) it can readily be shown that

$$I_{\xi^k}(x) = \begin{cases} \frac{(B_k-x)}{B_k-A_k} \log\left(\frac{(B_k-x)}{(1-p_k)(B_k-A_k)}\right) + \frac{(x-A_k)}{B_k-A_k} \log\left(\frac{(x-A_k)}{p_k(B_k-A_k)}\right) & \text{if } x \in [A_k, B_k] \\ +\infty & \text{otherwise.} \end{cases}$$

If job inter-arrival times for source $k \in \{1, \ldots, d\}$ are i.i.d. and exponentially distributed with rate $1/\nu_k$ then their rate function is readily calculated from the sCGF

$$\lambda(\theta) = \log \mathbb{E}(\exp(\theta \tau_1^k)) = \log\left(\frac{1}{1-\theta\nu_k}\right)$$

to be

$$I_{\tau^k}(x) = \frac{x}{\nu_k} - 1 - \log\left(\frac{x}{\nu_k}\right).$$

Thus $J_k(x,y) = yI_{\xi^k}(x/y) + yI_{\tau^k}(1/y)$ equals

$$\begin{cases} \frac{B_k y - x}{B_k - A_k} \log\left(\frac{B_k y - x}{(1-p_k)(B_k-A_k)}\right) + \frac{x - A_k y}{B_k - A_k} \log\left(\frac{x - A_k y}{p_k(B_k-A_k)}\right) & \\ \qquad + \frac{1}{\nu_k} - y + y\log(y\nu_k) & \text{if } x/y \in [A_k, B_k] \\ +\infty & \text{otherwise.} \end{cases}$$

Consider this system with two sources, parameterized as follows: $A_1 = 0.1, B_1 = 1.1, p_1 = 0.5, \nu_1 = 1.2$ and $A_2 = 0.1, B_2 = 0.2, p_2 = 0.7, \nu_2 = 0.4$. This gives $\rho = 0.925$ and we set $C = 1$,

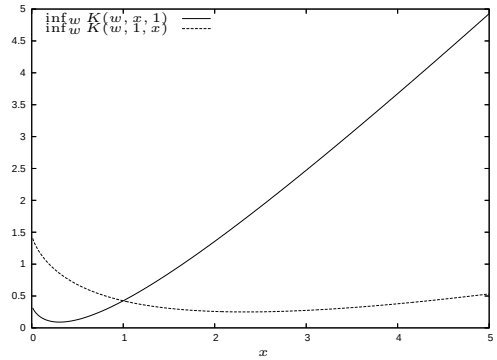

Figure 1: Bernoulli job sizes and exponential job inter-arrival times. Rate functions for the likelihood one queuelength is x times the other: $\inf_w K(w,x,1)$ and $\inf_w K(w,1,x)$ against x.

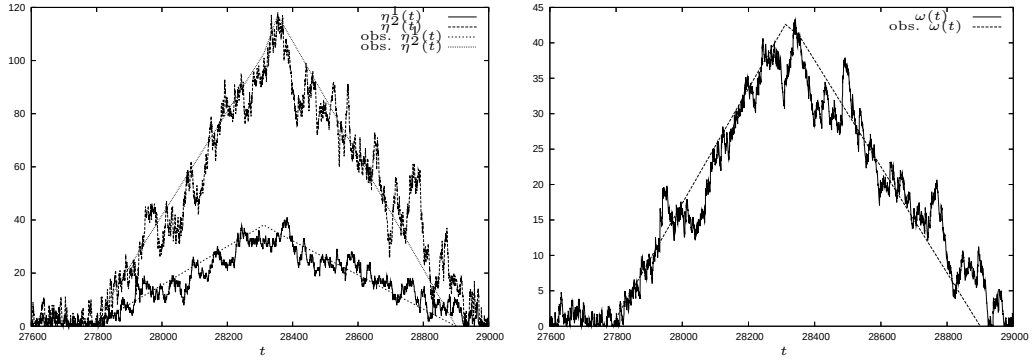

Figure 2: Bernoulli job sizes and exponential job inter-arrival times. Predicted paths for $\omega(t)$, $\eta^1(t)$ and $\eta^2(t)$ from equations (15) and (16) are overlaid with large excursions observed in simulation.

so the system is stable. Numerically solving equation (13) gives: $\delta_1 = 0.25, \delta_2 = 0.9, \delta_\omega = 0.236$, where these quantities are defined in equations (17) and (11). Note that these values are quite different to each other, despite the inter-dependency that comes about by sharing a common FIFO processor.

Figure 1 plots $\inf_w K(w,x,1)$ and $\inf_w K(w,1,x)$ against x. This captures the likelihood that given one queue length is long, the other is x times as long (without conditioning on the waiting time). This plot illustrates the asymmetry in the likelihood of these rare events. Note that $\arg\inf\{x : \inf_w K(w,x,1)\}$ determines the most likely ratio of η^1/η^2 given that η^2 is large.

We simulated this queueing system, processing of 2 million jobs from each source. Figure 2 plots the time history around the largest η^2 excursion observed. For this excursion we overlay the most likely paths as predicted by theory in equations (15) and (16). These are matched solely by setting n to the largest value of η^2; everything else is predicted by theory. The parameters for the $\eta^1(t)$ path are taken from the most likely ratio of η^1/η^2 given that η^2 is large. Similarly, for the $\omega(t)$ path we choose parameters determined by the most likely path given that η^2 is n. Both

the predicted times at which queue length and waiting time behavior change and the heights achieved at those times match well with the observed path. We have seen this for a large range of source statistics.

2. *An example with correlated job sizes and inter-arrival times.* To calculate the rate function for certain partial sums processes $\{n^{-1}\sum_{i=1}^{n} X_i\}$ it can be easiest to first calculate the rate function for the empirical laws $\{n^{-1}\sum_{i=1}^{n}\delta_{X_i}\}$ (where δ_x is the probability measure with mass 1 at x and zero elsewhere) in the weak topology on the space of probability measures. If the random variables X_n are bounded, then the rate function for $\{n^{-1}\sum_{i=1}^{n} X_i\}$ can be determined by the contraction principle.

