DECOUPLING BANDWIDTHS FOR NETWORKS:
A Decomposition Approach to Resource Managment[*]

G. DE VECIANA[**], C. COURCOUBETIS[†] AND J. WALRAND[‡]

[**]*Department of Electrical and Computer Engineering*
University of Texas at Austin
Austin, Texas 78712

[‡]*Department of Electrical Engineering and Computer Sciences*
University of California at Berkeley
Berkeley CA 94720

[†]*University of Crete, Heraklion, Crete*

UCB/ERL Technical Report M93/50

June 28, 1993; Revised January 1, 1995

Abstract

We consider large buffer asymptotics for feed-forward *networks* of discrete-time queues with deterministic service rate shared by multiple classes of streams subject to work conserving service policies. First we review the concept of *effective bandwidths* for traffic streams sharing a common buffer subject to subject to tail constraints on the workload distribution. Next, we obtain the effective bandwidth of the departure process from such a queue, proving that in fact the effective bandwidth of the output is at worst equal to that of the input, and depending on the service rate, strictly less than that of the input. We then define the notion of a *decoupling bandwidth* and the associated constraints, guaranteeing that asymptotics within the network are decoupled.

These results provide a framework for call admission schemes which are sensitive to constraints on the tail distribution of the workload or approximate cell loss probabilities. Our results require relatively weak assumptions on both the traffic streams and service policies. We consider the problem of "optimal" traffic shaping (via buffering) subject to a loss constraint. Finally, we discuss our results in the context of resource management for ATM networks.

1 Introduction

An important open problem in the context of BISDN/ATM is that of designing appropriate resource management schemes comprising call admission, routing, and network planning for a heterogeneous collection of users requiring multiple qualities of service. The difficulty in solving this problem, relative to traditional (circuit-switched) telephone networks, lies in the multiplexing of heterogeneous packetized traffic streams and messages via switches and communication links. In order for streams to share resources, one must guard against traffic fluctuations by inserting buffers.

[*] The research of G. de Veciana was partially supported by the National Science Foundation under Grant NCR-9409722. Part of this work was presented in Toronto at Infocom '94

To ease the task of managing such a network it is desirable to obtain an equivalent circuit-switched model. For example, suppose a collection of sources, n_j of type $j \in J$, which require a bandwidth α_j, share a link with capacity c. One can easily check for available bandwidth by considering whether

$$\sum_{j \in J} n_j \alpha_j \leq c.$$

This approach, extended to a network, is akin to traditional telephone systems where a connection is set up if indeed physical resources are available to link the source to the destination. Unfortunately, the interaction of multiple types of traffic and resources in networks is typically neither linear in the number of sources nor decoupled across the different types of streams.

There exists, however, a result for the case of heterogeneous streams sharing a *single buffered link* for which *effective bandwidths* and the accompanying linear constraint can be found such that an asymptotic constraint on the tail distribution of the buffers' workload is guaranteed (see §2). The goal herein is to investigate this idea for a *network* of queues. In this paper we consider the input/output map for the effective bandwidths of streams sharing a queue. We will show explicitly how the reduction in effective bandwidth due to buffering depends on the release rate. We then consider the resource management problem for a multi-class feed-forward network via the notion of *decoupling bandwidths*. Our results suggest that when the queues' service rates are selected appropriately the asymptotics for queues in the network reduce to that of the single buffer case, and thus, by way of effective bandwidths, a workable circuit-switched model is obtained.

An example with similar characteristics, is that of a Jackson network for which the steady state distribution is in fact product-form [28]. In this case the queue length distribution at queue i is geometric with a parameter $\rho_i = \lambda_i/\mu_i$ - the ratio of the aggregate arrival intensity λ_i to the service rate μ_i of the queue. In order to guarantee δ-constraints on the tail distributions of the queue lengths

$$\mathbb{P}(Q_i \geq B) = \rho_i^B \leq \delta^B,$$

we require that $\lambda_i \leq \mu_i \delta$ for all nodes. A new traffic stream with mean rate λ_{new} can be admitted, and still comply with the imposed tail constraints, if the additional traffic along its route remains within the prescribed intervals, i.e., $\lambda_i + \lambda_{new} \leq \mu_i \delta$ for all nodes i along the new call's route.

In this paper we extend this scenario to a feed-forward network with multiple classes of traffic streams, where the queues have deterministic service rates and work conserving service policies. The intuitive picture for our result is as follows: Consider a large accumulation of customers (packets) in a particular queue deep in the network. In a stable system this event is likely to be due to an increase in the empirical arrival rate of the traffic streams sharing that queue. When the asymptotics are "decoupled" these deviations from the mean rate are such that conditional on this overflow event, other queues shared by these streams are invisible, i.e., they will not be accumulating traffic. The likelihood of deviations in the empirical arrival rate for these traffic streams can then be computed at the network edge where the statistics are assumed to be known.

To our knowledge this is the first study of the effective bandwidth idea for large buffer asymptotics in networks. We note however, that previous and subsequent work by Chang et al. [4, 7], sheds further light upon this work. In particular the input-output characteristics of queues were first considered in [4] where an upper bound for the aggregate output is derived. Furthermore, in [5] a heuristic set of non-linear equations are proposed to investigate quick simulation of network asymptotics. These ideas were extended in [**?**] where an optimization problem is proposed as a means to study tail distributions in in-tree networks. Yaron and Sidi [30] also present exponential upper bounds, suitable for establishing exponential stability of such networks when streams satisfy exponentially bounded burstiness conditions. The results presented in this paper have been furthered by both O'Connell [24] and Chang and Zajic [6] both of which investigate in more detail the nature of the departure process, via large deviations for sample path processes. The key contribution of this paper is the observation that the asymptotic behavior of a rather general class of networks and work conserving policies can be simplified subject to decoupling constraints. This validates resource management schemes based on nodal or virtual circuit isolation - for a recent discussion see Towsley [27].

The rest of this paper is organized as follows. In §2 we briefly introduce concepts and notation related to the theory of large deviations and review the notion of effective bandwidths for a single buffered queue. In §3 we consider the large deviations of the departure process from a queue. Having identified the properties of the departure processes, we consider some examples of traffic shaping via buffering in §3.1. In §3.2 we consider queues in tandem. We turn to more general aspects of resource management for networks in §4. We argue therein that decoupling constraints may be naturally satisfied by future ATM networks. A summary of our results and conclusions can be found in §5.

2 Single buffer asymptotics via large deviations

In this section we state the effective bandwidth result for a multi-class discrete-time queue subject to constraints on the tail probability of the buffer occupancy. This discussion is based on an earlier paper [11] which reviewed and extended some of the results in Kelly [20]. The large deviations techniques we use here were inspired by a heuristic of Borovkov, see Walrand and Parekh [28, 26], and the results of Kesidis *et al.* [21] and Chang [4]. Note also the studies of Whitt [29] and Duffield *et al.* [14] which present several interesting results such as asymptotics for self-similar traffic. There exists much related work in this field. Notably, effective bandwidth results for Markov fluid sources were obtained via spectral expansions by Gibbens and Hunt [17] and Elwalid and Mitra [15]. In addition, early work on this topic can be found in Hui [19], and Guérin *et al.* [18].

2.1 Large deviations

We begin by reviewing the statement and possible requirements for large deviation results to hold. For a complete reference on the subject see Dembo and Zeitouni [12]. A sequence of measures $\{\mu_n\}$, on $\mathbb{R}$, will satisfy a Large Deviation Principle (LDP) with *good rate function*, $I(\cdot)$, if for every closed set F,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mu_n(F) \le -\inf_{x\in F} I(x),$$

and for every open set G,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \ge -\inf_{x\in G} I(x),$$

and $\{x : I(x) \le \alpha\}$ is compact for $\alpha < \infty$. We only consider the setting where $\{\mu_n\}$ denote the distributions of the partial sums $n^{-1}S_n^X = n^{-1}\sum_{i=1}^n X_n,\ \ n > 0$, for a sequence of real-valued random variables $\{X_n\}$. We then say that $\{X_n\}$ satisfies an LDP with good rate function $I(\cdot)$. In particular we will be using simpler bounds of the type

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n^X \ge n\alpha) \le -\inf_{x\ge\alpha} I(x) \qquad \text{and} \qquad \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n^X > n\alpha) \ge -\inf_{x>\alpha} I(x).$$

Below we briefly discuss when such bounds do indeed hold.

For example, the Gärtner-Ellis Theorem establishes the existence of an LDP with convex good rate function for a large class of sources. The requirements are that:

1. The limits $\Lambda(\theta) \overset{\triangle}{=} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\exp[\theta S_n^X]$ exist (possibly infinite) for all $\theta \in \mathbb{R}$;
2. The origin is in the interior, D_Λ^o, of the *effective domain* $D_\Lambda \overset{\triangle}{=} \{\theta : \Lambda(\theta) < \infty\}$ of $\Lambda(\cdot)$;
3. $\Lambda(\cdot)$ is differentiable throughout D_Λ^o and *steep*, i.e., $\lim_{n\to\infty} |\frac{d\Lambda(\theta_n)}{d\theta}| = \infty$ whenever $\{\theta_n\}$ is a sequence in D_Λ^o, converging to a boundary point of D_Λ^o.

