1

A Self-Similar Traffic Model for Network-on-Chip Performance Analysis Using Network Calculus

Yue Qian
School of Computer Science, National University of Defense Technology
China

1. Introduction

Since around year 2000, Network-on-Chip (NoC) has been proposed as a global communication paradigm to interconnect tens or hundreds of cores on a single chip (Bjerregaard & Mahadevan, 2006). One key challenge for NoCs has been Quality of Service (QoS), which is concerned about performance guarantees or bounds. To achieve QoS, formal performance analysis is essential because it overcomes the uncertainty in results and lengthiness in time of simulation-based approaches (Lu, 2007).

Network calculus (NetCal) (Chang, 2000; Cruz, 1991; Le Boudec & Thiran, 2004) is a mathematical framework to derive worst-case bounds on maximum latency and backlog. The beauty of NetCal relies on two abstraction models, an *arrival curve* for traffic, and a *service curve* for network elements (router, relay node, interface, channel, server etc.). Arrival curves bound the accumulated amount of traffic. Service curves describe minimal service levels of network elements. With these two models, the delay and backlog buffer bounds can be calculated. NetCal has been extremely successful when applied to ATM and IP networks with both differentiated and integrated services to achieve predictable performance without over-dimensioning network architectures (Le Boudec & Thiran, 2004). Recently NetCal has also been applied to wireless LAN (Agharebparast & Leung, 2005), sensor networks (Schmitt & Roedig, 2005), and on-chip networks (Qian et al., 2010) etc.

Our intention is to use NetCal for communication performance analysis of self-similar traffic in on-chip networks. ATM, Ethernet and Internet traffic has shown *self-similar* characteristics (Park & Willinger, 2000). In on-chip networks, it turns out also to be true for many applications, particularly, multimedia traffic, as supported by (Scherrer et al., 2005; Soteriou et al., 2006; Varatkar & Marculescu, 2004). By analyzing on-chip traffic traces, they demonstrate that packets injected from routing nodes possess scale-invariant burstiness over time. However, existing self-similar traffic models (Mao & Panwar, 2006; Park & Willinger, 2000) are not directly subject to NetCal analysis. The reason is simply because they do not comply with the arrival curve model. Therefore the purposes of our work are triple-folded: (1) to find an arrival curve for self-similar traffic, if it exists; (2) otherwise, propose an arrival curve to envelop the self-similar traffic; (3) to perform analysis based on the proposed arrival model using the NetCal framework. Performing these tasks should keep the beauty of NetCal and still enable us to apply known NetCal analysis methods and results to analyze the performance and buffering cost of networks transporting self-similar traffic flows.

The remainder of the chapter is organized as follows. Section 2 summarizes related work and our contributions. In Section 3, we first introduce the property of self-similar traffic. Then we present the Fractional Brownian Motion (FBM) model (Norros, 1995), which is used to characterize the self-similarity of traffic, and how to estimate FBM parameters. In Section 4, we present our main findings in the form of theorems, proposing an *extended arrival curve* to constrain self-similar traffic. Afterwards, in Section 5, we present formulas to calculate delay and backlog bounds. Assuming the latency-rate server model (Stiliadis & Varma, 1998) for network elements, we give closed-form equations. Moreover, to give a complete picture of our method, we describe a performance analysis flow to show how to conduct performance analysis for self-similar traffic. Experiments and results are reported in Section 6. Finally we draw conclusions in Section 7.

2. Related work

Since being initially identified in Ethernet by Leland et al. (Leland et al., 1994), traffic self-similarity has far-reaching influence on traffic modeling and performance analysis. Explorations of the nature of self-similarity and applications of this complex phenomenon have been extensively studied and summarized (Park & Willinger, 2000). In the context of NoCs, researchers have found the evidence of self-similarity from on-chip communication traces. In (Varatkar & Marculescu, 2004), Varatkar et al. first introduced self-similarity as a fundamental property exhibited by the bursty traffic between on-chip modules in multimedia video applications. This work captured the traffic characteristics between pair-wise nodes rather than for the entire network. Later, Soteriou et al. (Soteriou et al., 2006) empirically studied a large set of traffic traces gathered from the execution of SPEC, MediaBench and bit-parallel benchmarks over the entire on-chip network with different architectures and showed the presence of self-similar phenomena in on-chip traffic flows.

Cruz (Cruz, 1991) has pioneered the network calculus, which is based on bounds of traffic flows. A useful family of bound functions for concise descriptions has the form $\alpha(t) = rt + b$, where r is the rate and b limits the burstiness of the flow. Based on Cruz's foundation, Chang (Chang, 2000) and Le Boudec (Le Boudec & Thiran, 2004) have further developed the network calculus theory and based it on min-plus algebra. The basic elements in this algebra are arrival curves as an abstraction of application traffic and service curves as an abstraction for components (network elements). A well-defined service curve is the so-called latency-rate function $\beta_{R,T}$, where R is the service rate and T the maximum response delay of the node (Stiliadis & Varma, 1998).

Stochastic network calculus (Ciucu et al., 2005; Jiang, 2006; Starobinski & Sidi, 2000; Yin et al., 2002) is the probabilistic version of the (deterministic) network calculus. It has recently been developed for stochastic service guarantee analysis. Stochastic network calculus combines the deterministic network calculus with statistical multiplexing. For this, several stochastic versions of arrival curve have been proposed by extending the concept of arrival curve to the stochastic case based on the traffic amount property or virtual backlog property. Among the existing stochastic arrival curves, Sum of Exponentials, Weibull Bounded Burstiness (WBB), Fractional Brownian Motion (FBM) and Multifractal Brownian Motion (MBM) envelope processes consider the self-similar traffic (Mao & Panwar, 2006). In contrast to the deterministic arrival curves, stochastic arrival curves envelop traffic tighter but have higher implementation complexity.