   In [7], Duffy and Metcalfe calculate the empirical law rate function for the two state Markov chain $\{X_n, n \geq 1\}$ with transition matrix:

$$\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}, \text{ where } \alpha, \beta \in (0,1).$$

   Let $\{0,1\}$ denote the two states, then the empirical law rate function H is infinite unless the measure has the form $\Upsilon = (1-c)\delta_0 + c\delta_1$, in which case

$$H(\Upsilon) = H(c) = \begin{cases} -(1-c)\log(1-\alpha+\alpha K) - c\log(1-\beta+\beta/K) & \text{if } c \in (0,1), \\ -\log(1-\beta) & \text{if } c = 1, \\ -\log(1-\alpha) & \text{if } c = 0, \end{cases}$$

   where

$$K = \frac{-\alpha\beta(1-2c) + \sqrt{(\alpha\beta(1-2c))^2 + 4\alpha\beta c(1-\alpha)(1-\beta)(1-c)}}{2\alpha(1-\beta)(1-c)}.$$

   From this we can calculate the rate function for the partial sums process $\{n^{-1}\sum_{i=1}^{n} f(X_i)\}$ by the contraction principle. For source k, when the Markov chain is in state 0 a job of size π_0^k is generated with an inter-arrival time to the next job of υ_0^k and when the chain is in state 1 a job of size π_1^k is generated with an inter-arrival time of υ_0^k, with $\pi_1^k > \pi_0^k$. That is the state of the Markov chain completely determines both the job inter-arrival time and size, making the cumulative arrival process highly correlated. Then:

$$J_k(x,y) = \begin{cases} H\left(\dfrac{\pi_2^k(y-\upsilon_1^k) + \pi_1^k(\upsilon_2^k - y)}{(\pi_2^k - \pi_1^k)(\upsilon_2^k - \upsilon_1^k)}\right) & \text{if } y \in [\upsilon_1^k, \upsilon_2^k] \text{ and } x = \dfrac{y(\pi_2^k - \pi_1^k) + \pi_1^k\upsilon_2^k - \pi_2^k\upsilon_1^k}{\upsilon_2^k - \upsilon_1^k}, \\ +\infty & \text{otherwise,} \end{cases}$$

   and

$$\inf_{y>0} J_k\left(y, \frac{q_k}{x}\right) = \begin{cases} H\left(\dfrac{\pi_2^k\left(\frac{q_k}{x} - \upsilon_1^k\right) + \pi_1^k\left(\upsilon_2^k - \frac{q_k}{x}\right)}{(\pi_2^k - \pi_1^k)(\upsilon_2^k - \upsilon_1^k)}\right) & \text{if } x \in [q/\upsilon_2^k, q/\upsilon_1^k], \\ +\infty & \text{otherwise.} \end{cases}$$

   Consider this system with two sources having parameters: $\alpha^1 = 3/16, \beta^1 = 1/16, \nu_1^1 = 0.00001, \nu_2^1 = 0.5, \pi_1^1 = 0.0072, \pi_2^1 = 0.27$ and $\alpha^2 = 3/16, \beta^2 = 1/16, \nu_1^2 = 0.00008, \nu_2^2 = 0.7, \pi_1^1 = 0.16, \pi_2^1 = 0.22$. These parameters make job arrivals very bursty. We have $\rho = 0.978$ and set $C = 1$.

   Figure 3 is the analogue of Figure 1, capturing the likelihood that given one queue length is long, the other is x times as long. Despite the strong correlation structure in the sources here, the results display the same features.

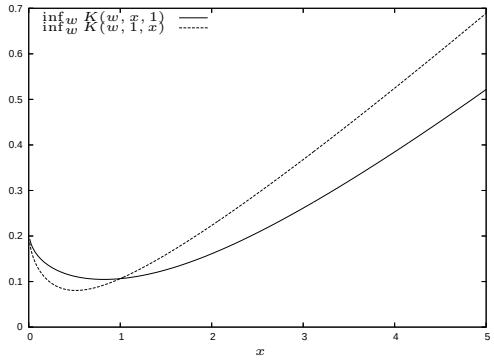

Figure 3: Markov driven job sizes and inter-arrival times.Rate functions for the likelihood one queue-length is x times the other: $\inf_w K(w,x,1)$ and $\inf_w K(w,1,x)$ against x.

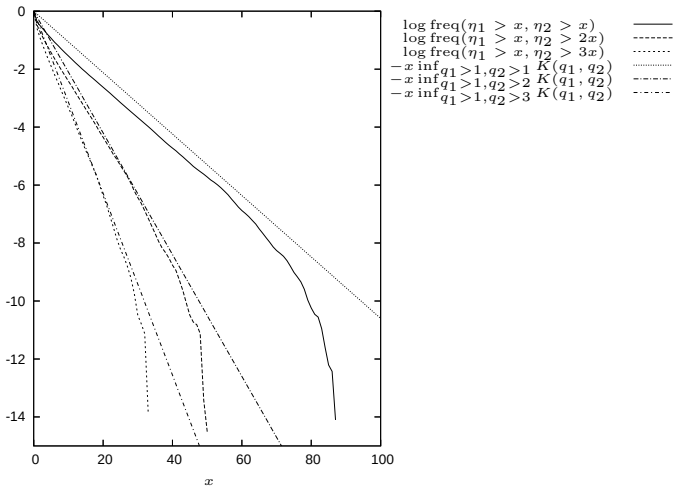

Figure 4: Markov driven job sizes and inter-arrival times. Plot of $-x \inf_{q_1>1,q_2>a} K(q_1,q_2)$, which the theory says should be approximately $\log \mathbb{P}(\eta^1 > x, \eta^2 > ax)$, and observed empirical frequencies from simulation.

