A Service Abstraction with Applications to Network Calculus *

Rene L. Cruz
Dept. of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA
cruz@ece.ucsd.edu

Alberto P. Blanc
Dept. of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA
ablanc@ucsd.edu

Abstract

In this paper, we define a general service abstraction for a network element. The service abstraction is defined in terms of a "service mapping," which is a montotone operator that maps an arrival process to a lower bound on the corresponding departure process. We consider service mappings which are shift invariant, which includes the previously studied service curve model defined in terms of convolution in the "min-plus" algebra. For a network element with a shift invariant service mapping, we obtain bounds on the maximum delay, maximum backlog, and a traffic envelope for the departure process, assuming that the arrival process to the network element conforms to a traffic envelope. These bounds have a unified graphical interpretation and reduce to previously known bounds in the case of a service curve model.

Next, we apply the service mapping abstraction to a First-In First-Out (FIFO) multiplexer whose arrival processes conform to traffic envelopes. The corresponding service mapping is *non-linear* in the min-plus algebra. We note that the quality of service bounds reduce to previously known tight bounds. We then consider a tandem configuration of FIFO multiplexers. Analysis of the tandem network with the service mapping abstraction yields achievable upper bounds on end-to-end delay which are smaller than those that can be obtained with previously proposed methods.

1 Introduction

Over the past several years, a deterministic theory for analysis of networks of queues has been developed, now commonly termed as "network calculus." We now briefly review some of the important developments in this area of research. The interested reader is referred to [4] and [10], which documents much of this work in detail.

Networks of queues operating with FIFO or fixed priority scheduling were analyzed by Cruz using a deterministic traffic model in [5][6]. Subsequently, Parekh and Gallager [12][13] analyzed networks of queues operating with generalized processor sharing using

*This work was partially supported by the U.S. Department of Energy under grant no.DE-FG03-99ER25394

this deterministic traffic model, and thereby stimulated a large body of work by other researchers on “fair queueing” scheduling algorithms [15]. A key concept in Parekh and Gallager’s work was that of a “universal service curve.” This concept was extended to a *service abstraction* for general network elements by Cruz in [7]. This service model was refined independently by Le Boudec [9] and Sariowan [14]. Le Boudec formalized the notion of the convolution and deconvolution operators, and C. S. Chang [3] first observed that a service curve is analogous to an “impulse response” in the theory of linear time invariant systems. Chang exploited this analogy to provide simple explanations for results on traffic regulators developed in [5].

The insight obtained from these developments are apparent in a framework [2] for deterministic quality of service guarantees proposed for the Internet by the *intserv* working group of the Internet Engineering Task Force (IETF). In the intserv model, traffic is managed on a per-flow basis, leading many researchers to question its capability to scale to large networks. Another working group of the IETF, *diffserv*, has aimed to address scalability by developing a framework where traffic is managed at a coarser granularity than the level of flow [11]. This led to a renewed interest in performance results for FIFO queueing, since FIFO queues do not require per flow traffic management.

In [8], Cruz presented a service curve characterization for a FIFO multiplexer. In this paper, we generalize and explore this characterization further. As we will see, this leads to a simplified understanding of previous results in network calculus, as well as the tightening of bounds for delay in networks of FIFO queues for deterministic traffic models.

The remainder of this paper is organized as follows. In the next section, we define a general service abstraction for network elements. We derive performance bounds in the context of this service abstraction. These performance bounds have a simple unified graphical interpretation that we illustrate. We see that the service curve model is a special case of this this general service abstraction, and the performance bounds we obtain reduce to previously known results. In Section 4, we apply our service abstraction to the context of FIFO multiplexing. We will see in Section 4.1 that our performance bounds for a single FIFO multiplexer reduce to those tight bounds previously obtained in [5]. In Section 4.2 we apply the general service abstraction to analyze a tandem configuration of FIFO multiplexers. We will see that this results in improved bounds to end-to-end delay which are achievable.

