Computable Exponential Bounds for Intree Networks with Routing

Cheng-Shang Chang* Jay Cheng

Dept. of Electrical Engineering
National Tsing Hua University
Hsinchu 30043, Taiwan, R.O.C.

Abstract

In this paper, we refine the calculus proposed in [5, 8, 9]. The new calculus, including network operations for multiplexing, input-output relation, and routing, allows us to compute tighter exponential bounds for the tail distributions of queue lengths in intree networks with routing. In particular, if external arrival processes and routing processes are either Markov arrival processes or autoregressive processes, the stationary queue length at a local node is stochastically bounded above by the sum of a constant and an Erlang random variable. The decay rate of the Erlang random variable is not greater than (in some cases equal to) the decay rate of the tail distribution of the stationary queue length. The number of stages of the Erlang random variable is the number of external arrival processes and routing processes contributing to its queue length. For the single queue case, both the lower and upper bounds are derived.

1 Introduction

Future high speed digital networks will carry various classes of traffic, such as voice, video, fax, data, etc. Some of them, e.g., voice and video, require grade-of-service (GOS) guarantees. These GOS guarantees are usually in terms of the end-to-end delay and the overall loss probability. As discussed in Nagarajan, Kurose and Towsley [30], these GOS can be guaranteed indirectly through allocating proper GOS at the local nodes. In queueing terminology, the problem of local GOS guarantees is then equivalent to finding or bounding the tail distributions of queue lengths and delays at the local nodes.

One of the most interesting approaches in dealing with this problem is the recently developed theory of effective bandwidth (see e.g., [21, 23, 18, 5, 25, 8, 16, 33, 19, 6, 24, 30, 34, 35, 14], and references therein). This theory, strongly connected to the large deviation principle, provides the asymptotic decay rate of the tail distribution of the queue length at a local node, i.e.,

$$\lim_{x\to\infty} \frac{\log Pr(q \geq x)}{x} = -\theta^*, \tag{1}$$

where q is the stationary queue length and θ^* is the decay rate. Equation (1) suggests the approximation for $Pr(q \geq x)$ by $e^{-\theta^* x}$. As in Choudhury, Lucantoni and Whitt [10], one might wonder whether the approximation is good or not. In fact, Choudhury et al. [10] showed that the approximation can be very far from its true value if the node has a large number of homogeneous sources.

To provide further support for the theory of effective bandwidth, in this paper we refine the calculus proposed in Chang [5], Chang, Heidelberger, Juneja and Shahabuddin [8], and Chang, Thomas and Kiang [9]. The new calculus, including network operations for multiplexing, input-output relation, and routing, allows us to compute tighter exponential bounds for the tail distributions of queue lengths in intree networks with routing. In particular, if external arrival processes and routing processes are either Markov arrival processes or autoregressive processes, the stationary queue length at a local node is stochastically bounded above by the sum of a constant and an Erlang random variable. The decay rate of the Erlang random variable is not greater than (in some cases equal to) the real decay rate θ^* in (1). The number of stages of the Erlang random variable is the number of external arrival processes and routing processes contributing to its queue length.

Our calculus is built upon the characterization of sources developed in Chang, Thomas and Kiang [9], which in turn is influenced by Cruz [12, 13] and Kurose [26]. This characterization, called linear upper envelope processes (UEP), basically measures the difference between an arrival process and a constant rate process. A surprising discovery in this paper, though simple in derivation, is that the difference is essentially the stationary queue length of a queue with a constant

*Research partially supported by the National Science Council, Taiwan, R.O.C., under Contracts NSC 83-0208-M007-091 and NSC 83-0404-E007-052.

capacity subject to the arrival process. This discovery reduces the characterization problem to finding or bounding the queue length of a queue with a constant capacity.

For (finite state) Markov arrival processes or (bounded) autoregressive processes, it is shown in Chang, Heidelberger, Juneja, and Shahabuddin [8] that the likelihood ratio of the original process to the corresponding conjugate process is uniformly bounded for every sample path. The likelihood ratio, known as a martingale, allows us to modify Ross's martingale proof for Kingman's exponential bounds of $GI/GI/1$ queues ([31], pp. 237). This modification leads to both the lower and upper bounds for a single queue subject to Markov arrival processes or autoregressive processes. To be precise, we derive

$$c_1 e^{-\theta^* x} \leq Pr(q \geq x) \leq c_2 e^{-\theta^* x} \tag{2}$$

for some explicit constants c_1 and c_2. The upper bounds naturally provide the linear UEPs needed for the calculus.

For an intree network with routing, stochastic upper bounds can be derived via the closure properties of stochastic ordering once the characterizations of external arrival processes and routing processes are obtained. Intuitively, our stochastic upper bound at a local node in an intree network is obtained by comparing each external process and each routing process with a constant rate process and then pasting the difference together.

This paper is organized as follows: we briefly review the notion of linear UEP and derive the characterization by a stationary queue length in Section 2. We show that both Markov arrival processes and autoregressive processes have a family of linear UEPs in Section 3. We then derive the closure properties for network operations in Section 4. As a result, we provide a recursive algorithm for obtaining stochastic upper bounds for queue lengths in intree networks. The paper is concluded in Section 5.

2 Stochastic ordering and envelope processes

Chang, Thomas and Kiang [9] defined the following characterization of point processes based on the notion of stochastic ordering. Let A be a point process with $A(t_1, t_2)$ being the number of arrivals in the interval $(t_1, t_2]$, $-\infty < t_1 \leq t_2 < \infty$.

Definition 2.1 *A stochastic process $\{\hat{A}(t), t \geq 0\}$ is an* upper envelope process *(UEP) of a point process A if for all $t \geq 0$*

$$\{A(t-s, t), 0 \leq s \leq t\} \leq_{st} \{\hat{A}(s), 0 \leq s \leq t\}, \tag{3}$$

where $\leq_{st}$ denotes the stochastic ordering. Similarly, a stochastic process $\{\breve{A}(t), t \geq 0\}$ is a lower envelope process *(LEP) of a point process A if for all $t \geq 0$*

$$\{A(t-s, t), 0 \leq s \leq t\} \geq_{st} \{\breve{A}(s), 0 \leq s \leq t\}. \tag{4}$$

In particular, if an UEP has the form, $\hat{A}(t) = \alpha t + \sigma$, for some constant α and some nonnegative random variable σ, then it will be called a linear UEP (α, σ). The notion of linear LEP's is defined similarly when an LEP has the form $\breve{A}(t) = \alpha t - \sigma$.