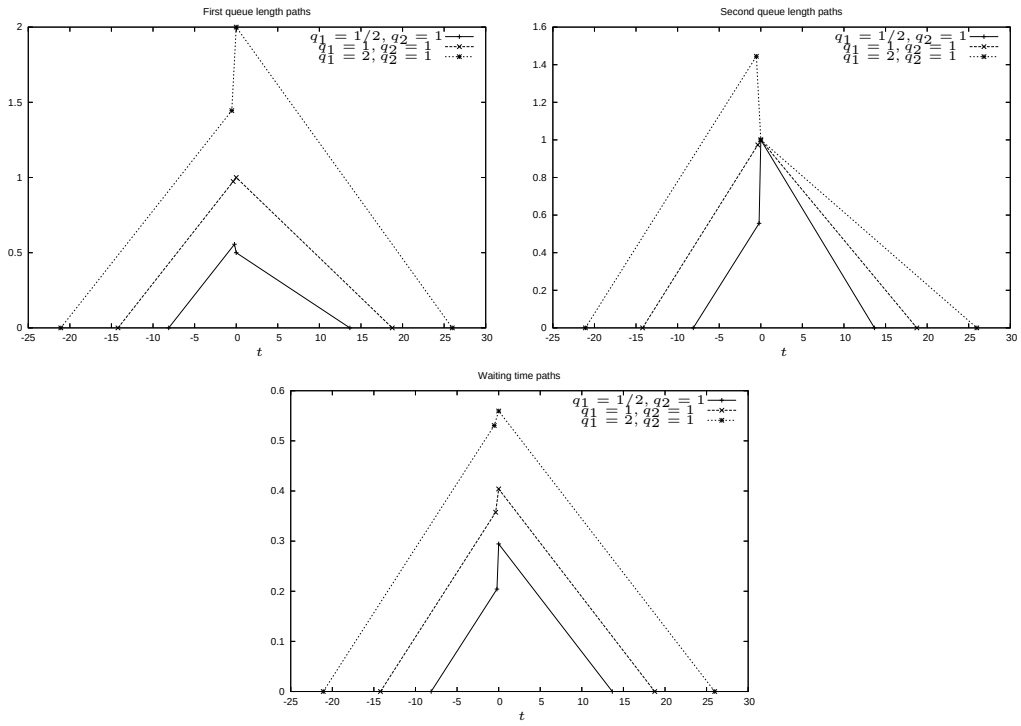

Figure 5: Markov driven job sizes and inter-arrival times. Predicted most likely paths to rare events.

Again we simulated this system with 2 million jobs from each source. We empirically observed the frequency with which $\{\eta^1 > x, \eta^2 > ax\}$ for a range of values of x and $a \in \{1, 2, 3\}$. Corollary 9 predicts that $\log \text{freq}\ \{\eta^1 > x, \eta^2 > ax\}$ should be approximately $-x \inf_{q_1>1, q_2>a} K(q_1, q_2)$ for large x. Figure 4 demonstrates the accuracy of this approximation by plotting these frequencies. Note that for large values of x, the tail of the empirically observed curves drop suddenly. This is due to a finite amount of data from simulation. Running the simulations for longer results in a similar drop off, but for larger x values. Again, we have observed similar output for a range of process statistics.

Figure 5 demonstrates the variability in the nature of the predicted most likely paths for the queue lengths and waiting time, as a function of the conditioning event. Here, for large n, we condition on η^1/n exceeding q_1 and η^2/n exceeding q_2. Note the predicted sharp changes at the time x^*, which can result in either an increase or decrease.

4.3 An application

We return to the question in section 1.4: how can one divide input buffer space between sources to ensure that the likelihood losses occur (any buffer overflows) is minimized?

Consider the case where we have a buffer that can store a maximum of N MTU sized packets. One

must divide up this storage space, where the divided space is allocated on a per-source basis. In other words, we wish to choose $\{\alpha_k, k \in \{1, \ldots, d\}\}$ so that source k can buffer $\alpha_k N$ packets, subject to the constraint that $1 = \sum_{k=1}^{d} \alpha_k$.

Two naive rules for splitting the buffer space would be based on: (1) ratios of the mean arrival rates of the sources; or (2) ratios of the mean inter-arrival times. Rule (1) gives $\alpha_k = (\mu_k/\nu_k)/(\sum_{j=1}^{d}(\mu_j/\nu_j))$, while rule (2) gives $\alpha_k = (1/\nu_k)/(\sum_{j=1}^{d}(1/\nu_j))$, where μ_k and ν_k are defined in assumption 3.

From the large deviations analysis and the comment on finite waiting space after Corollary 9, we know that for each source k, we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\eta^k > n) = -\delta_k,$$

where δ_k is defined in equation (17). By the principle of the largest term (Lemma 1.2.15 [5]),

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\{\eta^1 > \alpha_1 n\} \cup \ldots \cup \{\eta^d > \alpha_d n\}\right) = -\min_{k=1,\ldots,d} \alpha_k \delta_k.$$

Thus, from a large deviation approach, we wish to identify the collection of non-negative real numbers α_k that maximizes the minimum of $\alpha_k \delta_k$, as this would ensure that for large N the likelihood any source is experiencing loss is minimized.

Lemma 11 *We have that*

$$\max\left\{\min_{k=1,\ldots,d} \alpha_k \delta_k : 1 = \sum_{k=1}^{d} \alpha_k\right\} = \left(\sum_{k=1}^{d} \frac{1}{\delta_k}\right)^{-1},$$

where this maximum is obtained with

$$\alpha_k = \frac{1}{\delta_k}\left(\sum_{k=1}^{d} \frac{1}{\delta_k}\right)^{-1}.$$

For the two Markov sources in section 4.2 we chose $N = 100$ and ran simulations for the buffer divided as $(\alpha_1 N, (1-\alpha_1)N)$, for a range of α_1 and observed the percentage of dropped packets (lost jobs) over the course of the simulation. The same traffic traces were used in every simulation. Figure 6 plots these empirically observed values as a function of $\alpha_1/(1-\alpha_1)$. The vertical lines show the $\alpha_1/(1-\alpha_1)$ values suggested by the two naive rules and the large deviation rule. Rule (1) is furthest from minimizing the probability of job loss due to lack of storage space. Rule (2), based on mean inter-arrival times, is better. However the value suggested by the large deviation analysis does best. We have seen similar outcomes for a range of source statistics.