2 Service Models

Consider a network element, which is an abstration of a queueing system. For example, the queueing system might represent a packet switch, or more generally an entire network. In this paper, a network element is an abstraction defined for the purposes of describing a single stream of information passing through the associated queueing system. Specifically, the network element has an arrival process and a departure process, described by two functions of time $R_{in}(\cdot)$ and $R_{out}(\cdot)$, respectively. The value of $R_{in}(t)$ is defined as the number of bits that have arrived to the network element up to time t, and similarly $R_{out}(t)$ is the number of bits that have departed the network element up to time t. The *backlog* $B(t)$ of the network element at time t is defined as $B(t) = R_{in}(t) - R_{out}(t)$, i.e. it is the number of bits stored inside the network element at time t. The *virtual delay* of the network element at time t is defined as $D(t) = \inf\{d : d \geq 0 \text{ and } R_{out}(t + d) \geq R_{in}(t)\}$. For example, if the arriving bits depart the network element in first-in first-out (FIFO) order, then a bit that arrives at time t waits no longer than $D(t)$ seconds before departing the network element.

In general, the departure process $R_{out}(\cdot)$ may not be determined solely by the arrival process R_{in}, but also by external events in the associated queueing system. However, we can partially characterize the network element by bounding the departure process in terms of the arrival process. In general, we assume that the arrival process $R_{in}(\cdot)$ and departure process $R_{out}(\cdot)$ can be arbitrary non-decreasing functions. Formally, $R_{in}(\cdot)$ and $R_{out}(\cdot)$ are elements of $\mathcal{M}$, which is defined as the set of all non-decreasing functions whose domain is the set of all real numbers $\mathcal{R}$ and whose range is the extended set of real numbers $\mathcal{R} \cup \{+\infty\}$.

2.1 Minimum Service Mappings

Let $\mathcal{S} : \mathcal{M} \to \mathcal{M}$ be a given operator, which maps elements of $\mathcal{M}$ into elements of $\mathcal{M}$. Given functions of time $F(\cdot)$ and $G(\cdot)$, we use the notation $F \leq G$ if $F(t) \leq G(t)$ for all t. We say that $\mathcal{S}$ is *monotone* if $F \leq G$ implies that $\mathcal{S}(F) \leq \mathcal{S}(G)$ for all F and G.

Definition 1 *Suppose a network element is such that $R_{out} \geq \mathcal{S}(R_{in})$ for all possible arrival processes R_{in}, for some monotone operator $\mathcal{S}$. In this case, say that $\mathcal{S}$ is a (minimum) service mapping for the network element, and we write $R_{in} \to \mathcal{S} \to R_{out}$.*

The service model in the previous definition is composable in the sense that if several network elements are configured in tandem, each being described by a service mapping, then the composition of the service mappings is a service mapping of the entire system. This is stated formally in the following theorem for the case of two network elements in tandem.

Theorem 1 (Network Elements in Tandem) *Suppose $R_0 \to \mathcal{S}_1 \to R_1$ and $R_1 \to \mathcal{S}_2 \to R_2$. Then $R_0 \to (\mathcal{S}_1 \circ \mathcal{S}_2) \to R_2$, where $(\mathcal{S}_1 \circ \mathcal{S}_2)(F) = \mathcal{S}_2(\mathcal{S}_1(F))$.*

Proof. Fix any t. We have

$$\begin{aligned} R_2 &\geq \mathcal{S}_2(R_1) \\ &\geq \mathcal{S}_2(\mathcal{S}_1(R_0)) \\ &= (\mathcal{S}_1 \circ \mathcal{S}_2)(R_0) \ . \end{aligned}$$

The second inequality above follows since $\mathcal{S}_2$ is monotone. $\diamond$

2.2 Shift Invariant Service Mappings

An operator $\mathcal{S}$ is said to be *time invariant* if $\mathcal{S}(F) = G$ implies that $\mathcal{S}(F_\Delta)) = G_\Delta$ for all $\Delta \in \mathcal{R}$, where $F_\Delta(t) = F(t - \Delta)$ and $G_\Delta(t) = G(t - \Delta)$ for all t. An operator $\mathcal{S}$ is said to be *space invariant* if $\mathcal{S}(F) = G$ implies that $\mathcal{S}(k + F) = k + G$ for all constants k. An operator $\mathcal{S}$ that is both time invariant and space invariant is called *shift invariant*.