In (Norros, 1995), Norros introduced the FBM model to capture the long-range dependence within the self-similar traffic. This model inspires WBB envelope process and is the basis for the FBM and MBM envelope processes (Mao & Panwar, 2006). Since the stochastic properties of the FBM process retain well when the traffic is multiplexed, randomly split, or goes through a buffering system, the FBM model serves well for the objective of concatenating single-hop analysis into an end-to-end analysis (Cheng et al., 2007).

We link self-similar traffic to deterministic network calculus. We develop an extended linear arrival model as its arrival curve, and then apply NetCal analysis on it. Our arrival curve is also constructed based on the FBM process. In contrast to other stochastic arrival curves, it is coupled with deterministic network calculus. Also, it is an extension of the traditional linear expression, thus easy to use and understand and simple in implementation. We summarize our contributions as follows:

- We prove that self-similar traffic cannot be enveloped by any deterministic arrival curve.
- We extend the linear arrival curve $\alpha_{r,b}(t) = rt + b$ with an excess probability ε as ε-$\alpha_{r,b}(t) = rt + b(\varepsilon)$, where ε reflects the probability of traffic burstiness surpassing its arrival curve. We prove that self-similar traffic can be characterized by the extended linear arrival curve ε-$\alpha_{r,b}$.
- Based on the extended self-similar traffic model, we derive delay and backlog bounds for self-similar traffic served by one or a series of concatenated network elements. Furthermore, we give closed-form equations to compute the bounds assuming the network elements are modeled by the latency-rate server (Stiliadis & Varma, 1998).
- We present a performance analysis flow starting from self-similar traffic and ending with results of delay and backlog bounds.

3. Self-similarity and FBM

In this section we give a definition of self-similar traffic (Park & Willinger, 2000), describe the FBM model (Fonseca et al., 2000; Norros, 1995), and introduce the estimation of FBM parameters, $(\bar{a}, \sigma, H)$ (Norros, 1995; Park & Willinger, 2000).

3.1 Self-similarity

Let $X(t)$ denote the traffic volume arriving in the tth time unit. Let $A(t)$ be the cumulative process indicating the total traffic volume from time 0 up to time t. $X(t)$ is also termed as the increment process of $A(t)$ as $X(t) = A(t) - A(t-1)$.

Given a stationary time series $X = (X(t),\ t = 1, 2, 3, ...)$, we define the m-aggregated series $X^{(m)} = (X^{(m)}(k),\ k = 1, 2, 3, ...)$ by summing the original series X over non-overlapping blocks of size m. The time series process X is called *asymptotically second-order self-similar* (as-s), if the autocorrelation function of $X^{(m)}$ and X follows

$$r^{(m)}(k) \sim r(k),\ \text{as } m \to \infty,\ k \to \infty. \tag{1}$$

That is, at all scales the aggregated autocorrelation structures agree asymptotically to the autocorrelation structure of the entire series X.

The crucial feature of self-similar processes is that they exhibit *long-range dependence* (LRD). These LRD processes have an autocorrelation function $r(k)$ that decays with time lag k, i.e., $r(k) \sim k^{-\gamma}$ as $k \to \infty$, where $0 < \gamma < 1$. The *Hurst parameter* H is commonly used to measure the degree of LRD, and is related to the parameter γ by $H = 1 - \gamma/2$. In fact, with $1/2 < H < 1$, as-s and LRD imply each other, and self-similarity and LRD are often used interchangeably in practice.

3.2 FBM and its envelope process

Many different models are widely used to represent self-similarity. We use Fractional Brownian Motion (FBM) (Norros, 1995) to model the cumulative input traffic $A(t)$. The FBM input $\{A(t) : t \geq 0\}$ can be represented by

$$A(t) = \bar{a}t + \sigma Z(t), \tag{2}$$

where the mean arrival rate $E\{A(t)/t\} = \bar{a}$, and σ^2 is the variance of traffic in a time unit, and $\{Z(t) : t \geq 0\}$ is the standard (normalized) FBM process with Hurst parameter $H \in [1/2, 1)$.

The basic known property of FBM model is its marginal distribution (Norros, 1995), which allows computing an envelope process. For an FBM process $A(t)$ with mean $\bar{a}$ and variance σ^2, the envelope process $\hat{A}(t)$ can be defined as

$$\hat{A}(t) \stackrel{def}{=} \bar{a}t + k\sqrt{\sigma^2 t^{2H}} = \bar{a}t + k\sigma t^H, \tag{3}$$

where the parameter k determines the probability that $A(t)$ will exceed $\hat{A}(t)$ at time t as follows:

$$P\left(A(t) > \hat{A}(t)\right) = P\left(\frac{A(t) - \hat{a}t}{\sigma t^H} > k\right) = \varepsilon = \Phi(k), \tag{4}$$

where $\Phi(y)$ is the residual distribution function of the standard Gaussian distribution, using the approximation $\Phi(y) = \exp(-y^2/2)$, k is given by $k = \sqrt{-2\ln\varepsilon}$.

The FBM envelope process is advantageous: (1) It is parsimonious, i.e., only three parameters $(\bar{a}, \sigma, H)$ are required to completely characterize a self-similar source; (2) The input parameters $(\bar{a}, \sigma, H)$ can be estimated in real-time from the incoming traffic samples with minimal computational complexity (Fonseca et al., 2000).