A Proofs

Proof of Theorem 1: cumulative and counting processes sample path LDP.

Proof: Lemma 3 of Duffy and Rodgers-Lee [8] proves that $f : \mathcal{Y}^1 \times \mathcal{Y}^1_\uparrow \mapsto \mathcal{Y}^1 \times \mathcal{Y}^1_\uparrow$, $f(\phi, \psi) = (\phi(\psi^{-1}), \psi^{-1})$, is continuous. As $f(B^k_n, T^k_n) = (A^k_n, N^k_n)$, the result follows from an application of the contraction principle, using the fact that $\psi \mapsto \psi^{-1}$ is one to one in $\mathcal{Y}^1_\uparrow$.

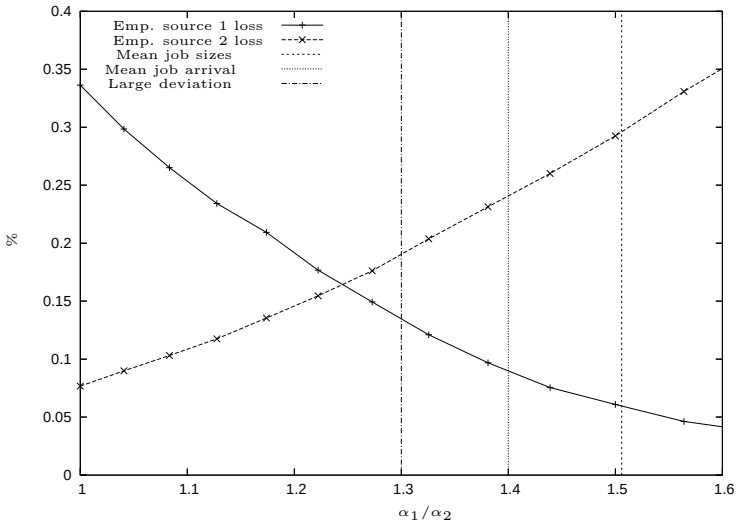

Figure 6: Markov driven job sizes and inter-arrival times. Loss rates for given buffer division ratio.

■

Proof of Lemma 2: exponential equivalence of interlaced and summed sample paths.

Proof: We have that

$$||\hat{A}_n(\cdot) - A_n(\cdot)|| \leq \sup_{t>0} \frac{1}{1+t} \left| \hat{A}_n(t) - \frac{A(nt)}{n} \right| = \sup_{t>0} \frac{1}{1+t} \sum_{l=1}^{d} \left(t - \frac{1}{n} \sum_{i=0}^{N^l(nt)} \tau_i^l \right) \frac{\xi^l_{N^l(nt)+1}}{\tau^l_{N^l(nt)+1}}$$

$$\leq \sup_{t>0} \frac{1}{1+t} \sum_{l=1}^{d} \xi^l_{N^l(nt)+1} \ \leq \sum_{l=1}^{d} \sup_{m\geq 1} \frac{\xi^l_{m+1}}{1+T^l(m)}.$$

As

$$\mathbb{P}\left(\sum_{l=1}^{d} \sup_{m\geq 1} \frac{\xi^l_{m+1}}{1+T^l(m)} > \delta n \right) \leq \max_{k=1,\ldots,d} \mathbb{P}\left(\sup_{m\geq 1} \frac{\xi^l_{m+1}}{1+T^l(m)} > \frac{\delta}{d} n \right),$$

exponential equivalence then follows from assumption 2 and the principle of the largest term (Lemma 1.2.15 [5]).

■

Proof of Lemma 3: equivalence of waiting time and functional representation.

Proof: Let $m > 0$ be given and consider:

$$\sup_{t>0} \left(\frac{A_m(t)}{C} - t \right) \quad = \quad \sup_{t>0} \left(\frac{A(mt)}{Cm} + \left(t - \frac{T^*(m)}{m} \right) \frac{\xi^*_{m+1}}{C\tau^*_{m+1}} - t \right).$$

The supreumum over t must be attained at interlaced arrival times, as the polygonal approximation less service process is linear between job arrivals. Thus recalling the definition of the waiting time, ω, from (1),

$$\sup_{t>0}\left(\frac{A_m(t)}{C}-t\right)=\frac{\omega}{m}.$$

■

Proof of Lemma 5: exponential equivalence of queue lengths and the functional representation.

PROOF: We know that $W(A_m(\cdot))=\omega/m$. Define $\sigma:=\sup\{s:A(s)\le C\omega\}$ and consider

$$\begin{aligned}\sigma_m &:= \sup\left\{s:A_m(s)\le\frac{C\omega}{m}\right\}\\ &= \frac{1}{m}\sup\left\{s:A(s)+\left(s-\sum_{i=0}^{N^*(s)}\tau_i^*\right)\frac{\xi^*_{N^*(s)+1}}{\tau^*_{N^*(s)+1}}\le C\omega\right\}.\end{aligned}$$

As $N_m^k(\cdot)$ is non-decreasing and $N_m^k(t)\in[N(mt)/m,(N(mt)+1)/m)$,

$$Q^k(A_m(\cdot),\vec{N}_m(\cdot))=N_m^k(\sigma_m)\in\left[N_m^k\left(\frac{\sigma}{m}\right)-\frac{1}{m},N_m^k\left(\frac{\sigma}{m}\right)\right)\subset\left[\frac{N^k(\sigma)-1}{m},\frac{N^k(\sigma)+1}{m}\right).$$

Thus

$$\left|Q^k(A_m(\cdot),\vec{N}_m(\cdot))-\frac{\eta^k}{m}\right|\le\frac{1}{m},$$

and $\{Q^k(A_n(\cdot),\vec{N}_n(\cdot))\}$ and $\{\eta^k/n\}$ are exponentially equivalent.

■

Proof of Lemma 6: continuity of the queue length functional.