Under conditions 1-3 an LDP holds with the good rate function given by the convex dual $\Lambda^*(\cdot)$ of $\Lambda(\cdot)$:

$$I(x) = \Lambda^*(x) = \sup_\theta [\theta x - \Lambda(\theta)].$$

This result applies to i.i.d. sequences with $\mathbb{E}e^{\theta X_1} < \infty$ for all θ, which corresponds to the original large deviation estimate of Cramér. The result also applies to sequences with weak dependencies.

Further cases where LDPs hold can be found in [12]. For example, coordinate functions of Markov chains satisfying strong uniformity conditions on the transition kernel and tail will satisfy an LDP, see [13]. In this case, the rate function can usually be interpreted in terms of the relative entropy rate of a deviant Markov chain with respect to the original process. For stationary sequences satisfying appropriate mixing and tail conditions similar results hold, see [3].

2.2 Effective bandwidths

Theorem 2.1 [See [11] or [21, 4]] *Let $\{X_n\}$ be a stationary ergodic process with $\mathbb{E}X_1 < 0$, which, either satisfies an LDP with convex good rate function $I(\cdot)$ such that for all $\theta < \infty$*

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp[\theta S_n^X] < \infty,$$

and where $\Lambda^*(\cdot)$ is strictly convex, or satisfies the requirements for the Gärtner-Ellis Theorem[1]*. Then the Lindley process*

$$W_{n+1} = [W_n + X_n]^+,$$

has a stationary distribution, say that of a random variable W, and for $\delta > 0$,

$$\Lambda(\delta) \le 0 \iff \lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(W > B) \le -\delta.$$

Theorem 2.1 can be used to establish the following corollary, by letting $X_n = A_{n+1} - c$, where $A_n = \sum A_n^j$ denotes aggregate arrivals from multiple streams on the n^{th} slot to a queue with service rate c.

Corollary 2.1 *Consider a collection of independent sources, n_j of each type $j \in J$, with discrete-time arrival processes $\{A_n^j\}$, each satisfying the conditions in Theorem 2.1 where*

$$\Lambda_j(\delta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp[\delta S_n^{A^j}].$$

Suppose they share a deterministic buffer with any work conserving service policy at rate c. Then the following effective bandwidth result holds:

$$\sum_{j\in J} n_j \alpha_j(\delta) \le c \iff \lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(W \ge B) \le -\delta,$$

where $\alpha_j(\delta) = \Lambda_j(\delta)/\delta$ and where W denotes the stationary workload.

We will say that a buffer satisfying such a constraint satisfies a δ-constraint, to be used and interpreted as a first order performance guarantee on nodal overflows. The usefulness of this result is predicated on being able to compute or estimate (possibly on-line) the effective bandwidth of a source. For a summary of some analytical formulae see Kesidis et al. [21]. These include the usual i.i.d. sources, as well as Markov modulated fluids or Poisson processes and Gaussian processes. We further note a study by de Veciana and Kesidis [10] exhibiting effective bandwidth results for buffers sharing bandwidth via a generalized processor sharing policy. This is a possible step towards a managing networks which guarantee multiple qualities of service.

3 Effective and decoupling bandwidths for departures from a queue

From here on, in addressing the large deviations characteristics of networks we make the following assumption:

[1] Note that the Gärtner-Ellis Theorem does not require finite log-moment generating functions.

Assumption 3.1 *We assume that all traffic processes, entering sharing and leaving our network, satisfy pathwise LDPs.*

Establishing the conditions under which this assumption holds for multi-class networks, with the various service disciplines currently under consideration for ATM networks is at this point an open and difficult problem. Discuss special cases: where we know what is true. For FCFS networks with Bernoulli splitting the results in provide a sound basis for As even the stability of general multi-class networks is an issue we will not attempt to address the existance of LDP result. On the other hand we focus on qualititative charactersitics in particular decoupling that will hold when the above assumption is satisfied.

The above assumption simplifies our work significantly as we need only identify the large deviation rate function of the different streams rather than prove the existence of the LDPs. We begin this section with a key lemma that shows the simple idea we will use repeatedly.

Lemma 3.1 *Let $\{D_n\}$ be a stationary departure process from a discrete-time queue with deterministic service rate c satisfying A-3.1, with with convex good rate function (?) $\Lambda_D^*(\cdot)$. Suppose we can show that*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n^D \geq n\alpha) \leq - \inf_{x\geq\alpha} I(x) \qquad \textit{and} \qquad \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n^D > n\alpha) \geq - \inf_{x>\alpha} I(x)$$

for some function $I(x)$ and $\alpha \geq \mathbb{E}D_1$. Then the limits

$$\Lambda_D(\delta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\exp[\delta S_n^D] < \infty,$$

exist and $\Lambda_D^*(\alpha) = I(\alpha)$ for $\alpha \geq \mathbb{E}D_1$ whence for $\delta \geq 0$ we have

$$\Lambda_D(\delta) = \sup_{\alpha\geq\mathbb{E}D_1} [\alpha\delta - I(\alpha)].$$

Proof: The service rate is c, so $S_n^D \leq nc$, and it follows that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\exp[\delta S_n^D] < \infty.$$

Now applying Theorem 4.5.10 in [12], we can conclude that the limit $\Lambda_D(\delta)$ exists, is finite, and since the rate function is unique $\Lambda_D^*(\alpha) = I(\alpha)$ for $\alpha \geq \mathbb{E}D_1$. Thus for $\delta \geq 0$ we have

$$\Lambda_D(\delta) = \sup_\alpha [\alpha\delta - \Lambda_D^*(\alpha)] = \sup_{\alpha\geq\mathbb{E}D_1} [\alpha\delta - \Lambda_D^*(\alpha)] = \sup_{\alpha\geq\mathbb{E}D_1} [\alpha\delta - I(\alpha)], \tag{1}$$

where the second equality follows from the fact that $\delta \geq 0$ and $\Lambda^*(\cdot)$ is strictly convex and non-negative on the set $[\mathbb{E}D_1, c]$, with $\Lambda^*(m) = 0$. □

Our approach will be to partially identify the rate function $\Lambda_D^*(\cdot)$ from which by way of the previous lemma we can partially determine $\Lambda_D(\delta)$. This in turn permits us to study the output's effective bandwidth $\alpha_D(\delta) = \Lambda_D(\delta)/\delta$.

Theorem 3.1 *Let $\{A_n, D_n\}$ be stationary and ergodic arrival and departure processes satisfying A-3.1 of a stable discrete-time queue with service rate c, i.e. $\mathbb{E}A_1 < c$. the arrival process is bounded and satisfies the Gärtner-Ellis Theorem; so in particular for all $\theta < \infty$ the limits*

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\exp[\theta S_n^A] < \infty,$$

exist and $\Lambda^*(\cdot)$ is strictly convex. Then the Lindley process

$$W_{n+1} = [W_n + A_{n+1} - c]^+$$

has a stationary distribution, say that of a random variable W, and when the associated departure process $\{D_n\}$ satisfies an LDP with a convex good rate function it is given by $\Lambda^*(\cdot)$ on $[\mathbb{E}A_1, c]$ and infinite on $[0, c]^c$.

We state the theorem subject to these conditions in order to avoid technical details, however, they are much more restrictive than necessary. Our proof relies on a large deviations result for the sample path process shown in Dembo and Zajic [12]. This result holds for a rather general class of Markov or mixing processes satisfying the technical assumptions discussed therein. In particular the arrivals per time slot need not be bounded, though this will typically be the case in practice, and the Gärtner-Ellis Theorem alone is not sufficient to prove these results.

An sketch of the proof can be found in the appendix at the end of this paper. The intuition is however quite clear, that is, in order to get a given empirical rate at the output, the input traffic must provide that flow, whence the rate functions are equal. Now having identified the rate function of $\{D_n\}$ we determine the effective bandwidth of the output.