3.3 Estimation of FBM parameters $(\bar{a}, \sigma, H)$

The FBM parameters $(\bar{a}, \sigma, H)$ can be estimated from a sample of traffic traces. To estimate $\bar{a}$ and σ, we first get the traffic cumulative process $A(t)$ from the sample. The mean arrival rate is derived as $\bar{a} = E\{A(t)/t\}$ and the variance of traffic in a time unit is given as $\sigma = \frac{\sqrt{\mathrm{Var}\{A(t)\}}}{t^H}$ (Norros, 1995).

To estimate Hurst parameter H, there are a number of methods: analysis of R/S (Range/Scale, rescaled adjusted range) statistic, analysis of the variance-time plot, the Whittle estimation and analysis based on wavelet function (Park & Willinger, 2000). We adopt the R/S method summarized as follows.

Given a sample of n observations in the time series (X_k, $k = 1, 2, ..., n$), the R/S statistic is denoted as $M\left[\frac{R(n)}{S(n)}\right] \sim cn^H$ as $n \to \infty$ and c is a positive constant. Taking the logarithm of the two parts gives $\log\left\{M\left[\frac{R(n)}{S(n)}\right]\right\} \sim H\log(n) + \log(c)$ as $n \to \infty$. Thus the H parameter can be estimated by placing the graph of the $\log\{M[R(n)/S(n)]\}$ on $\log(n)$ and using the obtained points to select a straight line with slope H based on the least-squares method (Park & Willinger, 2000).

4. Self-similar traffic model ε-$\alpha_{r,b}$

In Theorem 1, we prove that a self-similar traffic flow cannot be bounded by any deterministic function.

Theorem 1. *For a self-similar traffic flow, whose FBM envelope process is $\hat{A}(t) = \bar{a}t + k\sigma t^H$, there does not exist any wide-sense increasing deterministic function $\alpha(t)$ ($t > 0$) to envelope the flow.*

Proof. Using reduction ad absurdum, we assume there exists such $\alpha(t)$ for all $t > 0$ that $\alpha(t) \geq A(t)$, hence

$$P\{A(t) > \alpha(t)\} = 0, \tag{5}$$

where $A(t)$ denotes the cumulative function of the self-similar traffic flow. For any specified time t, the volume of $\alpha(t)$ is deterministic.

Since the self-similar flow is modeled by FBM, with the concept of the FBM envelope process, we can get $k = \sqrt{-2\ln\varepsilon}$ when $\varepsilon \to 0$, $k \to \infty$.

As $\bar{a}$ and σ are all positive and $t > 0$, there exists some $\varepsilon^* > 0$ which makes $k > \frac{\alpha(t)}{\sigma t^H}$, i.e., $\bar{a}t + k\sigma t^H > \alpha(t)$, at the same time, $P\{A(t) > \bar{a}t + k\sigma t^H\} = \varepsilon^*$.

Therefore

$$P\{A(t) > \alpha(t)\} > P\{A(t) > \bar{a}t + k\sigma t^H\} = \varepsilon^* > 0, \tag{6}$$

which conflicts Eq. (5). This means the condition can not be true, i.e., $\alpha(t)$ does not exist. □

Note that, in Theorem 1, $\alpha(t)$ covers any deterministic arrival curve, linear and nonlinear. However, in order to use NetCal theory for performance analysis of self-similar traffic, we develop in Theorem 2 an extended arrival curve for self-similar traffic, which is an ε-enhanced linear arrival curve.

Theorem 2. *For a self-similar traffic flow, whose FBM envelope process is $\hat{A}(t) = \bar{a}t + k\sigma t^H$, there exists a deterministic linear arrival curve ε-$\alpha_{r,b}(t) = rt + b(\varepsilon)$, having values exceeded by the traffic flow for any t with the upper excess probability $\varepsilon = \Phi(k)$, where $r > \bar{a}$, $b(\varepsilon) = (r - \bar{a})^{\frac{H}{H-1}}(\Phi^{-1}(\varepsilon)\sigma)^{\frac{1}{1-H}}H^{\frac{H}{1-H}}(1 - H)$.*

Proof. Since the traffic flow exceeds the arrival curve ε-$\alpha_{r,b}$ with the upper excess probability ε ($0 < \varepsilon \leq 1$), we have

$$P\{A(t) > \varepsilon\text{-}\alpha_{r,b}(t)\} \leq \varepsilon = P\{A(t) > \hat{A}(t)\}, \tag{7}$$

hence

$$\varepsilon\text{-}\alpha_{r,b}(t) = rt + b(\varepsilon) \geq \hat{A}(t) = \bar{a}t + k\sigma t^H. \tag{8}$$

By Eq. (8) for all t, we get

$$(r - \bar{a})t - k\sigma t^H + b(\varepsilon) \geq 0. \tag{9}$$

Since the Hurst parameter $1/2 < H < 1$, Eq. (9) is satisfied for the stable case only $r - \bar{a} > 0$, therefore $r > \bar{a}$.

To proceed further it is sufficient to note that Eq. (9) has to be met for the worst case and therefore, the minimum value of the left side of Eq. (9) in turn must be equal to zero (as of a weak inequality).

Let $f(t) = (r - \bar{a})t - k\sigma t^H + b(\varepsilon)$, in order to compute the minimum value of $f_{\min}$, it is necessary to find t^* such that $\dfrac{\mathrm{d}f(t)}{\mathrm{d}t} = 0$. Hence we have $(r - \bar{a}) - Hk\sigma t^{H-1} = 0$, t^* is given by $t^* = \left[\dfrac{k\sigma H}{(r - \bar{a})}\right]^{\frac{1}{1-H}}$.