PROOF: Assume that $(\phi_n,\psi_n)\to(\phi,\psi)$ in $(\mathcal{Y}^1_\mu\cap\mathcal{Y}_\uparrow)\times\mathcal{Y}^1$. Define the positive real valued function $\sigma(\phi):=\sup\{s:\phi(s)\le CW(\phi)\}$. We have that

$$\begin{aligned}|\sigma(\phi_n)-\sigma(\phi)| &\le |\sup\{s:\phi_n(s)\le CW(\phi_n)\}-\sup\{s:\phi(s)\le CW(\phi_n)\}|\\ &\quad+|\sup\{s:\phi(s)\le CW(\phi_n)\}-\sup\{s:\phi(s)\le CW(\phi)\}|\\ &= \left|\phi_n^{-1}(CW(\phi_n))-\phi^{-1}(CW(\phi_n))\right|+\left|\phi^{-1}(CW(\phi_n))-\phi^{-1}(CW(\phi))\right|,\end{aligned}$$

where $\phi_n^{-1}(\phi_n(t))=\phi_n(\phi_n^{-1}(t))=\phi^{-1}(\phi(t))=\phi(\phi^{-1}(t))=t$. As $\phi_n\to\phi$ in $\mathcal{Y}^1_\mu\cap\mathcal{Y}^d_\uparrow$, $\phi_n^{-1}\to\phi^{-1}$ in $\mathcal{Y}^1_{1/\mu}\cap\mathcal{Y}^d_\uparrow$. In particular, ϕ^{-1} is continuous and ϕ_n^{-1} converges to ϕ^{-1} uniformly on compact sets. As $W(\phi_n)\to W(\phi)$, σ is continuous. It remains to show that $\psi_n(\sigma(\phi_n))\to\psi(\sigma(\phi))$, but as $\sigma(\phi_n)\to\sigma(\phi)$ and $\psi_n\to\psi$ uniformly on compact sets, this follows immediately. Thus Q^k is continuous.

■

Proof of Theorem 7: main result for dependent sources.

PROOF: Theorem 1, Lemma 2, Lemma 4 and Lemma 6 show that the map

$$\begin{aligned}(\vec{\phi},\vec{\psi}) &\mapsto (\phi_1(\psi_1^{-1}),\ldots,\phi_d(\psi_d^{-1}),\psi_1^{-1},\ldots,\psi_d^{-1})\\ &\mapsto \left(\sum_{k=1}^{d}\phi_k(\psi_k^{-1}),\psi_1^{-1},\ldots,\psi_d^{-1}\right)\\ &\mapsto \left[W\left(\sum_{k=1}^{d}\phi_k(\psi_k^{-1})\right),Q\left(\sum_{k=1}^{d}\phi_k(\psi_k^{-1}),\psi_1^{-1},\ldots,\psi_d^{-1}\right)\right]\end{aligned}$$

is continuous. Thus an application of the contraction principle, using and the exponential equivalence in Lemma 2 and Lemma 5, gives

$$\begin{aligned}K(w,\vec{q}) &= \inf\left\{I_\infty(\vec{\phi},\vec{\psi}) : W\left(\sum_{k=1}^{d}\phi_k(\psi_k^{-1})\right)=w, Q\left(\sum_{k=1}^{d}\phi_k(\psi_k^{-1}),(\vec{\psi})^{-1}\right)=\vec{q}\right\}\\ &= \inf\left\{J_\infty(\vec{\phi},\vec{\psi}) : W(\phi)=w, Q(\phi,\vec{\psi})=\vec{q},\ \vec{\phi}\in(\mathcal{Y}_{\vec{\kappa}}^d\cap\mathcal{Y}_\uparrow^d),\ \vec{\psi}\in(\mathcal{Y}_{\vec{\nu}^{-1}}^d\cap\mathcal{Y}_\uparrow^d)\right\}\end{aligned}$$

where $\phi=\sum_{k=1}^{d}\phi_k$, $\vec{\kappa}=(\mu_1/\nu_1,\ldots,\mu_d/\nu_d)$ and $\vec{\nu}^{-1}:=(1/\nu_1,\ldots,1/\nu_d)$.

■

Proof of Theorem 8: main result for independent sources.

PROOF: Starting from the result in Theorem 7, we first show under assumptions 3 and 4 that J_∞ is of integral form with integrand

$$J(x_1,\ldots,x_d,n_1,\ldots,n_d)=\sum_{k=1}^{d}J_k(x_k,n_k),$$

where $J_k(x_k,y_k)=y_kI_k\left(x_k/y_k,1/y_k\right)$. Consider $(\phi_1,\ldots,\phi_d,\psi_1,\ldots,\psi_d)$ such that $J(\phi_1,\ldots,\phi_d,\psi_1,\ldots,\psi_d)<$

∞. Then using equation (10)

$$\begin{aligned}
J_\infty\left(\phi_1,\ldots,\phi_d,\psi_1,\ldots,\psi_d\right) &= I_\infty\left(\phi_1(\psi_1^{-1}),\ldots,\phi_d(\psi_d^{-1}),\psi_1^{-1},\ldots,\psi_d^{-1}\right)\\
&= \int_0^\infty\left[\sum_{k=1}^d I_k\left(\frac{\dot\phi_k(\psi^{-1}(s))}{\dot\psi_k(\psi^{-1}(s))},\frac{1}{\dot\psi_k(\psi^{-1}(s))}\right)\right]ds\\
&= \sum_{k=1}^d\left[\int_0^\infty I_k\left(\frac{\dot\phi_k(\psi^{-1}(s))}{\dot\psi_k(\psi^{-1}(s))},\frac{1}{\dot\psi_k(\psi^{-1}(s))}\right)ds\right]\\
&= \sum_{k=1}^d\left[\int_0^\infty I_k\left(\frac{\dot\phi_k(s)}{\dot\psi_k(s)},\frac{1}{\dot\psi_k(s)}\right)\dot\psi_k(s)ds\right]\\
&= \sum_{k=1}^d\left[\int_0^\infty J_k\left(\dot\phi_k(s),\dot\psi_k(s)\right)ds\right]\\
&= \int_0^\infty\left[\sum_{k=1}^d J_k\left(\dot\phi_k(s),\dot\psi_k(s)\right)\right]ds,
\end{aligned}$$

where $J_k(x,y) = yI_k\left(x/y,1/y\right)$ and we used the substitutions $s\mapsto\psi_k(s)$.