Recall that a random variable X is stochastically smaller than Y if $Ef(X) \leq Ef(Y)$ for all f increasing. Also, a stochastic process $\{X(t), t_1 \leq t \leq t_2\}$ is stochastically dominated by $\{Y(t), t_1 \leq t \leq t_2\}$ if for all n and $t_1, \ldots, t_n \in [t_1, t_2]$

$$Ef(X(t_1), \ldots, X(t_n)) \leq Ef(Y(t_1), \ldots, Y(t_n)) \tag{5}$$

for all f increasing. For other properties of stochastic ordering, we refer to Stoyan [32], Ross [31] and references therein.

Though the original intent of Chang, Thomas and Kiang [9] was to provide a simple and unified proof for various stability results (loose bounds) for open networks of queues, we were surprised that this characterization could also be used to provide tight bounds for intree networks with routing like those in Cruz [12, 13], Kurose [26] and Chang [3, 4]. This discovery is mainly based on the following lemma that links the characterization to a stationary queue length distribution.

Lemma 2.2 *$\{A(t-s, t), 0 \leq s \leq t\} \leq_{st} \alpha s + \sigma$ if and only if $\max_{0 \leq s \leq t}[A(t-s, t) - \alpha s] \leq_{st} \sigma$.*

Proof. We first prove the "only if" part. Since $\{A(t-s, t), 0 \leq s \leq t\} \leq_{st} \{\alpha s + \sigma, 0 \leq s \leq t\}$, we can construct two processes $\{\tilde{A}(t-s, t), 0 \leq s \leq t\}$ and $\{\alpha s + \tilde{\sigma}, 0 \leq s \leq t\}$ on the same probability space such that (i) $\{\tilde{A}(t-s, t), 0 \leq s \leq t\} =_{st} \{A(t-s, t), 0 \leq s \leq t\}$, (ii) $\tilde{\sigma} =_{st} \sigma$, and (iii) $\tilde{A}(t-s, t) \leq \alpha s + \tilde{\sigma}, 0 \leq s \leq t$, a.s. From (iii), it follows that $\max_{0 \leq s \leq t}[\tilde{A}(t-s, t) - \alpha s] \leq \tilde{\sigma}$, a.s. and thus $\max_{0 \leq s \leq t}[A(t-s, t) - \alpha s] \leq_{st} \sigma$.

Now we prove the "if" part. For any increasing function $f \in I\!R^n$, $0 \leq s_i \leq t$, $i = 1, \ldots, n$,

$$\begin{aligned} & f(A(t-s_1, t) - \alpha s_1, \ldots, A(t-s_n, t) - \alpha s_n) \\ & \leq f(\max_{0 \leq s \leq t}[A(t-s, t) - \alpha s], \ldots, \max_{0 \leq s \leq t}[A(t-s, t) - \alpha s]). \end{aligned} \tag{6}$$

View the right hand side of (6) as a function of $\max_{0 \leq s \leq t}[A(t-s, t) - \alpha s]$. This function is also increasing. Thus, it follows from $\max_{0 \leq s \leq t}[A(t-s, t) - \alpha s] \leq_{st} \sigma$ that

$$\begin{aligned} & Ef(\max_{0 \leq s \leq t}[A(t-s, t) - \alpha s], \ldots, \\ & \max_{0 \leq s \leq t}[A(t-s, t) - \alpha s]) \leq Ef(\sigma, \ldots, \sigma). \end{aligned} \tag{7}$$

In conjunction with (6), one has $\{A(t-s,t)-\alpha s, 0\le s\le t\}\le_{st}\sigma$. But this is equivalent to $\{A(t-s,t), 0\le s\le t\}\le_{st}\alpha s+\sigma$ (by construction as in the proof of "only if" part). ■

Now consider a discrete-time queue with one class of customers as in Chang [5]. Let $a(t)$ and $q(t)$ be the number of arrivals at time t and the number of customers in the queue at time t respectively. (Throughout, assume that $a(t)\ge 0$ though in some cases the assumption is not really needed.) Assume that the buffer size is infinite and that the server can serve α customers per unit of time. The constant α will be referred to as the capacity of the server. Under a work-conserving policy, i.e., a policy that does not allow idling when there are customers in the queue, the queue is governed by the following Lindley's equation:

$$q(t+1)=(q(t)+a(t+1)-\alpha)^+ \qquad (8)$$

where $(x)^+\equiv\max(0,x)$. It is well known (e.g., see Loynes [28], Baccelli and Bremaud [2]) that $q(t)$ converges in distribution to a finite random variable q_α if $a(t)$ is a stationary and ergodic sequence with $Ea(t)<\alpha$.

Let $A(t_1,t_2)=\sum_{t=t_1+1}^{t_2}a(t)$. The following theorem provides a method of finding a characterization (actually the minimum one) of the arrival process.

Theorem 2.3 *(Queue length) Suppose that $a(t)$ is a stationary and ergodic sequence with $Ea(t)<\alpha$.*

(i) *(α,q_α) is a linear UEP of A.*

(ii) *If (α,σ) is a linear UEP of A, then $q_\alpha\le_{st}\sigma$. Thus, (α,q_α) is the minimum linear UEP of A with rate α.*

Proof. (i) If the queue is empty at time 0, then

$$q(t)=\max_{0\le s\le t}[A(t-s,t)-\alpha s]. \qquad (9)$$

Since $a(t)$ is a stationary and ergodic sequence with $Ea(t)<\alpha$, it follows from Loynes' construction [28] that $q(t)\le_{st}q_\alpha$ for all t. In view of Lemma 2.2, $\{A(t-s,t), 0\le s\le t\}\le_{st}\alpha s+q_\alpha$ for all t.

(ii) If (α,σ) is a linear UEP of A, it follows from (9) and Lemma 2.2 that $q(t)\le_{st}\sigma$. Since $q(t)$ converges in distribution to q_α, $q_\alpha\le_{st}\sigma$. ■

From Theorem 2.3, the characterization of a stationary and ergodic process $a(t)$ is not unique since one can find the minimum linear UEP for any $\alpha>Ea(t)$. By varying α, one can then find a family of linear UEPs of $a(t)$.