Insert t^* into $f(t) = 0$, we get

$$\begin{aligned} b(\varepsilon) &= (\bar{a} - r)\left(\frac{k\sigma H}{r - \bar{a}}\right)^{\frac{1}{1-H}} + k\sigma\left(\frac{k\sigma H}{r - \bar{a}}\right)^{\frac{H}{1-H}} \\ &= (r - \bar{a})^{\frac{H}{H-1}}(\Phi^{-1}(\varepsilon)\sigma)^{\frac{1}{1-H}} H^{\frac{H}{1-H}}(1 - H). \end{aligned} \tag{10}$$

□

We can see that $b(\varepsilon)$ is a function of r ($r > \bar{a}$) and FBM parameters of $(\bar{a}, \sigma, H)$. Certainly, how closely the extended arrival curve constrains the traffic flow is sensitive to the excess probability ε, which is a measure of majorizing precision.

5. Performance analysis

Using the proposed arrival curve, we derive performance and backlog bounds based on the concepts of arrival and service curves (Le Boudec & Thiran, 2004).

5.1 General bounds

When a self-similar traffic flow with arrival curve $\varepsilon\text{-}\alpha_{r,b}$ is processed by a network element with service curve β, the maximum delay for the flow is bounded by:

$$D(\varepsilon\text{-}\alpha_{r,b}, \beta) = \sup_{t \geq 0}\{\inf\{\tau \geq 0 : \varepsilon\text{-}\alpha_{r,b}(t) \leq \beta(t + \tau)\}\}. \tag{11}$$

When a traffic flow is processed by a sequence of network elements, we could simply add the different maximum delays of each individual component together to obtain an end-to-end delay guarantee. However, in this case we can exploit the phenomenon known as Pay Bursts Only Once (Le Boudec & Thiran, 2004), and the end-to-end delay guarantee can be tightened by:

$$D(\varepsilon\text{-}\alpha_{r,b}, \beta_1 \otimes \beta_2 \otimes \ldots \otimes \beta_n). \tag{12}$$

The maximum buffer size that is required to buffer the traffic flow is bounded by:

$$B(\varepsilon\text{-}\alpha_{r,b}, \beta) = \sup_{t \geq 0}\{\varepsilon\text{-}\alpha_{r,b}(t) - \beta(t)\}. \tag{13}$$

And when the traffic flow traverses several consecutive elements, the total required buffer space can even be tightened by:

$$B(\varepsilon\text{-}\alpha_{r,b}, \beta_1 \otimes \beta_2 \otimes \ldots \otimes \beta_n). \tag{14}$$

Note that, strictly speaking, the delay and backlog "bounds" should be interpreted as "estimates" for maximum delay and backlog. Since the traffic is not entirely constrained by the arrival curve in our model due to ε, it is possible in theory that the calculated bounds may be exceeded, even though appearing only in extreme cases. However, to follow the terminology used in network calculus based performance analysis, we also use "bounds" for the estimated maximum delay and backlog in the chapter.

5.2 Bounds for latency-rate servers

In addition to the general performance bounds, we give equations to compute the bounds assuming the *latency-rate server* model for network elements (Stiliadis & Varma, 1998).

Consider a self-similar traffic flow with arrival model $\varepsilon\text{-}\alpha_{r,b}(t) = rt + b(\varepsilon)$ traversing a series of network elements, each element i ($i = 1, 2, 3, ..., n$) guarantees a latency-rate service curve $\beta_{R_i,T_i} = R_i(t - T_i)^+$, where R_i is the service rate and T_i delay to serve the flow. Notation $x^+ = x$, if $x \geq 0$; $x^+ = 0$, otherwise.

Let $R_{\min} = \bigwedge_{i=1}^{n} R_i$ and $T_{tol} = \sum_{i=1}^{n} T_i$. If $r \leq R_{\min}$, then the delay bound is

$$\begin{aligned} D(\varepsilon\text{-}\alpha_{r,b}, \beta_{R_1,T_1} \otimes \beta_{R_2,T_2} \otimes ... \otimes \beta_{R_n,T_n}) &= \frac{b(\varepsilon)}{R_{\min}} + T_{tol} \\ &= \frac{(r-\bar{a})^{\frac{H}{H-1}} (\Phi^{-1}(\varepsilon)\sigma)^{\frac{1}{1-H}} H^{\frac{H}{1-H}} (1-H)}{\bigwedge_{i=1}^{n} R_i} + \sum_{i=1}^{n} T_i, \end{aligned} \tag{15}$$

and the buffer bound is

$$\begin{aligned} B(\varepsilon\text{-}\alpha_{r,b}, \beta_{R_1,T_1} \otimes \beta_{R_2,T_2} \otimes ... \otimes \beta_{R_n,T_n}) &= b(\varepsilon) + rT_{tol} \\ &= (r-\bar{a})^{\frac{H}{H-1}} (\Phi^{-1}(\varepsilon)\sigma)^{\frac{1}{1-H}} H^{\frac{H}{1-H}} (1-H) + r\sum_{i=1}^{n} T_i. \end{aligned} \tag{16}$$

If $r > R_{\min}$, the bounds are infinite.