Moreover, if I_k is convex, then J_k is convex. This can be seen by considering, for $\alpha\in[0,1]$,

$$J_k(\alpha x_1+(1-\alpha)x_2,\alpha y_1+(1-\alpha)y_2) = (\alpha y_1+(1-\alpha)y_2)I_k\left(\frac{\alpha x_1+(1-\alpha)x_2}{\alpha y_1+(1-\alpha)y_2},\frac{1}{\alpha y_1+(1-\alpha)y_2}\right).$$

Set $\gamma = \alpha y_1/(\alpha y_1+(1-\alpha)y_2)$, note that $\gamma\in[0,1]$ as $y_1,y_2\geq 0$, and thus

$$\frac{\alpha x_1+(1-\alpha)x_2}{\alpha y_1+(1-\alpha)y_2} = \gamma\frac{x_1}{y_1}+(1-\gamma)\frac{x_2}{y_2}\ \text{ and }\ \frac{1}{\alpha y_1+(1-\alpha)y_2} = \gamma\frac{1}{y_1}+(1-\gamma)\frac{1}{y_2}.$$

Hence, using the convexity of $I_k(x,y)$,

$$\begin{aligned}
J_k(\alpha x_1+(1-\alpha)x_2,\alpha y_1+(1-\alpha)y_2) &\leq \alpha y_1 I_1\left(\frac{x_1}{y_1},\frac{1}{y_1}\right)+(1-\alpha)y_2 I_1\left(\frac{x_2}{y_2},\frac{1}{y_2}\right)\\
&= \alpha J_k(x_1,y_1)+(1-\alpha)J_k(x_2,y_2).
\end{aligned}$$

as required. It is also the case that $\inf_x J_k(x,y)$ is a convex function (e.g Theorem 5.3 of [17]) and as L is defined as an inf-convolution of convex functions, it too is convex (e.g. Theorem 5.4 of [17]).

We will see that $K(\vec{q})$ breaks into two parts. Firstly we have the "cost" of having the cumulative arrivals over an interval of length $t-s$ consume the service available over an interval of length t. The second part is the "cost" of having source k produce q_k jobs in an interval of length s while all the sources combine to generate a waiting time w. The rate function identifies the most likely t and s for

which this will happen. For $\vec{\phi} \in \mathcal{Y}^d$ define $\phi = \sum_{k=1}^d \phi_k$, then

$$\begin{aligned}
K(w,\vec{q}) &= \inf_{(\vec{\phi},\vec{\psi})} \left\{ J_\infty(\vec{\phi},\vec{\psi}) : W(\phi) = w, Q^1(\phi,\vec{\psi}) = q_1, \ldots, Q^d(\phi,\vec{\psi}) = q_d \right\} \\
&= \inf_{(\vec{\phi},\vec{\psi})} \{ J_\infty(\vec{\phi},\vec{\psi}) : \sup_t(\phi(t)/C - t) = w, \\
&\qquad \psi_1(\sup\{s : \phi(s) \le Cw\}) = q_1, \ldots, \psi_d(\sup\{s : \phi(s) \le Cw\}) = q_d \} \\
&= \inf_{x \ge 0} \inf_{(\vec{\phi},\vec{\psi})} \left\{ J_\infty(\vec{\phi},\vec{\psi}) : \sup_t(\phi(t) - Ct) = Cw, \phi(x) = Cw, \psi_1(x) = q_1, \ldots, \psi_d(x) = q_d \right\} \\
&\ge \inf_{x \ge 0} \inf_{t \ge x} \inf_{(\vec{\phi},\vec{\psi})} \left\{ J_\infty(\vec{\phi},\vec{\psi}) : \phi(t) - \phi(x) = Ct, \phi(x) = Cw, \psi_1(x) = q_1, \ldots, \psi_d(x) = q_d \right\} \\
&\ge \inf_{x \ge 0} \inf_{t \ge x} \inf_{(\vec{\phi},\vec{\psi})} \left\{ J_\infty(\vec{\phi},\vec{\psi}) : \phi(t) - \phi(x) = Ct, \phi(x) = Cw, \psi_1(x) = q_1, \ldots, \psi_d(x) = q_d \right\} \\
&\ge \inf_{x \ge 0} \inf_{t \ge x} \inf_{(\phi,\vec{\psi})} \left\{ \sum_{k=1}^d \left(\int_0^x J_k(\dot{\phi}_k(s), \dot{\psi}_k(s)) ds \right) + \int_x^t L\left(\dot{\phi}(s)\right) ds : \right. \\
&\qquad \left. \phi(t) - \phi(x) = Ct, \phi(x) = Cw, \psi_1(x) = q_1, \ldots, \psi_d(x) = q_d \right\} \\
&\ge \inf_{x \ge 0} \inf_{t \ge x} \inf_{\vec{y}: \sum_{k=1}^d y_k = Cw} \left\{ \sum_{k=1}^d x J_k\left(\frac{y_k}{x}, \frac{q_k}{x}\right) + (t-x) L\left(\frac{Ct}{t-x}\right) \right\} \qquad (19) \\
&= \inf_{x \ge 0} \inf_{\{\vec{y}: \sum_{k=1}^d y_k = Cw\}} \left\{ \sum_{k=1}^d x J_k\left(\frac{y_k}{x}, \frac{q_k}{x}\right) + \inf_{z \ge 0} z L\left(\frac{C(z+x)}{z}\right) \right\}
\end{aligned}$$

where we are using Jensen's inequality to get (19).