An operator $\mathcal{S}$ is called *additive* if $\mathcal{S}(F_1) = G_1$ and $\mathcal{S}(F_2) = G_2$ imply that $\mathcal{S}(F_1 \wedge F_2) = G_1 \wedge G_2$, where we use the notation $H_1 \wedge H_2$ to denote the function defined by $(H_1 \wedge H_2)(t) = \min\{H_1(t), H_2(t)\}$. An operator that is both space invariant and additive is said to be *linear*.

As an example, suppose that $S(\cdot)$ is a minimum service curve [14][9][8], as we now define. The convolution of two functions $F(\cdot)$ and $G(\cdot)$, $F * G$, is first defined as

$$(F * G)(t) = \inf_{\tau}\{F(\tau) + G(t-\tau)\}$$

for all t. The function $S(\cdot)$ is said to be a (minimum) service curve for the network element if for any arrival process R_{in} we have $R_{out} \geq R_{in} * S$. This service model is special case of a service mapping where the associated operator is both linear and time invariant. Conversely, it can be seen that any service mapping that is both linear and time invariant can be equivalently be described in terms of a service curve.

Although the space of service models which are linear is adequate for analyzing many queueing systems of interest, in this paper we make the proposition that non-linear service models are useful as well. In particular, we will demonstrate how they can be used to analyze networks of FIFO queues. For example, suppose that an arrival process R_{in} is multiplexed in a FIFO manner with another arrival process R^x_{in}, such that the aggregate arrival process $R_{in} + R^x_{in}$ results in the aggregate departure process $R_{out} + R^x_{out}$ with a service curve of G, i.e. $(R_{out} + R^x_{out})(t) \geq ((R_{in} + R^x_{in}) * G)(t)$ for all t. If $R^x_{in} \leq R^x_{in} * E^x$, it is known [8] that

$$R_{out} \geq R_{in} * S_T$$

for all $T \geq 0$, where

$$S_T(t) = \begin{cases} [G(t) - E^x(t-T)]^+ & \text{if } t \geq T \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where we use the notation $x^+ = \max\{x, 0\}$. In this case, in fact we have $R_{in} \rightarrow \hat{\mathcal{S}} \rightarrow R_{out}$, where the operator $\hat{\mathcal{S}}$ is defined for $F \in \mathcal{M}$ as

$$\hat{\mathcal{S}}(F)(t) = \sup_{T:T\geq 0} \left[(F * S_T)(t)\right] . \tag{2}$$

It can be verified that $\hat{\mathcal{S}}$ is shift invariant, but not necessarily linear.

This motivates us to look at the class of service models characterized by shift invariant service mappings. For generality, we do not necessarily assume that service mappings are of the form given in (2). We shall obtain bounds on delay, backlog, and a traffic envelope (defined below) for the departure process, in the context of service models characterized by shift invariant service mappings.

To begin with, we first define the notion of a traffic envelope [5]. Given a function $E(\cdot)$, and an arrival process R, we say that R has envelope E if $R \leq R * E$. Note that the inequality $R \leq R * E$ is equivalent to

$$R(u) \geq R(t) - E(t-u) \text{ for all } u \tag{3}$$

for any fixed value of t. In other words, if R has envelope E, then for any fixed t we have

$$R \geq R_{E,t} \ , \tag{4}$$

where $R_{E,t}(u) = R(t) - E(t-u)$ for all u. Note that in general $E(x)$ may be non-zero for negative values of x, although it is common to assume that $E(x) = 0$ for $x < 0$. For $x < 0$, the value of $-E(x)$ represents a *lower* bound on the increments of a process over any interval of length $-x$.

Before proceeding further, let us make a few definitions. First, given any function $E(\cdot)$, define the "tilde" operator as follows:

$$\tilde{E}(t) = -E(-t) \text{ for all } t. \tag{5}$$

Given any $R \in \mathcal{M}$, define $D_{t,k}(R) = \inf\{d : d \geq 0 \text{ and } R(t+d) \geq k\}$. Note that the virtual delay in a system with arrival and departure processes R_{in} and R_{out} is $D(t) = D_{t,R_{in}(t)}(R_{out})$. Finally, for any $R \in \mathcal{M}$, define $D_0(R) = D_{0,0}(R)$.