We can see when ε ($0 < \varepsilon \leq 1$) is approaching to 1, the backlog and delay bounds are deceasing. In particular, when ε equals 1, the value of $b(\varepsilon)$ will be zero and the delay and buffer bounds will equal to $\sum_{i=1}^{n} T_i$ and $r\sum_{i=1}^{n} T_i$, respectively. The reason is that, as ε increases,

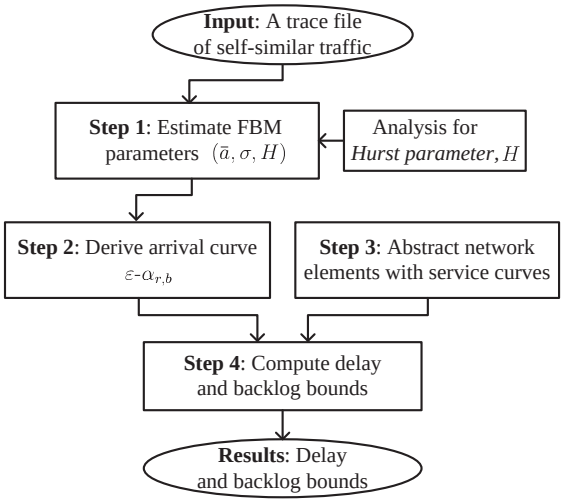


Fig. 1. Performance Analysis Flow Using Network Calculus on Self-Similar Traffic.

more bursty traffic exceeds the arrival curve. This is similar to the effect of lowering the traffic arrival curve. Thus the computed delay and backlog bounds become smaller.

5.3 Performance analysis flow

We illustrate the analysis flow in Figure 1. The input is a trace of self-similar traffic and output is delay and backlog bound results. The procedure contains four steps:

- Step 1: Estimate FBM parameters $(\bar{a}, \sigma, H)$ (Section 3.3). This step checks for self-similarity in the trace and performs, for example, the R/S analysis, to derive Hurst parameter H. With this step, we obtain its cumulative process.
- Step 2: Find its FBM envelope process, and further derive its ε-enhanced arrival model (Section 4).
- Step 3: Model network elements with service curves.
- Step 4: Compute delay and backlog bounds for its traversal through a single node or concatenated nodes. If the service models follow the latency-rate model, we can use the closed-form equations in Section 5.2 to compute the bounds.

6. Experiments and results

We devised experiments to (1) validate the proposed self-similar model; (2) show the correctness and tightness of calculated bounds via comparing them with simulated results. With the experiments, we also exemplify the performance analysis flow.

6.1 The simulation platform

We use a simulation platform in an open source simulation environment SoCLib (*SoCLib Simulation Environment*, n.d.) to collect application traces and to simulate their delay and backlog in on-chip networks.

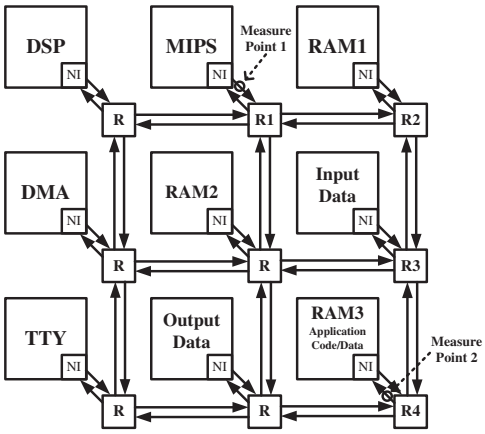


Fig. 2. The Simulation Platform.

As shown in Figure 2, the platform contains a MIPS R3000 processor, on-chip memories, a display component (TTY), and other components such as DSP and DMA. These components are interconnected with a 3×3 mesh network. The network performs wormhole flow control and uses XY routing. Routers are uniform, taking 5 cycles to deliver head flits and one cycle for other flits. Application code and data are stored in RAM3. The Network Interfaces (NIs) encapsulate transactions into flits and de-encapsulate flits into transactions.

We run four embedded multimedia programs on the MIPS: an MP3 audio decoder, an MPEG2 video decoder, a JPEG and a JPEG2000 decoder, respectively. The MP3 processes a 4KB audio stream, MPEG2 a 176×176 video frame, JPEG and JPEG2000 a 256×256 image. We set up two measurement points to observe the transactions between MIPS and RAM3 in the platform, as indicated in Figure 2. While application code running on the processor, at Point 1 we record the sequence number and timing of flits generated by MIPS in a trace file, and at Point 2 we observe the end-to-end delay experienced by each flit after traversing four routers, {R1, R2, R3, R4}, and the system backlog.

We have performed analysis and simulation for all the four application traces. For concise presentation, we only detail the analysis and simulation results of the MP3 application in Section 6.2 and Section 6.4, respectively. Section 6.3 discusses the derivation of the extended arrival curves for the MP3 application and the selection of parameters ε and r. Nevertheless, we report both analysis and simulation results on delay and backlog for all the applications in Section 6.5. For all results, the unit for delay is *cycle*, for backlog is $flit$. While examining traffic's self-similarity, we choose 100 cycles as the time window.

6.2 Analysis for MP3 application

The analysis of the MP3 application follows the four steps described in Section 5.3.

Step 1. The entire trace of MP3 application contains 1,697,249 flits in total and lasts for 46,696 hundreds of cycles as drawn in Figure 3. For such 100-cycle aggregated data series, we use the R/S analysis method to derive its Hurst parameter as illustrated in Figure 4. It turns out that H equals 0.86. This means the MP3 traffic exhibits good self-similarity. The FBM

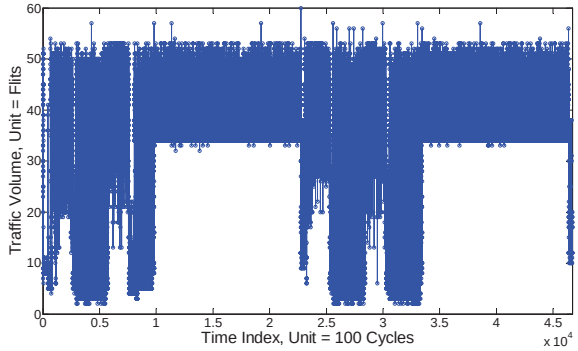


Fig. 3. Aggregated throughput trace obtained from the execution of MP3 application.