3 Examples of Envelope processes

In this section, we provide some examples that have linear UEPs as described in the previous section. These examples include Markov arrival processes and autoregressive processes. Our approach is based on a sample path criterion in [8].

Definition 3.1 *we call a triplet $(\tilde a(t;\theta),\hat a(\theta),\hat\sigma(\theta))$ a θ-upper envelope process ($\theta-$ UEP) of $a(t)$ if for all t and a_s, $s=1,\dots,t$,*

$$L_t(\omega)\frac{\exp(\theta\sum_{s=1}^t a_s)}{e^{\theta\hat a(\theta)t}}\le e^{\theta\hat\sigma(\theta)}, \qquad (10)$$

where $L_t(\omega)$ is the likelihood ratio with ω corresponding to the appropriate sample path, i.e.,

$$L_t(\omega)=\frac{Pr(a(s)=a_s, s=1,\dots,t)}{Pr(\tilde a(s;\theta)=a_s, s=1,\dots,t)}. \qquad (11)$$

Both $\hat a(\theta)$ and $\hat\sigma(\theta)$ are constant, and $0\le\hat\sigma(\theta)<\infty$. Similarly, we call a triplet $(\tilde a(t;\theta),\hat a(\theta),\hat\sigma(\theta))$ a θ-lower envelope process ($\theta-$ LEP) if the left hand side of (3) is bounded below by $\exp(-\theta\hat\sigma(\theta))$.

As in [8], we note that if $(\tilde a(t;\theta),\hat a(\theta),\hat\sigma(\theta))$ is a $\theta-$ UEP of $a(t)$, then

$$Ee^{\theta\sum_{s=1}^t a(s)}\le e^{\theta(\hat a(\theta)t+\hat\sigma(\theta))}. \qquad (12)$$

Thus, if $(\tilde a(t;\theta),\hat a(\theta),\hat\sigma(\theta))$ is both a $\theta-$ UEP and a $\theta-$ LEP of $a(t)$, then one has the Gärtner-Ellis limit [17, 15],

$$\lim_{t\to\infty}\frac{1}{t}\log Ee^{\theta\sum_{s=1}^t a(s)}=\theta\hat a(\theta). \qquad (13)$$

In this case, the stochastic process $\tilde a(t;\theta)$ is known as the θ-conjugate process of $a(t)$, and the function $\hat a(\theta)$ is called the effective bandwidth function of $a(t)$. For the use of these two notions, we refer to [5, 8, 6] and references therein. Examples that satisfy this criterion include Markov arrival processes and autoregressive processes (see [8]). Also, this criterion enjoys various closure properties such as superposition, input-output relation, and routing. In the following theorem, we show if an arrival process satisfies the criterion, then the stationary queue length can be bounded exponentially. In view of Theorem 2.3(i), the upper bound below can be used for obtaining a linear UEP of an arrival process.

Theorem 3.2 *Suppose that $a(t)$ is the arrival process to a queue governed by Lindley's equation in (8).*

(i) *(Upper bound) Assume $a(t)$ has a $\theta^* - UEP$ $(\tilde{a}(t;\theta^*), \hat{a}(\theta^*), \hat{\sigma}(\theta^*))$ such that $\theta^* > 0$ and $\hat{a}(\theta^*) = \alpha$. If both $a(t)$ and $\tilde{a}(t;\theta^*)$ are stationary and ergodic and $Ea(t) < \alpha$, then*

$$Pr(q_\alpha \ge x) \le e^{\theta^* \hat{\sigma}(\theta^*)} e^{-\theta^* x}. \tag{14}$$

(ii) *(Lower bound) Assume $a(t)$ is bounded above by a constant $a_{max} < \infty$, and $a(t)$ has a $\theta^* - LEP$ $(\tilde{a}(t;\theta^*), \hat{a}(\theta^*), \hat{\sigma}(\theta^*))$ such that $\theta^* > 0$ and $\hat{a}(\theta^*) = \alpha$. If both $a(t)$ and $\tilde{a}(t;\theta^*)$ are stationary and ergodic and $Ea(t) < \alpha$, then*

$$Pr(q_\alpha \ge x) \ge e^{-\theta^* a_{max}} e^{-\theta^* \hat{\sigma}(\theta^*)} e^{-\theta^* x}. \tag{15}$$

Note that the condition on the stationarity and ergodicity of $a(t)$ and $Ea(t) < \alpha$ ensures the existence of the stationary queue length q_α.

Proof. (i) Let $\{a^r(t), t \ge 0\}$ (resp. $\{\tilde{a}^r(t;\theta^*), t \ge 0\}$) be the reversed time process of $\{a(t), t \ge 0\}$ (resp. $\{\tilde{a}(t;\theta^*), t \ge 0\}$). Similarly, define

$$L_t^r(\omega) = \frac{Pr(a^r(s) = a_s, s = 1, \ldots, t)}{Pr(\tilde{a}^r(s;\theta^*) = a_s, s = 1, \ldots, t)}. \tag{16}$$

Note that

$$\frac{Pr(a^r(s) = a_s, s = 1, \ldots, t)}{Pr(\tilde{a}^r(s;\theta^*) = a_s, s = 1, \ldots, t)} = \frac{Pr(a(s) = a_{t-s}, s = 1, \ldots, t)}{Pr(\tilde{a}(s;\theta^*) = a_{t-s}, s = 1, \ldots, t)} \tag{17}$$

since we assume that both $a(t)$ and $\tilde{a}(t;\theta^*)$ are stationary. One can easily see that the triplet $(\tilde{a}^r(t;\theta^*), \hat{a}(\theta^*), \hat{\sigma}(\theta^*))$ is a $\theta^* -$ UEP of $a^r(t)$. From (9), it follows that

$$Pr(q_\alpha \ge x) = Pr(\max_{1 \le t < \infty} [\sum_{s=1}^{t} a^r(s) - \alpha t] \ge x). \tag{18}$$

Consider the stopping time $T = \inf\{t : \sum_{s=1}^{t} a^r(s) \ge \alpha t + x\}$. Clearly,

$$Pr(\max_{1 \le t < \infty} [\sum_{s=1}^{t} a^r(s) - \alpha t] \ge x) = Pr(T < \infty). \tag{19}$$