3 Quality of Service Guarantees for Shift Invariant Service Mappings

In this section we consider a single network element whose arrival process has envelope E. We suppose the network element has a shift invariant service mapping and derive bounds on virtual delay and backlog. We also find an envelope for the departure process.

Theorem 2 *Suppose $\mathcal{S}$ is a shift invariant service mapping for a network element. Suppose the arrival process to the network element has envelope E. Then the virtual delay $D(t)$ is upper bounded according to*

$$D(t) \leq D_0(\mathcal{S}(\tilde{E})) \textit{ for all } t. \tag{6}$$

Proof. Fix any t. We have

$$\begin{aligned}
D(t) &= \inf\{d : d \geq 0 \text{ and } R_{out}(t+d) \geq R_{in}(t)\} \\
&= D_{t,R_{in}(t)}(R_{out}) \\
&\leq D_{t,R_{in}(t)}(\mathcal{S}(R_{in})) \\
&\leq D_{t,R_{in}(t)}(\mathcal{S}((R_{in})_{E,t})) \\
&= D_{t,R_{in}(t)}(\mathcal{S}(R_{in}(t) - E(t - \cdot))) \\
&= D_{t,R_{in}(t)}(R_{in}(t) + \mathcal{S}(-E(t - \cdot))) \\
&= D_{t,0}(\mathcal{S}(-E(t - \cdot))) \\
&= D_{t,0}(\mathcal{S}(-E(-(\cdot - t)))) \\
&= D_{0,0}(\mathcal{S}(-E(-(\cdot)))) \\
&= D_0(\mathcal{S}(\tilde{E}))\ .
\end{aligned}$$

In the above sequence the first equality follows from the definition of virtual delay, the second equality follows from the definition of $D_{t,k}(R)$, the third inequality follows since $\mathcal{S}$ is a minimum service mapping and $D_{t,k}(R)$ is monotone decreasing in R. The fourth inequality follows from (4) and the monotonicity of $\mathcal{S}$. The remaining equalities follow from the shift invariance of $\mathcal{S}$. $\diamond$

Theorem 3 *Suppose $\mathcal{S}$ is a shift invariant service mapping for a network element. Suppose the arrival process to the network element has envelope E. Then the backlog $B(t)$ is upper bounded according to*

$$B(t) \leq \tilde{\mathcal{S}}(\tilde{E})(0) \textit{ for all } t. \tag{7}$$

Proof. Fix any t. We have

$$\begin{aligned}
B(t) &= R_{in}(t) - R_{out}(t) \\
&\leq R_{in}(t) - \mathcal{S}(R_{in})(t) \\
&\leq R_{in}(t) - \mathcal{S}((R_{in})_{E,t})(t) \\
&= R_{in}(t) - \mathcal{S}(R_{in}(t) - E(t - \cdot))(t) \\
&= -\mathcal{S}(-E(t - \cdot))(t) \\
&= -\mathcal{S}(-E(-(\cdot - t)))(t) \\
&= -\mathcal{S}(-E(-(\cdot)))(0) \\
&= -\mathcal{S}(\tilde{E})(0) \\
&= \tilde{\mathcal{S}}(\tilde{E})(0
\end{aligned}$$

In the above sequence the first equality follows from the definition of backlog, the second equality follows since $\mathcal{S}$ is a minimum service mapping. The third inequality follows from (4) and the monotonicity of $\mathcal{S}$. The remaining equalities follow from the shift invariance of $\mathcal{S}$. ◇

We say a network element is *conservative* if we always have $R_{out} \leq R_{in}$.

Theorem 4 *Suppose $\mathcal{S}$ is a shift invariant service mapping for a conservative network element. Suppose the arrival process to the network element has envelope E. Then the departure process has envelope E_{out} where*

$$E_{out} = \tilde{\mathcal{S}}(\tilde{E}) \ . \tag{8}$$

Proof. Fix any s, t. We have

$$\begin{aligned}
R_{out}(t) - R_{out}(s) &\leq R_{in}(t) - R_{out}(s) \\
&\leq R_{in}(t) - \mathcal{S}(R_{in})(s) \\
&\leq R_{in}(t) - \mathcal{S}((R_{in})_{E,t})(s) \\
&= R_{in}(t) - \mathcal{S}(R_{in}(t) - E(t - \cdot))(s) \\
&= -\mathcal{S}(-E(t - \cdot))(s) \\
&= -\mathcal{S}(-E(-(\cdot - t)))(s) \\
&= -\mathcal{S}(-E(-(\cdot)))(s - t) \\
&= -\mathcal{S}(\tilde{E})(s - t) \\
&= \tilde{\mathcal{S}}(\tilde{E})(t - s) \ . \qquad (9)
\end{aligned}$$