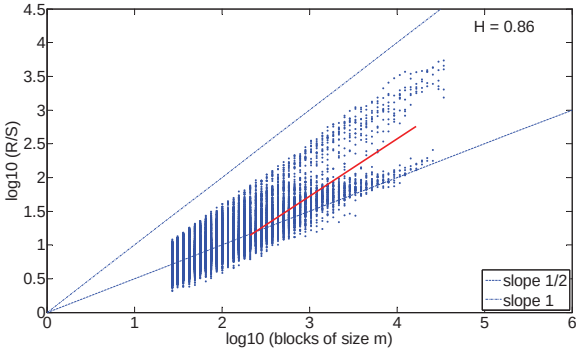


Fig. 4. Hurst Parameter Estimation via the R/S Method.

parameters of $\bar{a}$ and σ are also derived using the formulas presented in Section 3.3. We get the mean rate $\bar{a} = 36.35$ flits/100 cycles, and the variance in time unit of 100 cycles $\sigma = 0.33$.

Step 2. Assume the excess probability ε = 1E-4 (1×10^{-4}), with derived $(\bar{a}, \sigma, H) = (36.35, 0.33, 0.86)$, we have the FBM envelope process $\hat{A}(t) = 36.35t + 1.417t^{0.86}$. Now we compute its extended arrival curve of ε-$\alpha_{r,b}$. Let $r = 37$ flits/100 cycles $> \bar{a}$, then with Eq. (10), we get $b(\varepsilon) = 10$ flits, thus ε-$\alpha_{r,b}(t) = rt + b(\varepsilon) = 37t + 10$.

Together with the MP3 cumulative process, the two curves of ε-$\alpha_{r,b}(t)$ and $\hat{A}(t)$ are plotted in Figure 5. As we can see, the derived model ε-$\alpha_{r,b}(t)$ tightly bounds the cumulative process of the self-similar traffic. This validates the correctness of our proposed self-similar arrival model.

Step 3. The routers are modeled as latency-rate servers with the same service curve of $\beta(t) = 100(t - 0.05)^+$, which represents that the routers delay head flits for 5 cycles and forward 100 flits per 100 cycles.

Step 4. Flits generated by MIPS passing through a tandem of routers {R1, R2, R3, R4} before arriving at RAM3. Using Eq. (15) and (16), in Section 5.2, the delay and backlog bounds can be calculated as 30 cycles and 17.4 flits, respectively.

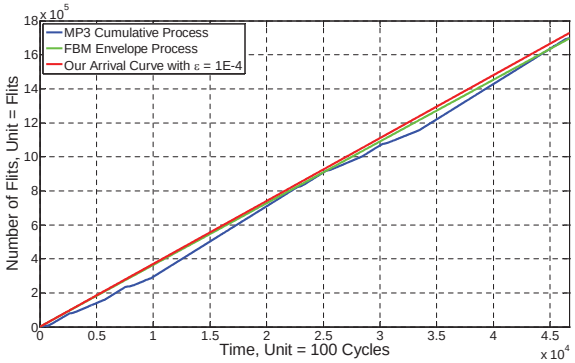


Fig. 5. Cumulative Process, FBM Envelope Process and ε-$\alpha_{r,b}$ Curve of Self-similar Traffic.

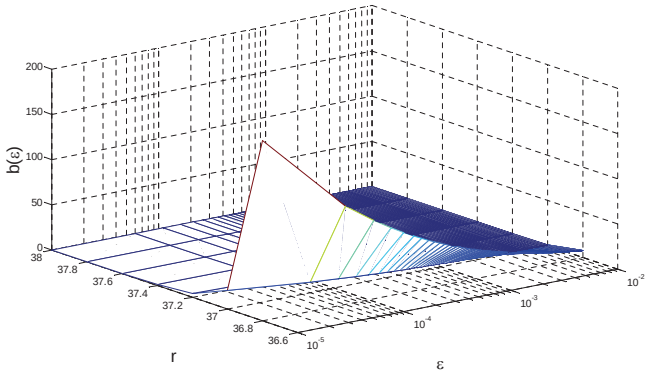


Fig. 6. $b(\varepsilon)$ with ε and r.

6.3 Discussions on extended arrival curves

6.3.1 Derivation of the extended arrival curves

For the MP3 application, we have obtained $(\bar{a}, \sigma, H) = (36.35, 0.33, 0.86)$. Using Equation 10, we get

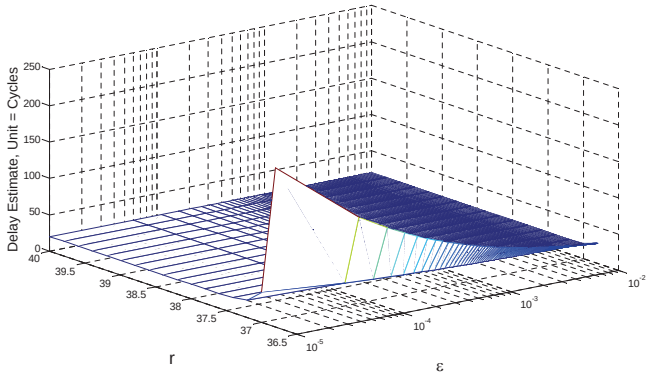
$$b(\varepsilon) = 0.0554 \cdot (r - 36.35)^{-6.1429} \cdot (0.33 \cdot \sqrt{-2 \ln \varepsilon})^{7.1429}. \quad (17)$$

This means that $b(\varepsilon)$ decreases as r or/and ε increases. The relation among b, r and ε is shown in the 3D Figure 6. With a small increase of r from 36.6 to 38, b is approaching 0. With an increase of ε, b is also decreasing and approaching to 0, but with a relatively less acceleration.