Let $T_n = \min[T, n]$ and $\mathcal{F}_t$ be the σ-algebra generated by $\{a^r(s;), s = 1, \ldots, t\}$. It is well known (e.g., [11], Exercise 7.4.2.) that $\{1/L_t^r(\omega), \mathcal{F}_t\}, t \ge 0\}$ is a nonnegative martingale (and hence a nonnegative supermartingale). Applying the optional sampling theorem (e.g. see Karlin and Taylor [22] Theorem 6.4.1) for nonnegative supermartingales and bounded stopping times yields $E(1/L_{T_n}^r(\omega)) \le E(1/L_0^r(\omega)) = 1$. Since the triplet $(\tilde{a}^r(t;\theta^*), \hat{a}(\theta^*), \hat{\sigma}(\theta^*))$ is $\theta^* -$ UEP of $a^r(t)$,

$$\begin{aligned} 1 &\ge E(\frac{1}{L_{T_n}^r(\omega)}) \\ &= E\Big(\frac{1}{L_{T_n}^r(\omega)} \frac{e^{\theta^* \hat{a}(\theta^*) T_n}}{e^{\theta^* \sum_{s=1}^{T_n} a^r(s)}} \frac{e^{\theta^* \sum_{s=1}^{T_n} a^r(s)}}{e^{\theta^* \hat{a}(\theta^*) T_n}}\Big) \\ &\ge e^{-\theta^* \hat{\sigma}(\theta^*)} E\Big(\frac{\exp(\theta^* \sum_{s=1}^{T_n} a^r(s))}{e^{\theta^* \hat{a}(\theta^*) T_n}}\Big). \end{aligned} \tag{20}$$

Since $\hat{a}(\theta^*) = \alpha$ and $\sum_{s=1}^{T_n} a^r(s) \ge \alpha T_n + x$ on the event $\{T_n < n\}$,

$$\begin{aligned} &E\Big(\frac{\exp(\theta^* \sum_{s=1}^{T_n} a^r(s))}{e^{\theta^* \hat{a}(\theta^*) T_n}}\Big) \\ &\ge E\Big(\frac{\exp(\theta^* \sum_{s=1}^{T_n} a^r(s))}{e^{\theta^* \hat{a}(\theta^*) T_n}} \mathbf{1}_{\{T_n < n\}}\Big) \\ &\ge e^{\theta^* x} Pr(T_n < n). \end{aligned} \tag{21}$$

In conjunction with (20), one has

$$Pr(T_n < n) \le e^{\theta^* \hat{\sigma}(\theta^*)} e^{-\theta^* x}. \tag{22}$$

Since $Pr(T_n < n) = Pr(T < n)$, letting $n \to \infty$ completes the argument.

(Due to page limitation, we will not present the proof for (ii) here.) ■

The right hand of (14) is the tail distribution of the random variable $\hat{\sigma}(\theta^*) + Z(\theta^*)$, where $Z(\theta^*)$ is exponentially distributed with mean $1/\theta^*$. As a direct consequence of Theorem 3.2(i) and Theorem 2.3(i), one has the following corollary.

Corollary 3.3 *Assume $a(t)$ has a $\theta - UEP$ $(\tilde{a}(t;\theta), \hat{a}(\theta), \hat{\sigma}(\theta))$ for some $\theta > 0$. If both $a(t)$ and $\tilde{a}(t;\theta)$ are stationary and ergodic, and $Ea(t) < \hat{a}(\theta)$, then $a(t)$ has a linear UEP $(\hat{a}(\theta), \hat{\sigma}(\theta) + Z(\theta))$, where $Z(\theta)$ is exponentially distributed with mean $1/\theta$.*

As discussed in [8], both Markov arrival processes and autoregressive processes satisfy the sample criterion in Definition 3.1. Thus, in Example 3.4 and Example 3.5 below, we use Corollary 3.3 to derive linear UEPs for Markov arrival processes and autoregressive processes.

Example 3.4 (MAP) Consider an M-state irreducible and aperiodic Markov chain $\xi(t) \in \mathbb{R}$. A process $a(t)$ is called a *Markov Arrival Process* (MAP) $a(t)$ if $(a(t), \xi(t))$ is a Markov chain with the transition kernel, $P_{i,j}(a) = Pr(a(t+1) \le a, \xi(t+1) = j | \xi(t) = i)$. Let $\Upsilon(\theta)$ be an $M \times M$ matrix with $\Upsilon_{i,j}(\theta) = \int e^{\theta a} dP_{i,j}(a)$. Also, let $sp(\Upsilon(\theta))$ and $f(\theta) = (f_1(\theta), \ldots, f_M(\theta))$ be the spectral radius of the matrix $\Upsilon(\theta)$ and the corresponding positive right eigenvector. Then

$$Pr(q_\alpha \ge x) \le \Gamma(\theta^*) e^{-\theta^* x}, \tag{23}$$

where θ^* is the solution of

$$\frac{1}{\theta} \log sp(\Upsilon(\theta)) = \alpha, \tag{24}$$

and $\Gamma(\theta) = \max_{i,j} f_i(\theta)/f_j(\theta)$. Thus, $a(t)$ has a family of linear UEPs

$$\{(\frac{1}{\theta} \log sp(\Upsilon(\theta)), \frac{1}{\theta} \log \Gamma(\theta) + Z(\theta)), \theta > 0\}, \tag{25}$$

where $Z(\theta)$ is exponentially distributed with mean $1/\theta$. The upper bound in (23) was previously presented in Duffield [14]. If $a(t) \leq a_{max}$ for some constant $a_{max} < \infty$, then

$$Pr(q_\alpha \geq x) \geq e^{-\theta^* a_{max}} \frac{1}{\Gamma(\theta^*)} e^{-\theta^* x}. \qquad (26)$$

Special cases of MAPs include Markov modulated processes (MMP), ON-OFF Markov processes, and bounded renewal processes (see [8] for details).