Corollary 3.1 *Let $\{A_n\}$ be a discrete-time arrival process with effective bandwidth $\alpha(\delta) = \Lambda(\delta)/\delta$ and rate function $\Lambda^*(\cdot)$ entering a discrete-time queue with service rate c and satisfying the conditions of Theorem 3.1. Then the effective bandwidth $\alpha_D(\delta)$ of the departure process $\{D_n\}$ satisfying an LDP with convex good rate function is given by*

$$\alpha_D(\delta) = \begin{cases} \alpha(\delta) & \text{if } \alpha^*(\delta) \le c, \\ c - \frac{1}{\delta}\,\Lambda^*(c) & \text{otherwise}, \end{cases}$$

where $\alpha^*(\delta)$ is defined implicitly through the convex duality relationship

$$\Lambda(\delta) = \sup_{\alpha}[\alpha\delta - \Lambda^*(\alpha)] = \alpha^*(\delta)\delta - \Lambda^*(\alpha^*(\delta)) \quad \text{or simply} \quad \alpha^*(\delta) = \left[\frac{d\Lambda^*}{d\alpha}\right]^{-1}(\delta) = \frac{d}{d\delta}\Lambda(\delta).$$

Proof: By Theorem 3.1 the rate function of the departure process is

$$\Lambda_D^*(\alpha) = \begin{cases} \Lambda^*(\alpha) & \text{if } \alpha \in [\mathbb{E}A_1, c], \\ \infty & \text{if } \alpha \in [0,c]^c. \end{cases}$$

The effective bandwidth of the output stream is given by $\alpha_D(\delta) = \Lambda_D(\delta)/\delta,\ \delta \ge 0$, so using Lemma 3.1 we have

$$\begin{aligned} \Lambda_D(\delta) &= \sup_{\alpha \ge \mathbb{E}A_1}[\alpha\delta - \Lambda_D^*(\alpha)] = \sup_{c \ge \alpha \ge \mathbb{E}A_1}[\alpha\delta - \Lambda^*(\alpha)], \\ &= \begin{cases} \Lambda(\delta) & \text{if } \alpha^*(\delta) \le c, \\ c\delta - \Lambda^*(c) & \text{otherwise.} \end{cases} \end{aligned}$$

□

Definition 3.1 *Referring to Corollary 3.1, we will call $\alpha^*(\delta)$ the* **decoupling bandwidth** *of a traffic stream. For the single buffer case the decoupling constraint $\alpha^*(\delta) \le c$ guarantees that the effective bandwidth of the output is the same as that of the input for the given δ-constraint.*

The intuition is as follows: When the service rate is very large the input and output traffic streams have the same characteristics. However as the service rate is reduced, the queueing delays incurred by the traffic and the deterministic release rate result in a reduction of the effective bandwidth of the departure process, see §3.1 for examples.

Fact 3.1 *If the stationary arrival process has a bounded "sustainable" peak arrival rate*[2]*, that is $\alpha(\infty) \stackrel{\triangle}{=} \lim_{\delta\to\infty}\alpha(\delta) < \infty$, then the decoupling bandwidth satisfies the following inequality:*

[2] Some care is needed in defining the notion of peak arrival rate for a discrete-time arrival process $\{A_n\}$. Drawing on the notation and ideas in Lemma 3.6 of [4] we define the "sustainable" peak arrival rate to be $\lim_{n\to\infty} n^{-1}\|S_n^A\|_\infty$ where $\|S_n^A\|_\infty = \inf\{x : \mathbb{P}(S_n^A > x) = 0\}$. An argument similar to that the aforementioned Lemma 3.6 shows that $\alpha(\infty) = \lim_{n\to\infty} n^{-1}\|S_n^A\|_\infty$. This is the "sustainable" peak arrival rate, in the sense that it will be the largest rate that can be achieved with non-zero probability for an arbitrary number of slots. For standard sources i.e., i.i.d. or Markov modulated sources, this definition matches the intuitive notion of peak arrivals per slot, however, in general this need not be so.

$$\alpha(\infty) \geq \alpha^*(\delta) \geq \alpha(\delta). \tag{2}$$

Proof: By definition $\alpha^*(\delta)$ satisfies $\Lambda(\delta) = \alpha^*(\delta)\delta - \Lambda^*(\alpha^*(\delta))$, so by rearranging terms we get

$$\alpha^*(\delta) = \frac{\Lambda(\delta)}{\delta} + \frac{\Lambda^*(\alpha^*(\delta))}{\delta} \geq \alpha(\delta),$$

since the rate function $\Lambda^*(\cdot)$ is non-negative. Recall that $\Lambda^*(\alpha) = \sup_\delta[\alpha\delta - \Lambda(\delta)]$; since we assumed that $\lim_{\delta\to\infty} \Lambda(\delta)/\delta = \alpha(\infty) < \infty$, if $\alpha > \alpha(\infty)$ then $\Lambda^*(\alpha) = \infty$. Thus suppose $\alpha^*(\delta) > \alpha(\infty)$ then the defining identity for $\alpha^*(\delta)$ would give $\Lambda(\delta) = -\infty$, which contradicts the fact that $\Lambda(\delta) \geq 0$. We conclude that $\alpha(\infty) \geq \alpha^*(\delta)$. □

The decoupling bandwidth $\alpha^*(\delta)$ is usually larger than the effective bandwidth $\alpha(\delta)$ of the source. While this is clear from the equations above, an intuitive interpretation sheds some light on these definitions. The effective bandwidth of a traffic stream can be interpreted as the minimum service rate required for a buffered link to satisfy a δ-constraint on the workload's distribution. Using LDPs for sample path processes, one can in fact show that the decoupling bandwidth corresponds to the most likely deviant empirical rate which the arrival traffic will sustain (in excess of the minimal service rate $\alpha(\delta)$) in order to accumulate a large amount of work in the queue. For example, consider a queue with mean arrival rate m and service rate c. Intuitively $c - m$ is the mean relaxation rate at which a large queued accumulation settles down. Suppose traffic builds up as if relaxations were reversed in time, then the deviant traffic rate offered by the source would have to be $2c - m$. Now, let the service rate for such a queue be the effective bandwidth of the arrival process, i.e., $c = \alpha(\delta)$. Then the build up rate, or decoupling bandwidth should roughly be $\alpha^*(\delta) \approx 2\alpha(\delta) - m$. This heuristic, based on a similar (exact) argument for M/M/1 queues, holds for a discrete-time queue with Gaussian arrivals (consider the examples in §3.1), and is approximately true for M/D/1 queues under heavy loads, see Frater [16].

As a further corollary to Theorem 3.1, consider the scenario in which the arrival process is an aggregate of independent traffic streams.

Corollary 3.2 *Consider a collection of independent sources, n_j of each type $j \in J$, with discrete-time arrival processes $\{A_n^j\}$ and aggregate departure process $\{D_n\}$ each satisfying the conditions in Theorem 3.1. Suppose the streams share a deterministic buffer according to a work conserving service policy with rate c. The rate function of the departure process is then given by*

$$\Lambda_D^*(\alpha) = \inf_{\sum_{j\in J} n_j\alpha_j = \alpha} \sum_{j\in J} n_j \Lambda_j^*(\alpha_j), \tag{3}$$

on the set $[\mathbb{E}A_1, c]$ and infinite on $[0,c]^c$. Moreover the effective bandwidth of the output traffic stream, $\alpha_D(\delta)$, is given by

$$\alpha_D(\delta) = \begin{cases} \sum_{j\in J} n_j\alpha_j(\delta) & \textit{if} \quad \sum_{j\in J} n_j\alpha_j^*(\delta) \leq c, \\ c - \frac{1}{\delta}\inf_{\sum_{j\in J} n_j\alpha_j = c} \sum_{j\in J} n_j\Lambda_j^*(\alpha_j) & \textit{otherwise.} \end{cases}$$

Thus, the decoupling constraint

$$\sum_{j\in J} n_j\alpha_j^*(\delta) \leq c, \tag{4}$$

is a sufficient and necessary condition for the effective bandwidth of the output to equal that of the input stream; otherwise, it is reduced and increases hyperbolically in δ to the service rate c.

Proof: In order to use Theorem 3.1 we begin by identifying the rate function for the arrival process. The latter is an aggregate sum of independent sources. Thus Eq. 3 follows from the contraction principle and the rate functions' convexity, see Dembo and Zeitouni [12, page 110].

The result follows from Corollary 3.1 if we can show that the decoupling bandwidth of the aggregate arrivals is indeed additive. Let $\Lambda(\cdot)$ denote the log moment generating function of the aggregate arrival stream. The decoupling bandwidth of the aggregate $\alpha^*(\delta)$ is defined implicitly by

$$\begin{aligned}\Lambda(\delta) &= \sup_{\alpha}\left[\alpha\delta - \inf_{\sum_{j\in J} n_j\alpha_j=\alpha}\sum_{j\in J} n_j\Lambda_j^*(\alpha_j)\right]\\ &= \alpha^*(\delta)\delta - \inf_{\sum_{j\in J} n_j\alpha_j=\alpha^*(\delta)}\sum_{j\in J} n_j\Lambda_j^*(\alpha_j).\end{aligned}$$

Alternatively we can compute

$$\begin{aligned}\Lambda(\delta) &= \sup_{\{\alpha,\alpha_j,\forall j\in J:\ \sum_{j\in J} n_j\alpha_j=\alpha\}}\left[\alpha\delta - \sum_{j\in J} n_j\Lambda_j^*(\alpha_j)\right]\\ &= \sum_{j\in J} n_j \sup_{\alpha_j}[\alpha_j\delta - \Lambda_j^*(\alpha_j)]\\ &= \sum_{j\in J} n_j\alpha_j^*(\delta)\delta - \sum_{j\in J} n_j\Lambda_j^*(\alpha_j^*(\delta)).\end{aligned}$$

Comparing the above expressions we note that $\alpha^*(\delta)=\sum_{j\in J} n_j\alpha_j^*(\delta)$ so the result follows from Corollary 3.1. □

In the sequel we consider some consequences of these results, however, first we add a final corollary. While in Corollaries 3.1 and 3.2 we examined the effective bandwidth for the *aggregate* departure process, it is of great interest to determine that of *individual* streams at the output of such a queue since streams may eventually follow different routes. Below we focus on a queue with two input streams, where the first is intended to represent a particular traffic stream of interest and the second represents the aggregation of other traffic sharing the queue.