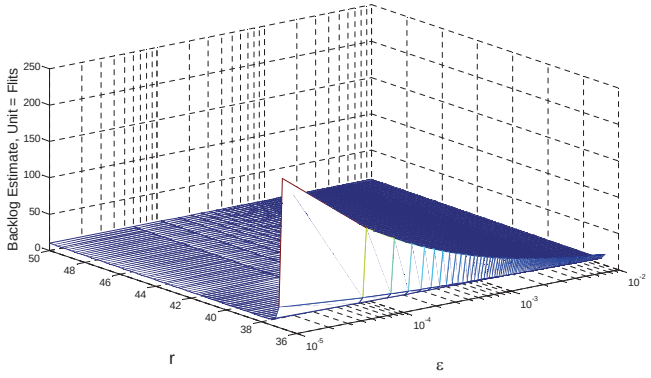
We also give the delay and backlog estimates as follows:

Delay Estimates:

$$D = 0.0554 \cdot (r - 36.35)^{-6.1429} \cdot (0.33 \cdot \sqrt{-2 \ln \varepsilon})^{7.1429} + 20. \quad (18)$$



(a) Delay Estimates with ε and r



(b) Backlog Estimates with ε and r

Fig. 7. Delay and Backlog Estimates with ε, r.

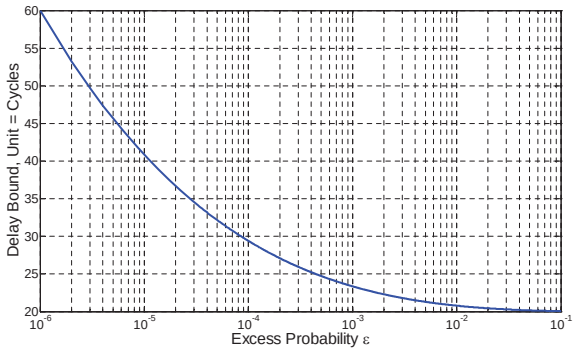
Backlog Estimates:

$$B = 0.0554 \cdot (r - 36.35)^{-6.1429} \cdot (0.33 \cdot \sqrt{-2 \ln \varepsilon})^{7.1429} + 0.2 \cdot r. \tag{19}$$

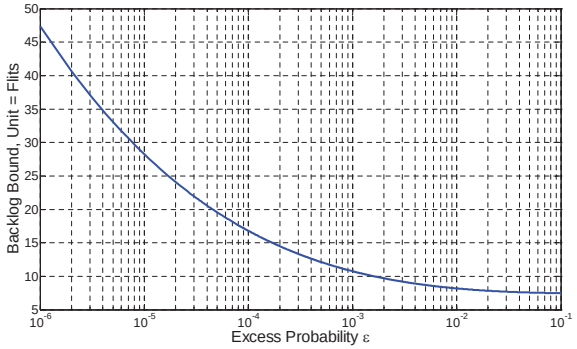
From the formulas, we can see that D/B decreases as r or/and ε increases, in a similar way as $b(\varepsilon)$. We draw two 3D figures for the delay and backlog estimates in Figure 7. We can see that the three figures are similar in shape.

6.3.2 Selection of ε and r

As can be observed from Figure 6 and 7, the burstiness b, delay and backlog estimates (D and B) are very sensitive to the value of $r > 36.35$. Staring from $r = 36.5$, a small increase of r sharply reduces b, D and B. We choose $r = 37$, since, from this point, the curves do not go down quickly. With this value, we plot a 2D figure to show how the delay and backlog estimates vary with ε in Figure 8.



(a) Delay Bounds with Excess Probability ε

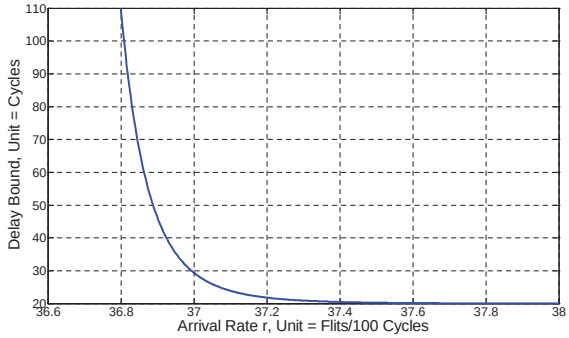


(b) Backlog Bounds with Excess Probability ε

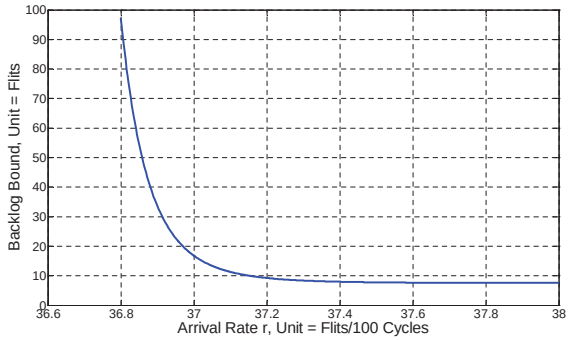
Fig. 8. Delay and Backlog Bounds Affected by Excess Probability ε when $r = 37$ flits/100 cycles.