In the above sequence the first equality follows since the network element is conservative, the second equality follows since $\mathcal{S}$ is a minimum service mapping. The third inequality follows from (4) and the monotonicity of $\mathcal{S}$. The remaining equalities follow from the shift invariance of $\mathcal{S}$. ◇

Note that (8) can be rewritten as $\tilde{E}_{out} = \mathcal{S}(\tilde{E})$, which has a simple intuitive appeal. The theorems of this section are illustrated graphically in Figure 1.

3.1 Linear Time Invariant Service Mappings

We conclude this section by considering a service mapping $\mathcal{S}$ that is linear and time invariant. In other words, the service mapping corresponds to a service curve guarantee, i.e.

$$\mathcal{S}(R) = R * S$$

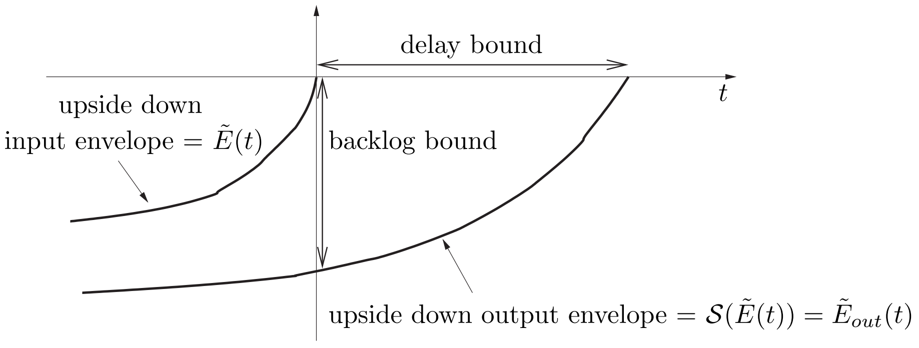
delay bound
t
upside down
input envelope = $\tilde{E}(t)$
backlog bound
upside down output envelope = $\mathcal{S}(\tilde{E}(t)) = \tilde{E}_{out}(t)$

Figure 1: $\mathcal{S}$-mapping: delay, backlog and output bounds

where S is a service curve for the network element. In this case we show that the theorems of the previous subsection reduce to previously known results. We assume that the arrival process conforms to the envelope E.

Observe that

$$\begin{aligned}
\mathcal{S}(\tilde{E})(-x) &= (\tilde{E} * S)(-x) \\
&= \inf_y \{\tilde{E}(-x-y) + S(y)\} \\
&= \inf_y \{-E(y+x) + S(y)\} \\
&= -\sup_y \{E(y+x) - S(y)\} \\
&= -(E \oslash S)(x) \ , \qquad (10)
\end{aligned}$$

where we use "$\oslash$" to denote the deconvolution operator, i.e. $(F \oslash G)(x) = \sup_y\{F(x+y) - G(y)\}$ for all x.

First, consider the result of Theorem 4. The departure process has envelope E_{out} where $E_{out} = \tilde{\mathcal{S}}(\tilde{E})$. In view of (10), we have $E_{out} = E \oslash S$, which agrees with the result in [1].

Second, consider the result of Theorem 3. The backlog is upper bounded by $-\mathcal{S}(\tilde{E})(0)$. In view of (10), we thus have $B(t) \leq (E \oslash S)(0) = \sup_y\{E(y) - S(y)\}$, which agrees with the result in [1].