Corollary 3.3 *Consider two independent discrete-time arrival processes $\{A_n^1, A_n^2\}$ satisfying the conditions in Theorem 3.1. Suppose they share a deterministic buffer with service rate c via a work conserving service policy and the system is not only stable but satisfies the effective bandwidth constraint*

$$\alpha_1(\delta)+\alpha_2(\delta) < c.$$

If the departures corresponding to the first stream $\{D_n^1\}$ satisfy an LDP with convex good rate function it is given by

$$\Lambda_{D^1}^*(\alpha) = \Lambda_1^*(\alpha) \tag{5}$$

on the set $[\mathbb{E}A_1^1, \min[\alpha_1^*(\delta), c-\mathbb{E}A_1^2]]$ and a sufficient condition for the effective bandwidth of the departures of the first stream to equal that of the input (at δ or below) is that

$$\alpha_1^*(\delta)+\alpha_2(0) = \alpha_1^*(\delta)+\mathbb{E}A_1^2 \le c. \tag{6}$$

An analogous result holds for the second departure stream.

The proof of this corollary has been relegated to the appendix. The intuition for this result is as follows. Suppose that stream 2 offers an arrival sample process close to its mean traffic rate $\mathbb{E}A_1^2$ and has service priority. In this case stream 1 sees an "effective" service rate of $c-\mathbb{E}A_1^2$. In order to observe an increased empirical output rate (say α) for stream 1, there must be an increased empirical input rate. Thus the rate function for the output process of stream 1 is the same as that of the arrivals, as long as the buffer is not overflowing, i.e., $\alpha \le c-\mathbb{E}A_1^2$, which is equivalent to the decoupling constraint Eq. 6.

We have not claimed that the effective bandwidth of *individual* output streams is necessarily less than or equal to that of the input, as is true for the *aggregate* departure process. We conjecture that, individual sources may interact when the decoupling constraint (Eq. 6) is not met, allowing for a shift in "burstiness" from one stream to another. This might happen when a bursty source shares a queue with a rather smooth source. Also note that the result is stated for work conserving service disciplines including strategies such as generalized processor sharing [25].

The decoupling constraint Eq. 6 is sufficient but not necessary to guarantee that output's effective bandwidth is the same as that at the input. In fact the following example illustrates that it may be more conservative than necessary. Suppose the service policy gave full priority to the first traffic stream. In this case the result of Corollary 3.1 would suffice to determine the effective bandwidth of the departure process for stream 1. The decoupling constraint for stream 1 would be $\alpha_1^*(\delta) \leq c$, rather than Eq. 6 above. This conservativeness is not surprising, as the range of possible service policies is vast, i.e., they need only be work conserving.

3.1 Examples: Departure processes and traffic shaping via buffering

In [11] we considered the impact of memoryless rejection (or marking) policies and filtering on the effective bandwidth. We showed that among memoryless rejection policies with the same throughput a threshold function is optimal in the sense of minimizing the effective bandwidth at the output, but *only* in the case of i.i.d. arrivals. An examination of the impact of linear filtering with unit gain, in the case of Gaussian source models, suggested that the effective bandwidth is invariant to filtering. In this section we consider the effect of traffic shaping via buffering.

Corollary 3.1 suggests what might be the impact of a buffered traffic shaping device. Suppose the goal of traffic shaping is to minimize the output's effective bandwidth subject to a δ_o-constraint on the tail of the overflow probability at the shaping device. The aim to select the optimal release rate c such that the workload satisfies the constraint, and the entire effective bandwidth characteristic of the departure process $\alpha_D(\cdot, c)$ is minimized, i.e.,

$$\begin{array}{ll} \min\limits_{c>0} & \alpha_D(\cdot, c) \\ \text{such that} & \lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}^c(W > B) \leq -\delta_o. \end{array} \tag{7}$$

The notation emphasizes the dependence on the service rate c of the departure's effective bandwidth via $\alpha_D(\cdot, c)$ and that of the workload distribution via $\mathbb{P}^c(\cdot)$.

In general it is unclear that minimization of the overall effective bandwidth characteristic is well defined. However, note that $\alpha_D(\cdot, c)$ is non-decreasing in c when evaluated at any QoS. Thus the optimal release rate is the smallest c consistent with the overflow δ_o-constraint, i.e., the effective bandwidth $\alpha(\delta_o)$ of the input. Thus the effective bandwdth at the output of an optimal (in the sense of (7) traffic shaping device is given by:

$$\alpha_D(\delta, \alpha(\delta_o)) = \begin{cases} \alpha(\delta) & \text{if } \alpha^*(\delta) \leq \alpha(\delta_o), \\ \\ \alpha(\delta_o) - \frac{\Lambda^*(\alpha(\delta_o))}{\delta}, & \text{otherwise.} \end{cases} \tag{8}$$

The fact that the overall effective bandwidth is minmimized is important in looking at multi-service networks where several QoS may be in effect and the overall characteristics of the output impact performance [10]. Further study of the impact of leaky-bucket throttles for traffic shaping subject to loss or delay constraints can be found in [9].

This result is quite intuitive: When the release rate is large, the stream is oblivious to the buffer and its effective bandwidth remains unchanged. When the service rate is reduced to the minimum acceptable level, i.e., the effective bandwidth of the input, then, queueing in combination with deterministic release, work to our advantage by smoothing the traffic stream entering the network.

Below we present two examples to make these observations concrete. The first, is the particularly simple case in which the input is modeled by a Gaussian process. Usually such models arise from approximations of the *net input* to a system and hence allow for what appear to be "negative" arrivals. This model does however provide quite a bit of insight. The second example is a standard discrete-time On/Off Markov source.

Example 1: Let $\{A_n\}$ be a Gaussian process with mean μ and finite asymptotic variability

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n} \text{Var}(\sum_{j=1}^{n} A_j) < \infty.$$

The log-moment generating function and its dual are given by:

$$\Lambda(\delta) = \mu\delta + \frac{\delta^2\sigma^2}{2}, \;\; \Lambda^*(\alpha) = \frac{(\alpha - \mu)^2}{2\sigma^2}.$$

The effective bandwidth of the arrival process is $\alpha(\delta) = \mu + \frac{\delta\sigma^2}{2}$, while the decoupling bandwidth is $\alpha^*(\delta) = \mu + \delta\sigma^2$. The effective bandwidth of the departure process from a deterministic server at rate c is

$$\alpha_D(\delta, c) = \begin{cases} \alpha(\delta) & \text{if } \alpha^*(\delta) \leq c, \\ \\ c - \frac{(c-\mu)^2}{2\sigma^2\delta} & \text{otherwise.} \end{cases}$$

Figure 1 exhibits the quantities we are considering as a function of δ. The effective bandwidth of the departure process is identical to that of the arrival process for $\delta \leq (c-\mu)/\sigma^2$, after which it is reduced and converges hyperbolically to a maximum of c, the service rate of the queue.

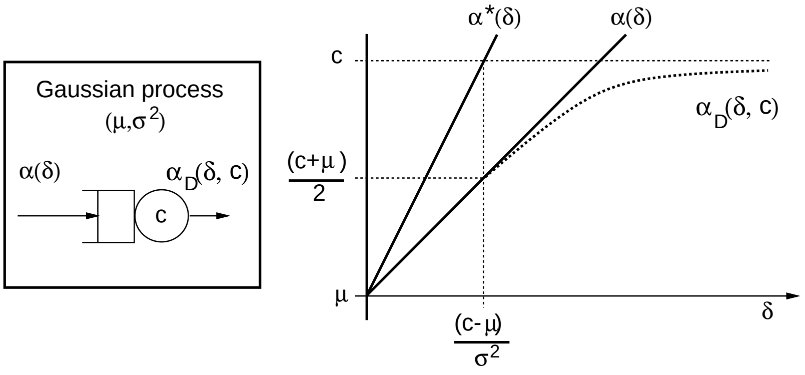

Figure 1: Input/Output effective bandwidths for a buffered for Gaussian source.