To show the lower bound is obtained we construct a collection of absolutely continuous functions $(\hat{\phi}_1, \ldots, \hat{\phi}_d, \hat{\psi}_1, \ldots, \hat{\psi}_d)$ with $\hat{\phi} := \sum_{k=1}^d \hat{\phi}_k$ that are in the set $\{(\phi, \vec{\psi}) : Q^1(\phi,\vec{\psi}) = q_1, \ldots, Q^d(\phi,\vec{\psi}) = q_d\}$. Moreover, by construction $I(\hat{\phi}_1 \ldots, \hat{\phi}_d, \hat{\psi}_1, \ldots, \hat{\psi}_d)$ will be arbitrarily close to the lower bound. Thus $K(\vec{q})$ has the form given in equation (13).

As all the functions in the infimum are convex and convex functions are continuous on the interior of where they're finite, given $\epsilon > 0$ we can select a points $(x^*, y_1^*, \ldots, y_d^*, z^*) \in (0,\infty)^{d+2}$ and $(a_1^*, \ldots, a_d^*, b_1^*, \ldots, b_d^*) \in (0,\infty)^{2d}$ with $\sum_{k=1}^d a_k^* = C(z^* + x^*)/z^*$ and $\sum_{k=1}^d y_k^* = Cw$, such that

$$\sum_{k=1}^d \left(x^* J_k\left(\frac{y_k^*}{x^*}, \frac{q_k}{x^*}\right)\right) + z^* \left(\sum_{k=1}^d J_k(a_k^*, b_k^*) \right)$$

is no greater than

$$\inf_{x \ge 0} \inf_{\{\vec{y}: \sum_{k=1}^d y_k = Cw\}} \left\{ \sum_{k=1}^d x J_k\left(\frac{y_k}{x}, \frac{q_k}{x}\right) + \inf_{z \ge 0} z L\left(\frac{C(z+x)}{z}\right) \right\} + \epsilon.$$

Let (μ_k, ν_k) be such that $J(\mu_k, \nu_k) = 0$ for each $k \in \{1, \ldots, d\}$. Set $\hat{\phi}(t) = \sum_{k=1}^d \hat{\phi}_k(t)$ and for each $k \in \{1, \ldots, d\}$ define the functions $(\hat{\phi}_1, \ldots, \hat{\phi}_d, \hat{\psi}_1, \ldots, \hat{\psi}_d)$ through their derivatives

$$\frac{d}{dt}\hat{\phi}_k(t) = \begin{cases} y_k^*/x^* & \text{if } t < x^* \\ a_k^* & \text{if } x^* \le t < x^* + z^* \\ \mu_k & \text{if } t \ge x^* + z^* \end{cases}$$

and

$$\frac{d}{dt}\hat{\psi}_k(t) = \begin{cases} q_k/x^* & \text{if } t < x^* \\ b_k^* & \text{if } x^* \le t < x^* + z^* \\ \nu_k & \text{if } t \ge x^* + z^*. \end{cases}$$

We have $(\hat{\phi}_1, \ldots, \hat{\phi}_d, \hat{\psi}_1, \ldots, \hat{\psi}_d) \in (\mathcal{Y}_{\vec{\mu}}^d \cap \mathcal{Y}_{\uparrow}^d) \times (\mathcal{Y}_{\vec{\nu}}^d \cap \mathcal{Y}_{\uparrow}^d)$, $W(\sum_{k=1}^d \hat{\phi}_k) = Cw$, $Q^k(\hat{\phi}, \hat{\psi}_1, \ldots, \hat{\psi}_d) = q_k$ for every $k \in \{1, \ldots, d\}$ and $J_\infty(\hat{\phi}_1, \ldots, \hat{\phi}_d, \hat{\psi}_1, \ldots, \hat{\psi}_d)$ is within ϵ of our lower bound. Thus it is possible to arbitrarily well approximate the lower bound, and the result follows.

To see that $K(w, \vec{q})$ is linear on rays, consider the expression in equation (13) for $K(\alpha w, \alpha\vec{q})$ and make the substitution $x' = x/\alpha$ to get $\alpha K(\vec{q})$.

To get the result for $\{\omega/n\}$, one contracts out by projection to get $K_\omega(w) = \inf_{\vec{q}} K(w, \vec{q})$.

■

Proof of Corollary 9: logarithmic asymptotics for independent sources.

Proof: As K is convex, it is continuous on the interior of the set where it is finite. Thus the result follows applying the large deviation lower bound to $\{(\omega, \eta^1, \ldots, \eta^d) > (nw, nq_1, \ldots, nq_d)\}$ and large deviation upper bound to $\{(\omega, \eta^1, \ldots, \eta^d) \ge (nw, nq_1, \ldots, nq_d)\}$.

■

Proof of Corollary 10: most likely paths for independent sources.

Proof: We appeal to a sub-result of Theorem 3.1 (b) of Lewis, Pfister and Sullivan [13]. This states that if $\{X_n\}$ satisfies the LDP with rate function I and the closed set D is such that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(X_n \in D) = -\inf_{x \in D} I(x) > -\infty, \text{ then } \lim_{n\to\infty} \mathbb{P}(X_n \in G_D \mid X_n \in D) = 1,$$

where G_D is any neighborhood of $\{x \in D : I(x) = \inf_{y \in D} I(y)\}$.

For $\vec{\phi} = (\phi_1, \ldots, \phi_d)$, define $\phi := \sum_{k=1}^d \phi_k$. In order to deduce the corollary, it suffices to show that

$$S_1 := \left\{(\vec{\phi}, \vec{\psi}) \in \mathcal{Y}^{2d} : W(\phi) \ge w, Q^1\left(\phi, \vec{\psi}\right) \ge q_1, \ldots, Q^d\left(\phi, \vec{\psi}\right) \ge q_d \text{ and } J_\infty(\vec{\phi}, \vec{\psi}) = K(\vec{q})\right\}$$

equals

$$S_2 := \left\{(\vec{\phi}, \vec{\psi}) \in \mathcal{Y}^{2d} : (\vec{\phi}, \vec{\psi}) = (\vec{\hat{\phi}}, \vec{\hat{\psi}}) \text{ apart from on a set of Lebesgue measure } 0\right\}.$$