Third, consider the result of Theorem 2. Using (10), the virtual delay $D(t)$ is upper bounded as follows:

$$\begin{aligned}
D(t) &\leq D_0(\mathcal{S}(\tilde{E})) \\
&= \inf\{d : d \geq 0 \text{ and } \mathcal{S}(\tilde{E})(d) \geq 0\} \\
&= \inf\{d : d \geq 0 \text{ and } -(E \oslash S)(-d) \geq 0\} \\
&= \inf\{d : d \geq 0 \text{ and } -\sup_y\{E(y-d) - S(y)\} \geq 0\} \\
&= \inf\{d : d \geq 0 \text{ and } \sup_y\{E(y-d) - S(y)\} \leq 0\} \\
&= \inf\{d : d \geq 0 \text{ and } E(y-d) \leq S(y) \text{ for all } y\} \ ,
\end{aligned}$$

which agrees with the result in [1], i.e. the delay in upper bounded by the "maximum horizontal distance" between the graphs of E and S.

These results are illustrated graphically in Figure 2 in the context of the graph of $E \oslash S$.

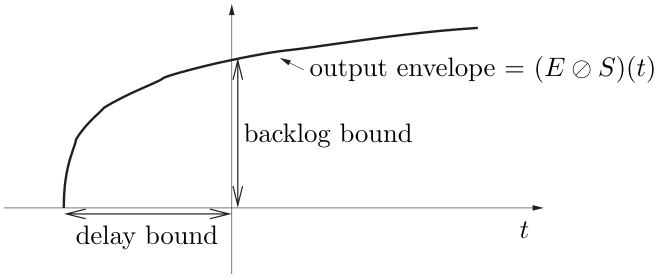

Figure 2: Service curves: delay, backlog and output bounds

4 FIFO Multiplexers

In this section, we apply the results in the previous section to the case in which network elements correspond to FIFO multiplexers. First, in the next section, we consider a single FIFO multiplexer.

4.1 A Single FIFO Multiplexer

Consider two arrival streams incident on a FIFO multiplexer, described by R_{in} and R^x_{in}, with traffic envelopes E and E^x, respectively. The multiplexer serves data in a FIFO manner as fast as possible with a maximum service rate of C bits per second. From the point of view of the aggregate arrival stream, $R_{in} + R^x_{in}$, a service curve of G is provided, where $G(x) = Cx$ if $x \geq 0$ and $G(x) = 0$ otherwise. In other words, if the corresponding departure streams are denoted as R_{out} and R^x_{out}, then we have $R_{out} + R^x_{out} \geq G*(R_{in} + R^x_{in})$.

If we assume that packets are served non-preemptively, then if we have non-zero packet sizes (i.e. a non-fluid model), then bits might not depart in exactly FIFO order. Therefore, for simplicity we assume a fluid model, which corresponds to "$L = 0$" as discussed in [5].

In this case, from Theorem 4.1 of [5], it is known that the delay for stream R_{in} is upper bounded by $\bar{D}_{FCFSMUX}$, i.e. $R_{out}(t + \bar{D}_{FCFSMUX}) \geq R_{in}(t)$ for all t, where

$$\bar{D}_{FCFSMUX} = \frac{1}{C} \max_{u \geq 0} \left[E(u) + E^x(u) - Cu\right] \ . \tag{11}$$

Since the system is FIFO, we have using the result from [8] that $R_{out} \geq R_{in} * S_T$ holds for any $T \geq 0$, with S_T given in (1). Thus we have $R_{in} \rightarrow \hat{\mathcal{S}} \rightarrow R_{out}$ holds, with $\hat{\mathcal{S}}$ given in (2). It can be shown that $D_0(\hat{\mathcal{S}}(\tilde{E})) \leq \bar{D}_{FCFSMUX}$. In fact, equality holds here since the delay bound from [5] is the best possible. Thus, in some sense Theorem 4.1 of [5] is a special case of Theorem 2.

Furthermore, from Theorem 4.4 of [5], it is known that R_{out} has envelope E_{out}, where

$$E_{out}(x) = \max_{\Delta \geq 0,\, D \geq 0} \left[\min\{E(x + D),\ E(x + D + \Delta) + E^x(\Delta) - C(\Delta + D)\}\right] \ . \tag{12}$$

We can show that in fact Theorem 4 reduces to this result in this case. For brevity we do not include the details here, but we note that we assumed continuity of $E(x)$ and $E^x(x)$ for $x > 0$. We conjecture that this assumption is un-necessary, however.