Example 3.5 Consider a standard autoregressive process $a(t)$ that is driven by a sequence of $M \times 1$ i.i.d. random vectors $u(t) = (u_1(t), \ldots, u_M(t))$:

$$a(t) = v x(t), \quad x(t) = \mathbf{r} x(t-1) + u(t),$$

where $\mathbf{r}$ is an $M \times M$ constant matrix, v is a $1 \times M$ constant vector, and $x(t)$'s are $M \times 1$ state vectors. We assume that $sp(\mathbf{r}) < 1$ so that $\mathbf{I} - \mathbf{r}$ is invertible. Also, the matrices v and $\mathbf{r}$ are nonnegative and $u(t)$ is nonnegative and bounded, i.e., there exists a vector u_{max} such that $\mathbf{0} \leq u(t) \leq u_{max}$ a.s., where $\mathbf{0}$ is a zero vector with the appropriate dimension. In this example, $a(t) \leq v(\mathbf{I} - \mathbf{r})^{-1} u_{max}$. Then

$$(\frac{1}{\Gamma(\theta^*)})^2 e^{-\theta^* x} \leq Pr(q_\alpha \geq x) \leq \Gamma(\theta^*) e^{-\theta^* x}, \qquad (27)$$

where θ^* is the solution of

$$\frac{1}{\theta} \log E e^{\theta v (\mathbf{I}-\mathbf{r})^{-1} u(t)} = \alpha, \qquad (28)$$

and $\Gamma(\theta) = exp(\theta v(\mathbf{I} - \mathbf{r})^{-1} u_{max})$. Thus, $a(t)$ has a family of linear UEPs

$$\{(\frac{1}{\theta} \log E e^{\theta v (\mathbf{I}-\mathbf{r})^{-1} u(t)}, \frac{1}{\theta} \log \Gamma(\theta) + Z(\theta)), \theta > 0\}, \qquad (29)$$

where $Z(\theta)$ is exponentially distributed with mean $1/\theta$.

4 Closure properties and intree networks with routing

Analogous to Chang [5], we derive a set of rules for network operations such as superposition, input-output relation of a single queue and routing. These rules will be used for deriving stochastic bounds for the stationary queue lengths in intree networks.

Lemma 4.1 *(Superposition)*
If two independent *point processes A_1 and A_2 have linear UEPs (α_1, σ_1) and (α_2, σ_2) respectively, then $A = A_1 + A_2$ has a linear UEP $(\alpha_1 + \alpha_2, \sigma_1 + \sigma_2)$, where σ_i, $i = 1$, and 2, are* independent.

Lemma 4.1 was derived in Chang, Thomas and Kiang [9], Lemma 2.6. Its proof follows directly from the closure property of stochastic ordering. Note that if A_1 and A_2 are not independent, then σ_1 and σ_2 are not independent. A worst case upper bound by maximal coupling could be derived. However, this is not related to the effective bandwidth theory and we will not pursue it further in this paper.

As discussed in Example 3.4 and Example 3.5, MAP and autoregressive processes have a family of linear UEPs parametrized in θ as in Corollary 3.3. From Lemma 4.1, the superposition of K independent MAP and autoregressive processes will result in a family of linear UEPs parametrized in $\theta_1, \ldots, \theta_K$. This increases the complexity of representing the linear UEPs of a process after superposition, and it could make our analysis difficult. To reduce the complexity of the problem, we simply choose the subset of linear UEPs when $\theta_1 = \ldots = \theta_K$. It can be shown that under certain technical conditions, this renders the maximum exponential decay rate.

Corollary 4.2 *Suppose two* independent *point processes A_1 and A_2 have families of linear UEPs $\{(\hat{a}_1(\theta), \hat{\sigma}_1(\theta) + Z_1(\theta)), \theta > 0\}$ and $\{(\hat{a}_2(\theta), \hat{\sigma}_2(\theta) + Z_2(\theta)), \theta > 0\}$), respectively, where $\{Z_1(\theta), \theta > 0\}$ and $\{Z_2(\theta), \theta > 0\}$ are families of random variables parametrized by θ. Then $A = A_1 + A_2$ has a family of linear UEPs $\{(\hat{a}_1(\theta) + \hat{a}_2(\theta), \hat{\sigma}_1(\theta) + \hat{\sigma}_2(\theta) + Z_1(\theta) + Z_2(\theta)), \theta > 0\}$, where $Z_1(\theta)$ and $Z_2(\theta)$ are* independent.

Let B be the departure process from the queue described in (8). Denote $b(t)$ as the number of departures at time t and $B(t_1, t_2)$ as the total number of departures in the interval $(t_1, t_2]$, i.e., $B(t_1, t_2) = \sum_{t=t_1+1}^{t_2} b(t)$.

Lemma 4.3 *(Input-output relation) Suppose the queue is empty at time 0, i.e., $q(0) = 0$. If the arrival process A has a linear UEP (γ, σ) and $\gamma \leq \alpha$, then the departure process B has a linear UEP (γ, σ).*

Proof. According to Lemma 2.2, it suffices to show that $\max_{0 \leq s \leq t}[B(t-s, t) - \gamma s] \leq_{st} \sigma$ for all t. Note that $B(t-s, t) \leq q(t-s) + A(t-s, t)$. From (9) and $\gamma \leq \alpha$, it follows that

$$q(t-s) \leq \max_{0 \leq \tau \leq t-s} [A(t-s-\tau, t-s) - \gamma\tau]. \qquad (30)$$

Thus,

$$\begin{aligned}
&B(t-s,t)\\
&\le \max_{0\le\tau\le t-s}[A(t-s-\tau,t-s)-\gamma\tau]+A(t-s,t)\\
&=\max_{0\le\tau\le t-s}[A(t-s-\tau,t)-\gamma(s+\tau)]\\
&+\gamma s\le \max_{0\le\ell\le t}[A(t-\ell,t)-\gamma\ell]+\gamma s. \qquad (31)
\end{aligned}$$

This implies

$$\max_{0\le s\le t}[B(t-s,t)-\gamma s]\le \max_{0\le s\le t}[A(t-s,t)-\gamma s]\le_{st}\sigma. \quad (32)$$

■

Note that the departure process is limited by the capacity, i.e., $B(t_1,t_2)\le \alpha t$. In conjunction with Lemma 4.3, one has the following corollary.