We can compute the solution (8) to the optimal traffic shaping problem (7) obtainint an output effective bandwidth given by

$$\alpha_D(\delta, \alpha(\delta_o)) = \begin{cases} \mu + \frac{\delta\sigma^2}{2} & \text{if } 2\delta \leq \delta_o, \\ \\ \mu + \frac{\delta_o\sigma^2}{2}(1 - \frac{\delta_o}{4\delta}), & \text{otherwise.} \end{cases}$$

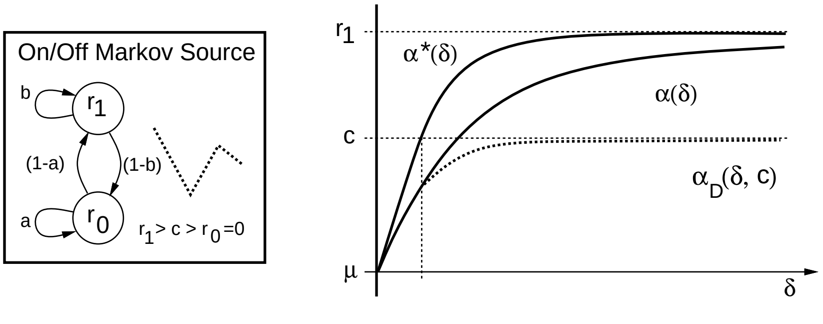

Figure 2: Input/Output effective bandwidths for a buffered discrete-time On/Off Markov source.

Example 2: Next, we consider a discrete-time On/Off Markov source having the following dynamics: the probability of making a transition from Off to On is $(1-a)$ and from On to Off is $(1-b)$. When the source is Off it generates no traffic, and when On r_1 units of traffic are produced. We will assume $r_1 > c > r_0 = 0$. The asymptotic log-moment generating function is given by [4]:

$$\Lambda(\delta) = \log\left[\frac{a + b\exp[\delta r_1] + \sqrt{(a + b\exp[\delta r_1])^2 + 4(1-a-b)\exp[\delta r_1]}}{2}\right],$$

and the effective bandwidth is $\alpha(\delta) = \frac{\Lambda(\delta)}{\delta}$. The rate function can be computed via $\Lambda^*(\alpha) = \sup_\theta[\theta\alpha - \Lambda(\theta)]$ and the decoupling bandwidth $\alpha^*(\delta) = \frac{d}{d\delta}\Lambda(\delta)$ is given by

$$\alpha^*(\delta) = r_1 \exp[\delta r_1 - \Lambda(\delta)] \left(\frac{b \exp[\Lambda(\delta)] + (1-a-b)}{2\exp[\Lambda(\delta)] - (a + b\exp[\delta r_1])} \right).$$

Fig. 2 shows a typical graph of the input and output effective bandwidths as well as the decoupling bandwidth as a function of δ for this traffic model. The effective bandwidth of the input stream increases to the peak rate with δ while that of the output process converges to the service rate c of the buffer. Of course, if $c > r_1$ the buffer will not affect the stream at all. Explicit calculation of (8) is somewhat involved and ommitted.

3.2 Example: Asymptotics for tandem queues

In this section we consider the large buffer asymptotics for a pair of (stable) queues in tandem. As in the single queue example there exists a stationary distribution, see Walrand [28, page 245].

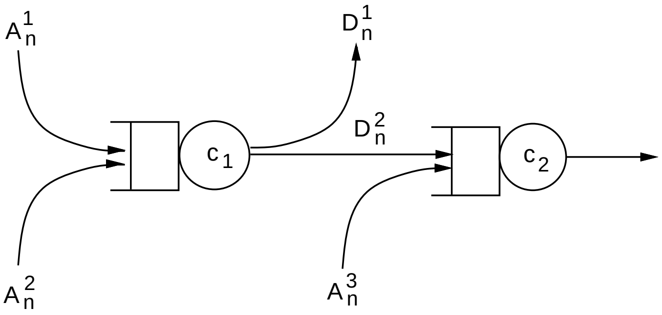

Figure 3: Tandem queues.

Using our characterization for the stationary output process we discuss resource management for a simple scenario shown in Figure 3. The figure shows two queues, shared by three traffic streams, we will assume the effective and decoupling bandwidths of the three streams are given by $\alpha_i(\delta),\ \alpha_i^*(\delta),\ i = 1,2,3$. We suppose the goal is to guarantee that

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(W_i > B) \le -\delta \text{ for } i = 1,2.$$

Using the result in Corollary 2.1 the above constraint is satisfied at the first queue if

$$\alpha_1(\delta) + \alpha_2(\delta) \le c_1.$$

Now consider the second queue. If the service rate of the first queue is such that

$$\alpha_1(0) + \alpha_2^*(\delta) \le c_1,$$

then by Corollary 3.3 the effective bandwidth of the stream $\{D_n^2\}$ entering the second queue is the same as the arrival process $\{A_n^2\}$ i.e., equals $\alpha_2(\delta)$. Thus the constraint on the second queue is guaranteed if

$$\alpha_2(\delta) + \alpha_3(\delta) \le c_2.$$

Two observations are worth making based on the simple tandem queue example. First, note that in order to guarantee decoupling we only need constraints on the aggregate traffic flow along routes sharing multiple queues (e.g., $\alpha_2^*(\delta) < c_1 - \alpha_1(0)$). Thus resource management based on effective bandwidths should be organized on a per route basis. Second, we note that the decoupling constraint may be subsumed by the effective bandwidth constraint if $\alpha_1(0) + \alpha_2^*(\delta) \le \alpha_1(\delta) + \alpha_2(\delta)$. This suggests that decoupling might in some instances come for free once performance requirements are satisfied.

4 Resource management for networks

In this section we consider more general networking scenarios and discuss some practical aspects of resource management via nodal decomposition.

The case of in-tree [**?**, 6] networks can be treated in a simple fashion by propagating the input-output characterization developed above up the tree. To provide further insight we discuss a heuristic for feed-forward networks. Such networks allow for multiple routes from a given origin to destination, as might be desirable to allow for alternative routing. In this case decoupling constraints appear to be significantly more complex, showing that guaranteeing decoupling can in practice be somewhat unwieldy.

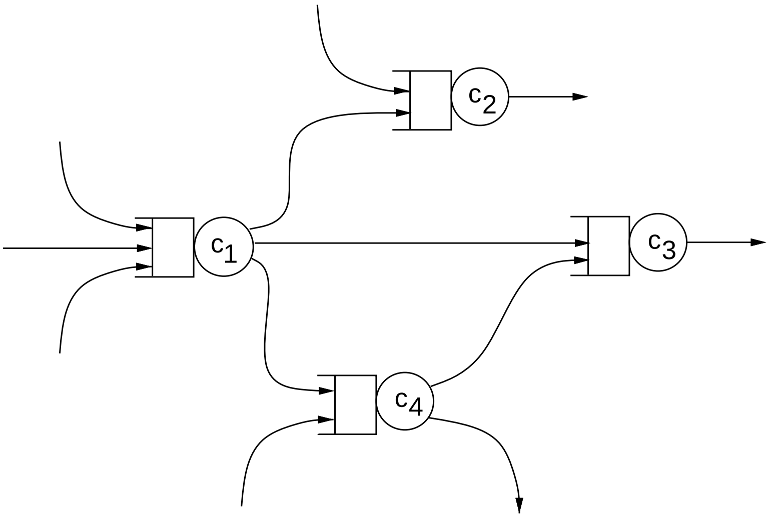

Figure 4: Resource management for networks.

Consider a stable feed-forward network of queues Q. Let R denote a collection of directed routes, where each route is denoted by an ordered subset of Q. We call the network feed-forward if any cutset, dividing the queues into disjoint sets, is such that all the routes in the cutset flow in the same direction. Let $R(i)$ denote the set of routes sharing queue i, and $DR(i)$ this same set of routes but truncated to include only nodes visited after i, i.e., we need only consider the downstream pieces for these routes. In order to guarantee decoupling we must ensure that the traffic flows on routes in $R(i)$ are decoupled with respect to all upstream queues, and in particular with respect to queue i. If none of the routes in $DR(i)$ intersect, i.e., they are such that, when $r_1, r_2 \in DR(i)$ then $r_1 \cap r_2 = \emptyset$, then the decoupling constraints at queue i would be

$$\alpha_r^*(\delta) + \sum_{j \in R(i)-r} \alpha_j(0) < c_i.$$

$\forall r \in R(i)$ where $\alpha_r(\delta), \alpha_r^*(\delta)$ denote the effective and decoupling bandwidths of the traffic flowing along the route r. We will call these nodal decoupling constraints, and note that these would suffice if we were dealing with a tree network. However, in order to have flexible routing permitting higher reliability and load distribution, networks are likely to have alternative routes leading to the same destinations, such is the case in Figure 4. In this case any collection of flows through node i which share a downstream queue, must jointly satisfy decoupling constraints at upstream queues, and typically these constraints are stronger than those above. For example the network in Figure 4 has seven possible routes $\{\{1,2\},\{1,3\},\{1,4\},\{1,4,3\},\{2\},\{4\},\{4,3\}\}$. Due to the downstream interaction of routes $\{1,3\}$ and $\{1,4,3\}$ we require that, in addition to nodal decoupling constraints at each node, the following constraint holds:

$$\alpha_{\{1,3\}}^*(\delta) + \alpha_{\{1,4,3\}}^*(\delta) + \sum_{j \in R(1)-\{\{1,3\},\{1,4,3\}\}} \alpha_j(0) < c_1.$$

Thus a complete set of decoupling constraints at a given node, would need to account for all possible downstream interactions among alternative routes.