Thus, we assert that Theorems 2 and Theorem 4 here are more general than Theorems 4.1 and Theorem 4.4 of [5].

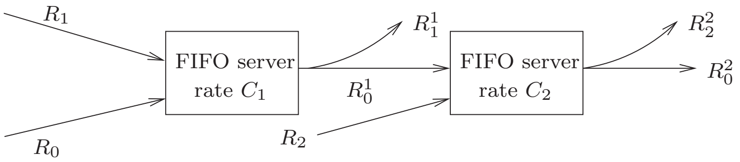

Figure 3: Two nodes in tandem

4.2 FIFO Multiplexers in Tandem

In this section we consider the system illustrated in Figure 3, where each arrival processes R_i has an envelope E_i of the form $E_i(t) = \sigma_i + \rho_i t$. We are interested in an upper bound for the total delay for flow 0, which is the sum of the delay through each node. We assume that the system is stable, that is $\rho_0 + \rho_1 \leq C_1$ and $\rho_0 + \rho_2 \leq C_2$, which ensures that the delay is bounded. We use the notation $\hat{\mathcal{S}}_1$ and $\hat{\mathcal{S}}_2$ to denote the corresponding minimum service mappings for flow 0 at the first and second multiplexer, respectively, as implied by (1) and (2).

It can be shown that

$$\hat{\mathcal{S}}_1(\tilde{E}_0(t)) = \begin{cases} \rho_0 t - \sigma_0 - \rho_0 \frac{\sigma_1}{C_1} & \text{if}\, t \leq \frac{\sigma_1}{C_1} \\ C_1 t - \sigma_1 - \sigma_0 & \text{if}\, \frac{\sigma_1}{C_1} \leq t \leq \frac{\sigma_1 + \sigma_0}{C_1} \\ 0 & \text{if}\, t \geq \frac{\sigma_1+\sigma_0}{C_1} \end{cases}$$

Moreover, it can be shown that if $C_2 - \rho_2 \geq C_1$ we have

$$\hat{\mathcal{S}}_2\left(\hat{\mathcal{S}}_1(\tilde{E}_0(t))\right) = \begin{cases} \rho_0\left(t - \frac{\sigma_2}{C_2} - \frac{\sigma_1}{C_1}\right) - \sigma_0 & \text{if}\, t \leq \frac{\sigma_1}{C_1} + \frac{\sigma_2}{C_2} \\ C_1\left(t - \frac{\sigma_2}{C_2} - \frac{\sigma_1}{C_1}\right) - \sigma_0 & \text{if}\, \frac{\sigma_1}{C_1} + \frac{\sigma_2}{C_2} \leq t \leq \frac{\sigma_1+\sigma_0}{C_1} + \frac{\sigma_2}{C_2} \\ 0 & \text{if}\, t \geq \frac{\sigma_1+\sigma_0}{C_1} + \frac{\sigma_2}{C_2} \end{cases}$$

whereas if $C_2 - \rho_2 \leq C_1$ we have

$$\hat{\mathcal{S}}_2\left(\hat{\mathcal{S}}_1(\tilde{E}_0(t))\right) = \begin{cases} \rho_0\left(t - \frac{\sigma_2}{C_2} - \frac{\sigma_1}{C_1}\right) - \sigma_0 & \text{if}\, t \leq \frac{\sigma_1}{C_1} + \frac{\sigma_2}{C_2} \\ \frac{C_2 C_1}{C_1+\rho_2}\left(t - \frac{\sigma_2}{C_2} - \frac{\sigma_1}{C_1}\right) - \sigma_0 & \text{if}\, \frac{\sigma_1}{C_1} + \frac{\sigma_2}{C_2} \leq t \leq \frac{\sigma_1}{C_1} + \frac{\sigma_2}{C_2} + \frac{\sigma_0(C_1+\rho_2)}{C_2 C_1} \\ 0 & \text{if}\, t \geq \frac{\sigma_1}{C_1} + \frac{\sigma_2}{C_2} + \frac{\sigma_0(C_1+\rho_2)}{C_2 C_1} \end{cases}$$