Corollary 4.4 *Suppose the queue is empty at time 0, i.e., $q(0)=0$. If the arrival processes A has a family of linear UEPs $\{(\hat{a}(\theta),\hat{\sigma}(\theta)+Z(\theta)),\theta>0\}$, where $\{Z(\theta),\theta>0\}$ is a family of random variables parametrized by θ, then B has a family of linear UEPs $\{(\hat{a}(\theta),\hat{\sigma}(\theta)+Z(\theta)),\theta>0\}$.*

For the routing operation, we need to introduce the notion of marked point processes as in Chang [5]. A discrete-time marked point process $\{(\tau(n),p(n)),n=0,1,2,\dots\}$ is a sequence of random vectors. The process $\{\tau(n)\}$ is called the arrival process with $\tau(n)$ being the arrival epoch of the n^{th} customer. Assume that $\tau(n)$ is increasing and that $\tau(n)\to\infty$ *a.s.* as $n\to\infty$. The random variable $p(n)$ is called the mark associated with the arrival epoch $\tau(n)$. (Throughout, we assume that $p(n)\ge 0$ though it may not be needed in some cases.) Let

$$a(t)=\sum_{n=0}^{\infty}\mathbf{1}_{\{\tau(n)=t\}}. \quad (33)$$

Then $a(t)$ is the number of arrivals at time t. Let

$$r(t)=\sum_{n=0}^{\infty}\mathbf{1}_{\{\tau(n)=t\}}p(n). \quad (34)$$

If $p(n)$'s are indicator random variables (e.g., routing variables), then $r(t)$ is a thinning process of $a(t)$ and can be viewed as the number of customers that are routed to a particular queue at time t.

Let $P(m_1,m_2)=\sum_{n=m_1+1}^{m_2}p(n)$ and $R(t_1,t_2)=\sum_{t=t_1+1}^{t_2}r(t)$. Note that

$$R(t_1,t_2)=P(A(0,t_1),A(0,t_2)). \quad (35)$$

Lemma 4.5 *If $\{a(t),t\ge 0\}$ and $\{p(n),n\ge 0\}$ are* independent *and they have linear UEPs (α,σ) and (γ,δ) respectively, then $\{r(t),t\ge 0\}$ has a linear UEP $(\alpha\gamma,\delta+\gamma\sigma)$, where δ and σ are* independent.

Proof. Analogous to the proof of Lemma 4.3, it suffices to show that $\max_{0\le s\le t}[R(t-s,t)-\alpha\gamma s]\le_{st}\delta+\gamma\sigma$. We will show this by construction as in the proof of Lemma 2.2(i). From (35), it follows that

$$\begin{aligned}
&\max_{0\le s\le t}[R(t-s,t)-\alpha\gamma s]\\
&=\max_{0\le s\le t}[P(A(0,t-s),A(0,t))-\gamma A(t-s,t)\\
&+\gamma A(t-s,t)-\alpha\gamma s]\\
&\le \max_{0\le s\le t}[P(A(0,t-s),A(0,t))-\gamma A(t-s,t)]\\
&+\gamma\max_{0\le s\le t}[A(t-s,t)-\alpha s]\\
&=\max_{0\le s\le A(0,t)}[P(A(0,t)-s,A(0,t))-\gamma s]\\
&+\gamma\max_{0\le s\le t}[A(t-s,t)-\alpha s]. \qquad (36)
\end{aligned}$$

Since we assume that $\{a(t),t\ge 0\}$ has a linear UEP (α,σ), we can construct two processes $\{\tilde{A}(t-s,t),0\le s\le t\}$ and $\{\alpha s+\tilde{\sigma},0\le s\le t\}$ on the same probability space such that (i) $\{\tilde{A}(t-s,t),0\le s\le t\}=_{st}\{A(t-s,t),0\le s\le t\}$, (ii) $\tilde{\sigma}=_{st}\sigma$, and (iii) $\tilde{A}(t-s,t)\le\alpha s+\tilde{\sigma},0\le s\le t$, a.s. From (iii), it follows that $\max_{0\le s\le t}[\tilde{A}(t-s,t)-\alpha s]\le\tilde{\sigma}$, a.s. Also, since $a(t)\ge 0$ for all t, $\tilde{A}(0,t)\ge 0$. From the independent assumption of $\{a(t),t\ge 0\}$ and $\{p(n),n\ge 0\}$ and the assumption that $\{p(n),n\ge 0\}$ has a linear UEP (γ,δ), we can also construct two processes $\{\tilde{P}(\tilde{A}(0,t)-s,\tilde{A}(0,t)),0\le s\le t\}$ and $\{\gamma s+\tilde{\delta},0\le s\le t\}$ on the same probability space such that (i) $\{\tilde{P}(\tilde{A}(0,t)-s,\tilde{A}(0,t)),0\le s\le t\}=_{st}\{P(A(0,t)-s,A(0,t)),0\le s\le t\}$, (ii) $\tilde{\delta}=_{st}\delta$, (iii) $\tilde{P}(\tilde{A}(0,t)-s,\tilde{A}(0,t))\le\gamma s+\tilde{\delta},0\le s\le t$, a.s., and (iv) $\tilde{\delta}$ is independent of $\tilde{\sigma}$. From these constructions, one has

$$\begin{aligned}
&\max_{0\le s\le\tilde{A}(0,t)}[\tilde{P}(\tilde{A}(0,t)-s,\tilde{A}(0,t))-\gamma s]\\
&+\gamma\max_{0\le s\le t}[\tilde{A}(t-s,t)-\alpha s]\le\tilde{\delta}+\gamma\tilde{\sigma}. \qquad (37)
\end{aligned}$$

Thus, one has

$$\begin{aligned}
&\max_{0\le s\le A(0,t)}[P(A(0,t)-s,A(0,t))-\gamma s]\\
&+\gamma\max_{0\le s\le t}[A(t-s,t)-\alpha s]\le_{st}\delta+\gamma\sigma, \qquad (38)
\end{aligned}$$

where δ and σ are independent. In conjunction with (36), one derives the desired result. ■

Corollary 4.6 *If $\{a(t),t\ge 0\}$ and $\{p(n),n\ge 0\}$ are* independent *and they have families of linear UEPs $\{(\hat{a}(\theta),\hat{\sigma}(\theta)+\sum_{i=1}^{k}Z_{1,i}(\theta)),\theta>0\}$ and $(\{\hat{p}(\theta),\hat{\delta}(\theta)+Z_2(\theta)),\theta>0\}$, where $Z_{1,i}(\theta)$, $i=1,\dots,k$, are independent and exponentially distributed with mean $1/\theta$, and $\{Z_2(\theta),\theta>0\}$ is a family of random variables* *parametrized by θ, then $\{r(t), t \geq 0\}$ has a family of linear UEPs*

$$\Big\{ \Big(\hat{p}(\theta)\hat{a}(\theta\hat{p}(\theta)), \hat{\delta}(\theta) + \hat{p}(\theta)\hat{\sigma}(\theta\hat{p}(\theta)) + \sum_{i=1}^{k} Z_{1,i}(\theta) + Z_2(\theta)\Big), \theta > 0\Big\}, \tag{39}$$

where $Z_{1,i}(\theta)$, $i = 1, \ldots, k$, and $Z_2(\theta)$ are independent.