In order to make this approach viable it is preferable to make guaranteeing decoupling a simple task. We believe that high-speed networks with sufficient "routing diversity" will naturally satisfy this requirement. A network with routing diversity is one in which the proportion of bandwidth taken by traffic sharing similar origins and destinations is small relative to the typical link capacity. If this is the case then nodal decoupling constraints should suffice; moreover, these can be ensured by guaranteeing that the peak rate of individual streams is small relative to the total link capacity. Below we explore a conservative rule along these lines, see [22] for a simulation study.

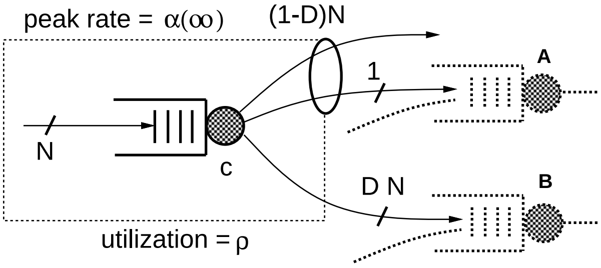

Figure 5: Diversity in routing and decoupling constraints.

Consider the scenario shown in Figure 5: N streams of a given type share a buffer with capacity c and utilization ρ. Assume all streams have an effective bandwidth $\alpha(\delta)$ and decoupling bandwidth $\alpha^*(\delta)$. The traffic streams are assumed to have a bounded sustainable peak rate, i.e., $\alpha(\infty) < \infty$.

Consider a single traffic stream (virtual circuit) sharing a downstream queue (labeled A in Figure 6) with other traffic in the network but no other traffic passing through one of the upstream buffers associated with its path. The decoupling constraint

$$\alpha^*(\delta) + (N-1)\alpha(0) < c$$

guarantees that this stream will have the same effective bandwidth downstream. Noting that $\alpha^*(\delta) \leq \alpha(\infty)$ and that $\alpha(0)$ is the mean arrival rate of a traffic stream, we see that the following conservative constraint guarantees decoupling:

$$\frac{\alpha(\infty)}{c} + \rho < 1. \tag{9}$$

Suppose the utilization of the network is below $\rho < 0.9$, then Eq. 9 means that $\alpha(\infty) < 0.1\ c$. That is, the peak rate of a single stream should be no more than 10% of the link capacity if it is to incur no distortion through the queue. Extensive simulations in [22] confirm that such a relationship is indeed approximately true, where the output distortion is measured with respect to output statistics rather than effective bandwidth. Satisfaction of such a constraint is not likely to be a problem since the capacity of links is expected to be orders of magnitude larger than the peak rate of single streams.

Guaranteeing decoupling per virtual circuit is however not sufficient in a network where several streams may in fact follow the same path; next we consider the decoupling requirement for such a scenario. Suppose a fraction D of N streams share a common downstream queue (labelled B in Fig. 6). D quantifies the fraction of streams entering a queue that might share the same downstream buffer or the "routing diversity." In this case the decoupling constraint would be

$$ND\alpha^*(\delta) + N(1-D)\alpha(0) < c,$$

which in turn gives a simpler, but conservative, constraint,

$$D\left[\frac{N\alpha(\infty)}{c} - \rho\right] + \rho < 1. \tag{10}$$

Note that while ρ represents the average utilization, $N\alpha(\infty)/c$ is the "worst case" utilization. Typically we would like the worst case to be much larger than 1, i.e., we expect to statistically multiplex streams composed of bursts, such that the aggregate peak rate exceeds the capacity of the link. The difference between these two quantities might be interpreted as a measure for the range of fluctuation in the system. Moreover, the magnitude of this difference, constrains the diversity in routing D required to maintain decoupling. Suppose, for example, that $\rho < 0.9$ and the worst case utilization satisfies $N\alpha(\infty)/c < 2.9$; in this case the constraint in Eq. 10 means that $D < 0.05$, i.e., no more than 5% of streams sharing a queue should share a downstream buffer.

The number 2.9 was chosen arbitrarily to match the simulation results in [22]; roughly this indicates that if all the streams were transmitting at their ”sustainable” peak rates the aggregate arrival rate would not exceed three times the capacity of the link. Typically we expect the worst case utilization might be larger than 2.9. In general the characteristics of the traffic and required performance constraints will determine both the achievable average and worst case utilizations and hence the required diversity. Once again the simulation study of Lau and Li [22] suggests that nodal decoupling will indeed hold when diversity of 5% is maintained within a network.

While guaranteeing decoupling may follow for free once an ATM network is managed to satisfy stringent performance constraints we should note that in practice nodal decomposition is likely to be a conservative approach. Indeed, the smoothing effect of queues (suggested by Corollary 3.2) is likely to work in our favor by reducing the effective bandwidth of traffic in the network when the decoupling constraints are not quite in place, however, this may in turn allow for transfer of burstiness from one stream to another. Furthermore the approximate nature of effective bandwidths in and of themselves can lead to over or under allocation of resources, see Choudhury *et al.* [8]. Undoubtedly one should be careful in interpreting these asymptotic results, however subject to verification these approximations provide a reasonably simple integrated approach to resource management. By incorporating these simple structural results with monitoring and estimation we believe we can overcome many of these shortcomings.

5 Summary

In this paper we have obtained several novel results on the large buffer asymptotics for multi-class traffic streams sharing a feed-forward network of queues. We began by identifying the large deviation rate function at the output of a multi-class queue, subject to a work conserving service policy, for both aggregate and individual streams. This result demonstrates the smoothing property of queues in terms of reducing the effective bandwidth of a traffic streams. We introduced the notion of *decoupling bandwidths*, to ensure that the asymptotics for every queue in a feed-forward network are essentially decoupled. This indicates that resource management may be carried out by assuming nodal decomposition, i.e., by considering queues individually and using the effective bandwidths of streams as specified by users at the network edge. These results are only valid for networks with “large” shared buffers, but allow arbitrary work conserving service policies. Further research will focus on the interplay between service policies which can successfully guarantee specific qualities of service to certain users and the asymptotics of large but finite buffers.

Acknowledgments: G. de Veciana would like to thank A. Dembo and N. O’Connell for many helpful comments.

Appendix

Outline of the proof for Theorem 3.1: The stability condition, $\mathbb{E}A_1 < c$, guarantees the existence of a stationary distribution, see Loynes [23] or Walrand [28, Chap. 7]. In particular, let

$$\begin{aligned} W_n^m &= 0 \quad n \le -m, \\ W_{n+1}^m &= [W_n^m + A_{n+1} - c]^+ \quad n \ge -m, \end{aligned}$$

then the distribution of W_0^m converges monotonically to that of W where $\mathbb{P}(W < \infty) = 1$. We will denote the stationary workload and departure processes by $\{W_n\}$ and $\{D_n\}$ respectively.

By assumption the arrival process satisfies an LDP with a rate function given by the convex dual of $\Lambda(\cdot)$, i.e.,

$$I(\alpha) = \Lambda^*(\alpha) = \sup_\theta [\theta\alpha - \Lambda(\theta)].$$

We will let S_n^D, S_n^A for $n > 0$ denote the the partial sums of the departure and arrival processes. We further use the convention that $S_0^A = 0$ and $S_n^A = \sum_{i=n+1}^0 A_i$, for $n < 0$. We begin by considering the departures for a stationary version of this queue. Note that for $n > 0$

$$S_n^D \leq W_0 + S_n^A, \;\text{ where }\; W_0 = \max_{i \geq 0}[S_{-i}^A - ic]$$

and W_0 is in general is not independent of S_n^A. Using an argument similar to that used to prove Theorem 2.1 (see [11, 4]) we have that for $\varepsilon > 0$ and large enough n,

$$\begin{aligned} \mathbb{E}\exp[\theta S_n^D] &\leq \mathbb{E}\exp[\;\theta\;(W_0 + S_n^A)\;] \\ &= \mathbb{E}\exp[\;\theta\;\max_{i\geq 0}[S_n^A + S_{-i}^A - ic]\;] \\ &\leq \sum_{i \geq 0} \exp[(\Lambda(\theta) + \varepsilon)n + (\Lambda(\theta) + \varepsilon - \theta c)i] \\ &\leq C\;\exp[(\Lambda(\theta) + \varepsilon)n], \end{aligned}$$

for some finite constant C as long as $\Lambda(\theta) + \varepsilon < c\theta$. Suppose $\mathbb{E}A_1 \leq \alpha \leq c$ then by Chebychev's bound we have that

$$\mathbb{P}\left(S_n^D \geq n\alpha\right) \;\leq\; \exp[-\theta n\alpha]\,C\exp[(\Lambda(\theta) + \varepsilon)n].$$

It follows by letting $n \to \infty$ and $\varepsilon \to 0$ that

$$\limsup_{n\to\infty} \frac{1}{n}\log\mathbb{P}\left(\frac{1}{n}S_n^D \geq \alpha\right) \leq -\sup_{\{\theta:\;\Lambda(\theta) < c\theta\}} [\theta\alpha - \Lambda(\theta)] = -\Lambda^*(\alpha).$$