Clearly $S_2 \subset S_1$. Assume $(\vec{\phi}, \vec{\psi}) \in S_1$ Let Ξ be such that

$$\left(\dot{\phi}_1, \ldots, \dot{\psi}_d\right)(s) \ne \left(\frac{d}{dt}\hat{\phi}_1, \ldots, \frac{d}{dt}\hat{\psi}_d\right)(s) \text{ for } s \in \Xi.$$

However $J_\infty(\vec{\phi}, \vec{\psi}) = J_\infty(\vec{\hat{\phi}}, \vec{\hat{\psi}})$ and therefore

$$\int_{x^*+z^*}^{\infty} J\left(\frac{d}{dt}\hat{\phi}_1(s), \ldots, \frac{d}{dt}\hat{\psi}_d(s)\right) ds = 0,$$

as $J(x_1, \ldots, x_d, y_1, \ldots, y_d) = 0$ if and only if $(x_1, \ldots, y_d) = (\mu_1, \ldots, \mu_d, \nu_1, \ldots, \nu_d)$, it must be the case that $\int_{\Xi \cap [z^*+x^*, \infty)} ds = 0$, as otherwise

$$\int_0^{x^*+z^*} J(\dot{\phi}_1(s), \ldots, \dot{\psi}_d(s))ds < \int_0^{x^*+z^*} J\left(\frac{d}{ds}\hat{\phi}_1(s), \ldots, \frac{d}{ds}\hat{\psi}_d(s)\right) ds$$

which would violate equation (19). Thus $\Xi \subset [0, z^* + x^*]$.

As the minimizing points $(x^*, y_1^*, \ldots, y_d^*, z^*) \in (0, \infty)^{d+2}$ and $(a_1^*, \ldots, a_d^*, b_1^*, \ldots, b_d^*) \in (0, \infty)^{2d}$ are unique, by Jensen's inequality it must be the case that

$$\phi_k(x^*) = y_k^*,\ \phi_k(x^* + z^*) = y_k^* + z^* a_k,\ \psi_k(x^*) = q_k,\ \text{ and } \psi_k(x^* + z^*) = q_k + z^* b_k.$$

Again by Jensen's inequality, given this constraint, $J(\vec{\phi}, \vec{\psi}) = K(\vec{q})$ only if $(\vec{\phi}, \vec{\psi}) = (\hat{\vec{\phi}}, \hat{\vec{\psi}})$ almost everywhere, so that $S_1 = S_2$.

■

References

[1] V. Anantharam, *How large delays build up in a $GI/G/1$ queue*, Queueing Systems Theory Appl. **5** (1989), no. 4, 345–367.

[2] S. Asmussen and J. F. Collamore, *Exact asymptotics for a large deviations problem for the $GI/G/1$ queue*, Markov Process. Related Fields **5** (1999), no. 4, 451–476.

[3] S. Aspandiiarov and E. A. Pechersky, *A large deviations problem for compound Poisson processes in queueing theory*, Markov Process. Related Fields **3** (1997), no. 3, 333–366.

[4] A. Dembo and T. Zajic, *Large deviations: from empirical mean and measure to partial sums process*, Stochastic Process. Appl. **57** (1995), no. 2, 191–224.

[5] A. Dembo and O. Zeitouni, *Large deviation techniques and applications*, Springer, 1998.

[6] K. Duffy, J. T. Lewis, and W. G. Sullivan, *Logarithmic asymptotics for the supremum of a stochastic process*, Ann. Appl. Probab. **13** (2003), no. 2, 430–445.

[7] K. Duffy and A. P. Metcalfe, *The large deviations of estimating rate functions*, J. Appl. Probab. **42** (2005), no. 1, 267–274.

[8] K. Duffy and M. Rodgers-Lee, *Some useful functions for functional large deviations*, Stoch. Stoch. Rep. **76** (2004), no. 3, 267–279.

[9] K. Duffy and W. G. Sullivan, *Logarithmic asymptotics for unserved messages at a FIFO*, Markov Process. Related Fields **10** (2004), no. 1, 175–189.

[10] A. Ganesh, N. O'Connell, and D. Wischik, *Big queues*, Lecture Notes in Mathematics, vol. 1838, Springer-Verlag, Berlin, 2004.

[11] A. J. Ganesh and N. O'Connell, *A large deviation principle with queueing applications*, Stoch. Stoch. Rep. **73** (2002), no. 1-2, 25–35.

[12] P. W. Glynn and W. Whitt, *Logarithmic asymptotics for steady-state tail probabilities in a single-server queue*, J. Appl. Probab. **31A** (1994), 131–156, Studies in applied probability.

[13] J. T. Lewis, C.-E. Pfister, and W. G. Sullivan, *Entropy, concentration of probability and conditional limit theorems*, Markov Process. Related Fields **1** (1995), no. 3, 319–386.

[14] R. M. Loynes, *The stability of a queue with non-independent interarrival and service times*, Proc. Cambridge Philos. Soc. **58** (1962), 497–520.

[15] T. Mikosch and A. V. Nagaev, *Large deviations of heavy-tailed sums with applications in insurance*, Extremes **1** (1998), no. 1, 81–110.

[16] A. A. Mogul′skiĭ, *Large deviations in the space $C(0, 1)$ for sums that are defined on a finite Markov chain*, Sibirsk. Mat. Ž. **15** (1974), 61–75, 237.

[17] R. Tyrrell Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970.

[18] J. S. Sadowsky, *The probability of large queue lengths and waiting times in a heterogeneous multiserver queue. II. Positive recurrence and logarithmic limits*, Adv. in Appl. Probab. **27** (1995), no. 2, 567–583.

[19] J. S. Sadowsky and W. Szpankowski, *The probability of large queue lengths and waiting times in a heterogeneous multiserver queue. I. Tight limits*, Adv. in Appl. Probab. **27** (1995), no. 2, 532–566.

[20] S. R. S. Varadhan, *Asymptotic probabilities and differential equations*, Comm. Pure Appl. Math. **19** (1966), 261–286.