This corollary is achieved by choosing $\theta_1 = \theta\hat{p}(\theta)$ in the family of UEPs of $a(t)$. Also, note that $Z_{1,i}(\theta) =_{st} \hat{p}(\theta)Z_{1,i}(\theta\hat{p}(\theta))$ $i = 1, \ldots, k$, since $Z_{1,i}$, $i = 1, \ldots, k$, are exponentially distributed.

Now consider a discrete-time intree network with J queues and multiple classes of customers as in [8, 7]. In such a network, there is a natural partial ordering among the J queues. If part of the departure process of queue i is an input process of queue j, we say queue j is a *successor* of queue i and queue i is a *predecessor* of queue j. Let $Suc(j)$ (resp. $Pre(j)$) be the set of queues that are *successors* (resp. predecessors) of queue j. We label the J queues in the *postorder* (left subtree, right subtree, root). Since the network is an intree, $Suc(i)$ only contains one element. Also, we have from the postorder labelling that $i < j$ for all $i \in Pre(j)$ and $Suc(j) > j$.

We assume that routing only depends on the queue that customers depart from. To model this, we let $\{p_j(n), n \geq 1\}, j = 1 \ldots, J-1$, be $J-1$ sequences of routing random variables, i.e., $p_j(n) = 1$ if the n^{th} departure from queue j is routed to the successor of queue j, i.e., queue $Suc(j)$, and $p_j(n) = 0$ if the n^{th} departure from queue j leaves the intree network. Besides the arrivals from the queues in the predecessors, there are k_j external arrival processes coming to queue j with $\{a_{j,k}(t), t = 0, 1, 2, \ldots\}$, $k = 1, \ldots, k_j$, denoting the number of arrivals at time t.

Denote $q_j(t)$ and c_j the queue length of queue j at time t and its capacity, respectively. We assume that each queue is governed by Lindley's equation, i.e.,

$$q_j(t+1) = (q_j(t) + a_j(t+1) - c_j)^+, \;\; j = 1, \ldots, J, \tag{40}$$

where $a_j(t)$ is the total number of arrivals at queue j at time t, and $(x)^+ \equiv \max(0, x)$.

To analyze such a network, we make the following assumptions on the arrival processes and the routing processes:

(A1) All the external arrival processes and routing processes are *independent* and independent of all else.

(A2) $\{a_{j,k}(t), t \geq 0\}, j = 1, \ldots, J, k = 1, \ldots, k_j$, has a family of linear UEPs $\{(\hat{a}_{j,k}(\theta), \hat{\sigma}_{j,k}(\theta) + Z_{j,k}(\theta)), \theta > 0\}$, where $Z_{j,k}(\theta)$ is exponentially distributed with mean $1/\theta$.

(A3) $\{p_j(n), n \geq 0\}, j = 1, \ldots, J-1$, has a family of linear UEPs $\{(\hat{p}_j(\theta), \hat{\delta}_j(\theta) + Y_j(\theta)), \theta > 0\}$, where $Y_j(\theta)$ is exponentially distributed with mean $1/\theta$.

(A4) $\{a_{j,k}(t), t = 0, 1, 2, \ldots\}$, $k = 1, \ldots, k_j, j = 1, \ldots J$ and $\{p_j(t), t = 0, 1, 2, \ldots, \}, j = 1, \ldots, J-1$, are *stationary* and *ergodic* processes. Moreover, there exist stationary queue lengths q_j, $j = 1, \ldots, J$.

Using Corollaries 4.2, 4.4 and 4.6, we have the following algorithm for bounding the tail distributions of queue lengths in such a network:

Algorithm 4.7 **1.** *For $j = 1, \ldots, J$, define the following recursively:*

$$\hat{a}_j(\theta) = \sum_{k=1}^{k_j} \hat{a}_{j,k}(\theta) + \sum_{k \in P(j)} \hat{p}_k(\theta)\hat{a}_k(\theta\hat{p}_k(\theta)) \tag{41}$$

$$\hat{\sigma}_j(\theta) = \sum_{k=1}^{k_j} \hat{\sigma}_{j,k}(\theta) + \sum_{k \in P(j)} [\delta_k(\theta) + \hat{p}_k(\theta)\hat{\sigma}_k(\theta\hat{p}_k(\theta))] \tag{42}$$

$$Z_j(\theta) = \sum_{k=1}^{k_j} Z_{j,k}(\theta) + \sum_{k \in P(j)} [Y_k(\theta) + Z_k(\theta)], \tag{43}$$

where $Z_{j,k}(\theta)$, $Y_k(\theta)$ and $Z_k(\theta)$ in (43) are independent.

2. *If there exists a $\theta_J^* > 0$ from the equation $\hat{a}_J(\theta) = c_J$, then*

$$q_J \leq_{st} \hat{\sigma}_J(\theta_J^*) + Z_J(\theta_J^*). \tag{44}$$

Note that if the intree network is connected, i.e., every external arrival process and routing process contributes to the J^{th} queue, then $Z_J(\theta)$ is simply an L-stage Erlang random variable, where $L = J - 1 + \sum_{j=1}^{J} k_j$. Thus,

$$Pr(q_J \geq x) \leq e^{\theta_J^*\hat{\sigma}(\theta_J^*)} \sum_{\ell=0}^{L-1} \frac{(\theta_J^*(x - \hat{\sigma}(\theta_J^*))^+)^\ell e^{-\theta_J^* x}}{\ell!}. \tag{45}$$

Also, the recursive equation in (41) is different from that in Chang [7], where the exact decay rate of q_J is computed. It can be shown that θ_J^* computed by this algorithm is not greater than the exact decay rate of q_J. The gap is due to the input-output relation obtained in this paper, where we simply choose $\hat{a}_k(\theta)$. Comparing this with the corresponding equation in [7], we note that θ^* matches the exact decay rate if it is small enough.