We obtain a lower bound for the cumulative departures by starting with an empty queue at time zero, i.e., $W_0^0 = 0$. The cumulative departures from this queue, denoted by $\bar{S}_n^D$, clearly lower bound those from the associated stationary version, i.e., $S_n^D \geq \bar{S}_n^D$. Note that

$$\begin{aligned} \bar{S}_n^D &= S_n^A - W_n^0 \\ &= S_n^A - \max_{0 \leq i \leq n}[S_n^A - S_i^A + [n-i]c] \\ &= \min[S_n^A,\, S_{n-1}^A + c,\; \ldots,\, S_1^A + [n-1]c,\, nc] \\ \bar{S}_n^D - nc &= \min[S_n^A - nc,\, S_{n-1}^A - [n-1]c,\; \ldots,\, S_1^A - c,\, 0]. \end{aligned}$$

Consequently we have that

$$\begin{aligned} \mathbb{P}\left(S_n^D > n\alpha\right) &\geq \mathbb{P}\left(\bar{S}_n^D - nc > n[\alpha - c]\right) \\ &\geq \mathbb{P}\left(\bar{S}_n^D - nc > n[\alpha - c] \mid S_n^A > n\alpha\right) \;\times\; \mathbb{P}(S_n^A > n\alpha) \\ &= \mathbb{P}\left(\min[S_n^A - nc,\; \ldots,\, S_1^A - c,\, 0] > n[\alpha - c] \mid S_n^A - nc > n[\alpha - c]\right) \;\times \\ &\qquad\qquad \mathbb{P}(S_n^A > n\alpha). \end{aligned}$$

Finally taking limits we find that

$$\begin{aligned} &\liminf_{n\to\infty} \frac{1}{n}\log\mathbb{P}\left(\frac{1}{n}S_n^D > \alpha\right) \geq \\ &\quad \liminf_{n\to\infty} \frac{1}{n}\log\mathbb{P}\left(\min[S_n^A - nc,\; \ldots,\, S_1^A - c,\, 0] > n(\alpha - c) \mid S_n^A - nc > n[\alpha - c]\right) \;+ \\ &\qquad \liminf_{n\to\infty} \frac{1}{n}\log\mathbb{P}\left(S_n^A > n\alpha\right) \;\geq\; 0 - \inf_{x > \alpha}\Lambda^*(x) = \Lambda^*(\alpha). \end{aligned}$$

The bound for the second term follows by a straightforward application of the large deviation principle for the arrival process, when $\alpha \geq \mathbb{E}A_1$, and the continuity of $\Lambda^*(\cdot)$ at α. The asymptotic probability of the first term can be estimated by way of a result for the conditional distribution of sample paths corresponding to the partial sum process. Indeed, as exhibited in Figure 6, one can show that the mass of the conditional distribution of paths leading to $S_n^A - nc > n[\alpha - c]$ concentrates on a specific path lying above the endpoint $n(\alpha - c)$. The "most likely" path for the partial sums $S_i - ic$

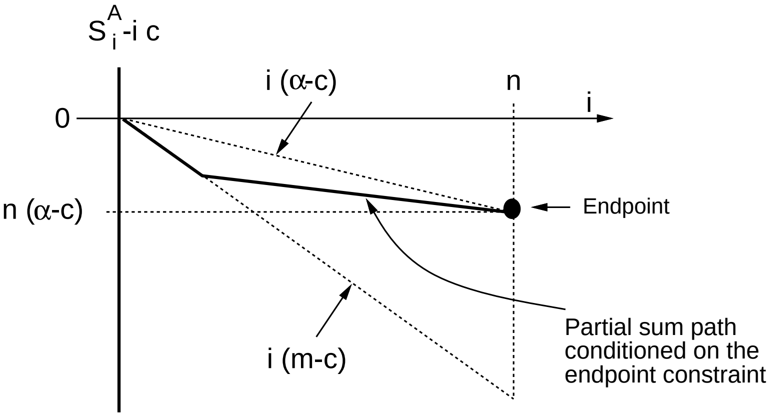

Figure 6: Conditional path subject to constraint on the endpoint.

follows the line $i(m-c)$, corresponding to the mean path, but then, due to the conditioning, breaks off toward the endpoint following a new line. From the figure it should be clear that the probability that the path's minimum remains above the endpoint goes to 1, and thus the log goes to zero. This result is essentially a consequence of the rate function's strict convexity. A discussion of this result is beyond the scope of this paper, we refer the reader to the work of Asmussen [2] and Anantharam [1], for an investigation of the normalized partial sum paths of i.i.d. random variables, and Dembo and Zajic [12] for a general result, which we use here, predicated on the existence of sample path LDP.

The theorem is proved by applying Lemma 3.1 and noting that rate function is clearly infinite on $[0,c]^c$.
□

Outline of the proof of Corollary 3.3: We will adapt the argument used in Theorem 3.1 to the departure process of the first stream. Let $S_n^{A^1}$ and $S_n^{D^1}$ denote the partial sums for the arrivals and departures of this stream, while $S_n^{A^2}$ and $S_n^{D^2}$ denote the cumulative arrivals and departures for the other streams sharing the queue.

As in Theorem 3.1 we have that

$$S_n^{D^1} \leq W_0 + S_n^{A^1}, \text{ where } W_0 = \max_{i\geq 0}[S_{-i}^{A^1} + S_{-i}^{A^1} - ic].$$

W_0 denotes the aggregate stationary workload at time 0. For $\varepsilon > 0$ and large enough n,

$$\begin{aligned} \mathbb{E}\exp[\theta S_n^{D^1}] &\leq \mathbb{E}\exp[\,\theta\,(W_0 + S_n^{A^1})\,] \\ &\leq \sum_{i\geq 0}\exp[(\Lambda_1(\theta)+\varepsilon)n + (\Lambda_1(\theta)+\Lambda_2(\theta)+\varepsilon-\theta c)i] \\ &\leq C\,\exp[(\Lambda_1(\theta)+\varepsilon)n], \end{aligned}$$

where C is finite as long as $\Lambda_1(\theta)+\Lambda_2(\theta) < c\theta$. Now suppose $\alpha \geq \mathbb{E}A_1^1$ then by Chebychev's bound we have that

$$\mathbb{P}\left(S_n^{D^1} \geq n\alpha\right) \;\leq\; \exp[-\theta n\alpha]\,C\exp[(\Lambda_1(\theta)+\varepsilon)n].$$

It follows by letting $n \to \infty$ and $\varepsilon \to 0$ that

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left(\frac{1}{n}S_n^{D^1} \geq \alpha\right) \leq -\sup_{\{\theta:\,\Lambda_1(\theta)+\Lambda_2(\theta)<c\theta\}}[\theta\alpha - \Lambda_1(\theta)]$$

Finally, recall that we have assumed $\Lambda_1(\delta)+\Lambda_2(\delta) < c\delta$ (the effective bandwidth constraint) and note that if we further constrain $\alpha \leq \alpha_1^*(\delta)$ then

$$\sup_{\{\theta:\,\Lambda_1(\theta)+\Lambda_2(\theta)<c\theta\}}[\theta\alpha - \Lambda_1(\theta)] = -\Lambda_1^*(\alpha).$$

So for $\alpha \in [\mathbb{E}A_1^1, \alpha_1^*(\delta)]$ we have that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} S_n^{D^1} \geq \alpha\right) \leq -\Lambda_1^*(\alpha).$$

To show a lower bound for all work conserving service policies we argue as follows. Since the aggregate queue is stable (empties in finite time) we need only consider the queue starting from empty. Thus, suppose the queue is empty at time zero, W_0^0 and denote the cumulative departures by $\overline{S}_n^{D^1}, \overline{S}_n^{D^2}$ and the aggregate by $\overline{S}_n^D$.

We fix $\varepsilon > 0$ and write

$$\mathbb{P}(\overline{S}_n^{D^1} > n\alpha) \;\geq\; \mathbb{P}(\overline{S}_n^{D^1} > n\alpha,\, S_n^{A^1} > n(\alpha + 2\varepsilon),\, \mathcal{E}_n),$$

where $\{S_n^{A^1} > n(\alpha + 2\varepsilon)\}$ suggests that stream 1 maintains an empirical arrival rate similar to that desired at the output, and $\mathcal{E}_n$ suggests that the second stream maintains an arrival rate close to its mean $m = \mathbb{E}A_1^2$. More precisely, $\mathcal{E}_n$ denotes the event

$$\mathcal{E}_n = \{S_i^{A^2} \in (im - n\varepsilon, im + n\varepsilon), \text{ for } i = 1, \ldots n\},$$

corresponding to the case where the partial sum path of stream 2 stays close to its mean arrival rate; this event is typically very likely to occur, in fact we have that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\mathcal{E}_n) \;=\; 0,$$

by the LDP results in Dembo and Zajic [12].