The upper bound on the end-to-end delay as given by Theorem 1 and Theorem 2 is: $D = D_0((\hat{\mathcal{S}}_1 \circ \hat{\mathcal{S}}_2)(\tilde{E}_0)) = D_0\left(\hat{\mathcal{S}}_2(\hat{\mathcal{S}}_1(\tilde{E}_0))\right)$. Carrying out this calculation, it can be verified that our upper bound D on end to end delay is given by

$$D = \begin{cases} \frac{\sigma_0+\sigma_1}{C_1} + \frac{\sigma_2}{C_2} & \text{if } C_2 - \rho_2 \geq C_1 \\ \frac{\sigma_1}{C_1} + \frac{\sigma_2}{C_2} + \frac{\sigma_0(C_1+\rho_2)}{C_2 C_1} = \frac{\sigma_1}{C_1} + \frac{\sigma_2}{C_2} + \frac{\sigma_0}{C_2} + \frac{\sigma_0 \rho_2}{C_2 C_1} & \text{if } C_2 - \rho_2 \leq C_1 \end{cases} \tag{13}$$

These bounds are indeed achievable. To see why this is true when $C_2 - \rho_2 \leq C_1$ consider the following arrival pattern: $R_0(t) = R_1(t) = 0$ for $t < 0$; $R_0(t) = \sigma_0$ and $R_1(t) = \sigma_1$ for $t \geq 0$. That is both flows at the first server have a burst at time 0. Suppose that flow 1 is served before flow 0, then suppose that $R_2(t) = 0$ for $t < \frac{\sigma_1}{C_1}$ and that $R_2(t) = \rho_2(t - \frac{\sigma_1}{C_1}) + \sigma_2$ otherwise. Under these assumptions the last bit of the burst from flow 0 will experience a total delay given by (13). If $C_2 - \rho_2 \leq C_1$ the arrival pattern for flow 0 and 1 is the same while in this case $R_2(t) = 0$ for $t < \frac{\sigma_1+\sigma_0}{C_1}$ and $R_2(t) = \sigma_2$ otherwise; again the last bit of the burst from flow 0 will experience a total delay given by (13).

References

[1] Rajeev Agrawal, Rene L. Cruz, Clayton Okino, and Rajendran Rajan. Performance bonds for flow control protocols. *IEEE/ACM Transactions on Networking (TON)*, 7(3):310–323, 1999.

[2] R. Braden, D. Clark, and S. Shenker. RFC 1633: Integrated services in the Internet architecture: an overview, June 1994.

[3] C.S. Chang. On deterministic traffic regulation and service guarantee: A systematic approach by filtering. *IEEE Transactions on Information Theory*, 44:1096–1107, 1998.

[4] C.S. Chang. *Performance Guarantees in Communcation Networks.* Springer Verlag, New York, 2000.

[5] R. L. Cruz. A calculus for network delay, part I: Network elements in isolation. *IEEE Transactions on Information Theory*, 37, 1:114–131, 1991.

[6] R. L. Cruz. A calculus for network delay, part II: Network analysis. *IEEE Transactions on Information Theory*, 37, 1:132–141, 1991.

[7] Rene L. Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE Journal on Selected Areas in Communications*, 13(6):1048–1056, 1995.

[8] Rene L. Cruz. SCED+: Efficient management of quality of service guarantees. In *IEEE Infocom '98, San Francisco*, March 1998.

[9] J. Y. Le Boudec. Application of network calculus to guaranteed service networks. *IEEE Transactions on Information Theory*, 44(3), May 1998.

[10] J. Y. Le Boudec and P. Thiran. *Network Calculus.* Springer Verlag, New York, June 2001.

[11] K. Nichols, S. Blake, F. Baker, and D. Black. RFC 2474: Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 headers, December 1998.

[12] Abhay K. Parekh and Robert G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, 1993.

[13] Abhay K. Parekh and Robert G. Gallagher. A generalized processor sharing approach to flow control in integrated services networks: the multiple node case. *IEEE/ACM Transactions on Networking (TON)*, 2(2):137–150, 1994.

[14] H. Sariowan. *A service curve approach to performance guarantees in integrated service networks.* PhD thesis, U.C. San Diego, 1996.

[15] H. Zhang. Service disciplines for guaranteed performance service in packet-switching networks. In *Proceedings of the IEEE, 83(10)*, October 1995.