5 Conclusions

In this paper, we refined the calculus proposed in [5, 8, 9] so that tighter exponential bounds for the tail distributions of queue lengths in intree networks with routing could be computed via Algorithm 4.7. The key discovery was that the characterization (linear UEP) of a point process in [9] could be obtained by feeding the point process into a queue with a constant capacity. By varying the capacity of a queue, each point process was then associated with a family of UEPs. For processes that satisfy a sample path criterion in [8], e.g., Markov arrival processes and autoregressive processes, the linear UEP characterization via the corresponding stationary queue length was established in Theorem 3.2. In the same theorem, a lower bound was also obtained. By establishing the closure properties of the linear UEPs for network operations such as multiplexing, input-output operation, and routing, computable exponential upper bounds were obtained.

We performed various numerical examples and found that the upper and lower bounds for a single queue in Theorem 3.2 are good when the number of arrival processes is small. (Due to page limitation, we are not able to present these examples in this conference paper.) However, they become loose when the number of arrival processes is large. The upper bound derived by Algorithm 4.7 was not as good as that in Theorem 3.2. However, to the best of knowledge of the literature, it is the only computable exponential upper bound for the intree network case. We note that for a single queue with Markov modulated inputs, tighter bounds than those in Theorem 3.2 were obtained in recent papers [1, 27].

References

[1] S. Asmussen and T. Rolski, "Risk theory in a periodic enviroment: the Cramér-Lundberg approximation and Lundberg's inequality," *Math. Operat. Res*, Vol. 19, No. 2, pp. 410-433, 1994.

[2] F. Baccelli and P. Bremaud, *Palm Probabilities and Stationary Queues*. New York: Springer-Verlag, 1987.

[3] C.S. Chang, "Stability, queue length and delay, Part I: Deterministic queueing networks," *IEEE Conference on Decision and Control*, Vol 1, pp. 999-1004, 1992.

[4] C.S. Chang, "Stability, queue length and delay, Part II: Stochastic queueing networks," *IEEE Conference on Decision and Control*, Vol 1, pp. 1005-1010, 1992.

[5] C.S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Transactions on Automatic Control*, Vol.39, pp. 913-931, 1994.

[6] C.S. Chang, "Approximations of ATM networks: effective bandwidths and traffic descriptors," *IBM RC 18954*, 1993.

[7] C.S. Chang, "Sample path large deviations and intree networks," to appear in *Queueing Systems*.

[8] C.S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin, "Effective Bandwidth and Fast Simulation of ATM Intree Networks," *Performance Evaluation*, Vol. 20, pp. 45-66, 1994.

[9] C.S Chang, J.A. Thomas, and S.-H. Kiang, "On the stability of open networks: an unified approach by stochastic dominance," *Queueing Systems*, Vol. 15, pp. 239-260, 1994.

[10] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the most out of ATM," *preprint*, 1993.

[11] Y.S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*. New York: Springer-Verlag, 1988.

[12] R.L. Cruz, "A calculus for network delay, Part I: Network elements in isolation," *IEEE Tran. Inform. Theory*, Vol. 37, pp. 114-131, 1991.

[13] R.L. Cruz, "A calculus for network delay, Part II: Network analysis," *IEEE Tran. Inform. Theory*, Vol. 37, pp. 132-141, 1991.

[14] N.G. Duffield, "Exponential bounds for queues with Markovian arrivals," *preprint*, 1993.

[15] R. Ellis, "Large deviations for a general class of random vectors, " *Ann. Probab.*, Vol. 12, pp. 1-12, 1984.

[16] A.I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, Vol. 1, pp. 329-343, 1993.

[17] J. Gärtner, "On large deviations from invariant measure," *Theory Probab. Appl.*, Vol. 22, pp. 24-39, 1977.

[18] R.J. Gibbens and P.J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Systems*, Vol. 9, pp. 17-28, 1991.

[19] P.W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," *J. Appl. Prob*, Vol. 31A, pp. 131-156, 1994.

[20] P.W. Glynn and W. Whitt, "Large deviations behavior of counting processes and their inverses," *Queueing Systems*, Vol. 17, pp. 107-128, 1994.

[21] R. Guérin, H. Ahmadi and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.*, Vol. 9, pp. 968-981, 1991.

[22] S. Karlin and H. Taylor. *A First Course in Stochastic Processes*. New York: Academic Press, 1975.

[23] F.P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Systems*, Vol. 9, pp. 5-16, 1991.

[24] G. Kesidis and J. Walrand, "Traffic policing and enforcement of effective bandwidth constraints in ATM networks," *preprint*.

[25] G. Kesidis, J. Walrand and C.S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, Vol. 1, pp. 424-428, 1993.

[26] J.F. Kurose, "On computing per-session performance bounds in high-speed multi-hop computer networks," to be presented in *SIGMETRICS and PERFORMANCE'92*.

[27] Z. Liu, P. Nain and D. Towsley, "Exponential bounds with an applications to call admission," *submitted*. Conference version in IEEE CDC'94.

[28] R.M. Loynes, "The stability of a queue with non-independent interarrival and service times," *Proc. Camb. Phil. Soc.*, Vol. 58, pp. 497-520, 1962.

[29] J.F. Kurose, "Open Issues and challenges in providing quality of service guarantees in high-speed networks," *preprint*.

[30] R. Nagarajan, J. Kurose and D. Towsley, "Local allocation of end-to-end quality-of-service in high speed networks," *preprint*.

[31] S.M. Ross, *Stochastic Processes*. New York: J. Wiley & Sons, 1983.

[32] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, Berlin: J. Wiley & Sons, 1983.

[33] W. Whitt, "Tail probability with statistical multiplexing and effective bandwidths in multi-class queues," *Telecommunication Systems*, Vol. 2, pp. 71-107, 1993.

[34] G. De. Veciana and J. Walrand, "Effective Bandwidths: call admission, traffic policing & filtering for ATM networks," Submitted to *IEEE/ACM Trans. Networking*, 1993.

[35] G. De. Veciana, C. Courcoubetis and J. Walrand, "Decoupling bandwidths for networks: a decomposition approach to resource management," Submitted to *IEEE/ACM Trans. Networking*, 1993.

[36] D. Yates, J. Kurose, D. Towsley and M. Hluchyj, "On per-session end-to-end delay distributions and the call admission problem for real-time applications with QOS requirements, " *preprint*.