Now by conditioning and using the independence of the input streams we obtain,

$$\mathbb{P}(\overline{S}_n^{D^1} > n\alpha) \geq \underbrace{\mathbb{P}(\overline{S}_n^{D^1} > n\alpha \mid S_n^{A^1} > n[\alpha + 2\varepsilon], \mathcal{E}_n)}_{\text{Term 1}} \times \underbrace{\mathbb{P}(S_n^{A^1} > n[\alpha + 2\varepsilon]) \;\times\; \mathbb{P}(\mathcal{E}_n)}_{\text{Term 2}}.$$

As for Theorem 3.1 limits to find a lower bound for all these terms. Term 2 results in

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} S_n^{A^1} > \alpha + 2\varepsilon\right) + \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\mathcal{E}_n) \;\geq\; -\Lambda_1^*(\alpha + 2\varepsilon) + 0.$$

To lower bound Term 1, we use the fact that (as shown in proof for Thm 3.1) for $n > 0$ we have,

$$\begin{aligned} \overline{S}_n^{D^1} + \overline{S}_n^{D^2} - nc &= \min[S_n^A - nc,\ \ldots,\ S_1^A - c,\ 0], \\ \overline{S}_n^{D^1} &= \min[S_n^{A^1} + S_n^{A^2} - nc,\ \ldots,\ S_1^{A^1} + S_1^{A^2} - c,\ 0] - \overline{S}_n^{D^2} + nc. \end{aligned}$$

Also note that $S_n^{A^2} \geq \overline{S}_n^{D^2}$ and that we have conditioned on $\mathcal{E}_n$, so that Term 1 is lower bounded by

$$\begin{aligned} &\mathbb{P}(\min[S_n^{A^1} + S_n^{A^2} - nc, \ldots, S_1^{A^1} + S_1^{A^2} - c, 0] > n[\alpha - c] + S_n^{A^2} \mid S_n^{A^1} > n[\alpha + 2\varepsilon],\, \mathcal{E}_n) \;\geq \\ &\quad \mathbb{P}(\min[S_n^{A^1} + n[m - \varepsilon] - nc, \ldots, S_1^{A^1} + m - n\varepsilon - c, 0] \;> \\ &\qquad\qquad n[\alpha - c] + n[m + \varepsilon] \mid S_n^{A^1} > n[\alpha + 2\varepsilon]) \;\geq \\ &\quad\quad \mathbb{P}(\min[S_n^{A^1} + n[m - c], \ldots, S_1^{A^1} + m - c, 0] > n[\alpha + 2\varepsilon + m - c] \mid S_n^{A^1} > n[\alpha + 2\varepsilon]). \end{aligned}$$

Once more taking limits we have that

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\min[S_n^{A^1} + n[m - c], \ldots, S_1^{A^1} + m - c, 0] \;> \\ &\qquad\qquad n[\alpha + 2\varepsilon + m - c] \mid S_n^{A^1} > n[\alpha + 2\varepsilon]) \;=\; 0. \end{aligned}$$

Indeed, as in the proof of Theorem 3.1 this conditional distribution converges to one as long as $\mathbb{E}A_1^1 \leq \alpha + 2\varepsilon \leq c - \mathbb{E}A_1^2$, in which case the partial sum process for stream 1 remains above the endpoint $n[\alpha + 2\varepsilon + m - c]$. Finally letting $\varepsilon \to 0$ to obtain

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\overline{S}_n^{D^1} > n\alpha) \geq -\liminf_{\varepsilon\to 0} \Lambda_1^*(\alpha + 2\varepsilon) = -\Lambda_1^*(\alpha)$$

The upper and lower bounds we have obtained coincide when $\alpha \in [\mathbb{E}A_1^1, \min[\alpha_1^*(\delta), c - \mathbb{E}A_1^2]\,]$. Furthermore assuming the decoupling constraint (Eq. 6) is in effect and adapting Lemma 3.1 to this case, we have identified the rate function of departures corresponding to the first stream as $\Lambda_{D^1}^*(\alpha) = \Lambda_1^*(\alpha)$ on the set $[\mathbb{E}A_1^1, \alpha_1^*(\delta)]$. A similar argument to that in Corollary 3.2 shows that the effective bandwidth of the departures for stream 1 is indeed equal to that of the input for the given δ-constraint:

$$\begin{aligned} \alpha_{D^1}(\delta) &= \frac{1}{\delta} \sup_{\alpha \geq \mathbb{E}A_1^1} [\alpha\delta - \Lambda_{D^1}^*(\alpha)] \\ &= \frac{1}{\delta}[\alpha_1^*(\delta)\delta - \Lambda_1^*(\alpha_1^*(\delta)] \\ &= \frac{\Lambda_1(\delta)}{\delta} = \alpha_1(\delta), \end{aligned}$$

where the second equalities is a consequence of the definition of the decoupling bandwidth and that $\Lambda_{D^1}^*(\alpha) = \Lambda_1^*(\alpha)$ on the set $[\mathbb{E}A_1^1, \alpha_1^*(\delta)]$. □

References

[1] V. Anantharam. How large delays build up in a GI/G/1 queue. *Queueing Syst.*, 5:345–68, 1988.

[2] S. Asmussen. Conditional limit theorems relating the random walk to its associate, with applications to risk processes and the GI/G/1 queue. *Adv. App. Prob.*, 14:143–70, 1982.

[3] W. Bryc and A. Dembo. Large deviations and strong mixing. *Preprint*, 1993.

[4] C.-S. Chang. Stability, queue length and delay, part II: Stochastic queueing networks. *Proc. IEEE CDC, Tucson, AZ*, pages 1005–1010, 1992.

[5] C.-S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin. Effective bandwidths and fast simulation of ATM networks, 1993.

[6] C.-S. Chang and T. Zajic. Effective bandwidths of departure processes from queues with time varying capacities. In *Proc. IEEE INFOCOM*, 1995.

[7] C.S. Chang. Approximations of ATM networks: Effective bandwidths and traffic descriptors. *IBM Research Report 18954*, 1993.

[8] G.L. Choudhury, D.M. Lucantoni, and W. Whitt. Squeezing the most out of ATM. Preprint, 1993.

[9] G. de Veciana. Leaky buckets and optimal self-tuning rate control. In *Proc. IEEE GLOBECOM*, pages 1207–11, 1994.

[10] G. de Veciana and G. Kesidis. Bandwidth allocation for multiple qualities of service using generalized processor sharing. *IEEE Transactions on Information Theory*, 42(1):268–72, January 1996.

[11] G. de Veciana and J. Walrand. Effective bandwidths: Call admission, traffic policing and filtering for ATM networks. *Queueing Systems*, 20:37–59, 1995.

[12] A. Dembo and T. Zajic. Large deviations for sample path of partial sums. *Preprint*, 1992.

[13] J.D. Deuschel and D.W. Stroock. *Large Deviations*. Academic Press, Boston, 1989.

[14] N.G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *preprint*, 1993.

[15] A.I. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking*, 1:329–43, 1993.

[16] M. R. Frater. *Estimation of the Statistics of Rare Events in Data Communications Systems*. PhD thesis, Dept. of Systems Engineering Research School of Physical Sciences, The Australian National University, 1990.

[17] R.J. Gibbens and P.J. Hunt. Effective bandwidths for multi-type UAS channel. *Queueing Systems*, 9(1):17–28, 1991.

[18] R. Guérin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE J. Select. Areas Commun.*, 9(7):968–81, 1991.

[19] J.Y. Hui. Resource allocation for broadband networks. *IEEE J. Select. Areas Commun.*, 6(9):1598–1608, 1988.

[20] F.P. Kelly. Effective bandwidths of multi-class queues. *Queueing Syst.*, 9(1):5–15, 1991.

[21] G. Kesidis, J. Walrand, and C.-S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Trans. Networking*, 1(4):424–428, 1993.

[22] W-C Lau and S-Q. Li. Traffic analysis in large-scale high-speed integrated networks: Validation of nodal decomposition approach. In *IEEE INFOCOM Proc.*, 1993.

[23] R.M. Loynes. The stability of a queue with non-independent inter-arrivals and service times. *Proc. Camb. Phil. Soc.*, 58:497–520, 1962.

[24] N. O'Connell. Large deviations in queueing networks. *DIAS Technical Report*, No. DIAS-APG-9413, 1994.

[25] A.K. Parekh and R.G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The single node case. *IEEE/ACM Trans. Networking*, 1(3):344–57, 1993.

[26] S. Parekh and J. Walrand. A quick simulation of excessive backlogs in networks of queues. *IEEE Trans. Automatic Control*, 34:54–66, 1989.

[27] D. Towsley. Providing quality of service in packet switched networks. *Performance Evaluation of Computer and Communications Systems (ed. L. Donatiello and R. Nelson)*, Springer-Verlag:560–586, 1993.

[28] J. Walrand. *An Introduction to Queueing Networks*. Prentice-Hall, 1988.

[29] W. Whitt. Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunication Systems*, 2:71–107, 1993.

[30] O. Yaron and M. Sidi. Generalized processor sharing networks with exponentially bounded burstiness arrivals. *IEEE INFOCOM Proc.*, 2:628–635, 1994.