Figure 8 clearly shows that, as ε increases from 1E-6 to 1E-1, the delay and backlog are both decreasing and the decrease is sharp until ε goes beyond 1E-4. From then on, the decrement of ε affects the bounds lightly. For smaller ε, the arrival curve allows less flits excess, and the bounds are certainly calculated larger. "ε = 1E-4 (1×10^{-4})" means that the tolerance of exceeding the arrival curve is one out of 10,000 flits. Note that the excess probability ε may come from application constraints. In such cases, ε is pre-determined and we only need to consider the relation between r and b.

With ε = 1E-4, we can look closer on how the selection of rate r influences the delay and backlog estimates, as shown in Figure 9. While varying r from 36.8 to 38, both the delay and backlog estimates decrease and the decrease is sharp until r exceeds 37. From then on, the increase of r affects the bounds lightly. For smaller r, the burstiness b is greater so as to guarantee that the ε-$\alpha_{r,b}$ envelopes the traffic for a certain excess probability, and the bounds are consequently calculated larger. Since $r = 37$ is the turning point, we have chosen $r = 37$ for the MP3 application.



(a) Delay Bounds with Arrival Rate r



(b) Backlog Bounds with Arrival Rate r

Fig. 9. Delay and Backlog Bounds Affected by Arrival Rate r when $\varepsilon = 1E\text{-}4$.

6.4 Simulation results of MP3 application

We present detailed simulation results for the MP3 application.

Figure 10(a) plots the flit delay for a sequence of 1E+4 (1×10^4) flits. The calculated delay bound (30 cycles) is plotted as a straight line. We can see that there is no point above the line. Similarly, in Figure 10(b), for the sequence of 1E+4 flits, we plot the backlog value at each observing time point when a flit arrives at RAM3 and the calculated backlog bound (17.4 flits) as a straight line. We can see that there are some points above the line, indicating there exist some points beyond the bound caused by the burstiness of self-similar traffic. This in fact validates one finding in this chapter: no deterministic arrival curves can fully bound self-similar traffic.

Figures 11(a) and 11(b) show the delay and backlog histogram, respectively, for the entire trace. We find the maximum delay is 24 cycles and there are no flits experiencing larger delay than the bound of 30 cycles, so the excess ratio equals zero. For the backlog, the observed maximum backlog is 20 flits. There are 6 points in total exceeding the bound of 17.4 flits. The real exceeding ratio equals $6/1697249 = 3.53E\text{-}6$, which is far smaller than the assumed

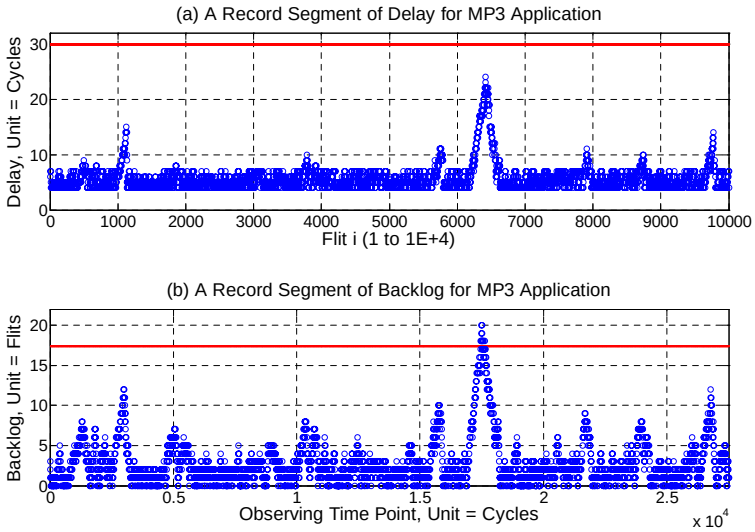


Fig. 10. Record Segment of Delay and Backlog for MP3 Application.

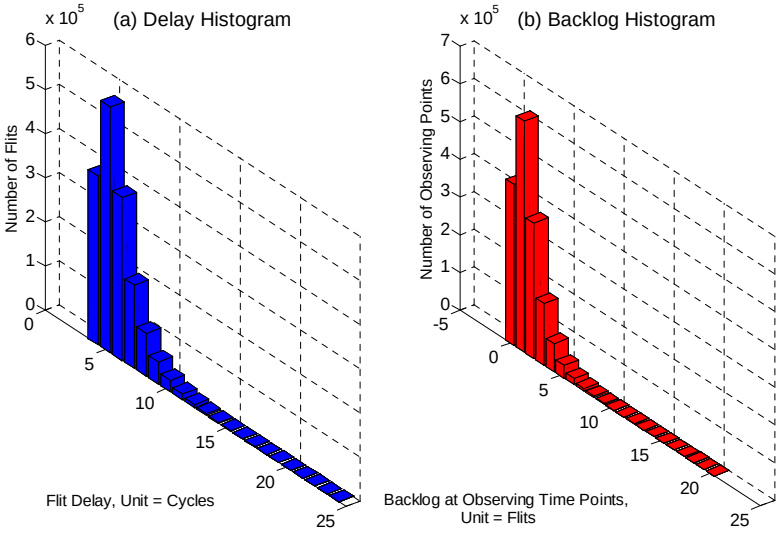


Fig. 11. Simulated Delay and Backlog Histograms for MP3 Application.

excess probability ε = 1E-4. This validates that our arrival curve with a predictive upper excess probability can well bound the self-similar traffic.

6.5 Summary of results for all applications

We summarize all calculated bounds and simulated results for the four applications, MP3, MPEG2, JPEG and JPEG2000 in Table 1, where we also list their FBM parameters and extended

Application	MP3	MPEG2	JPEG	JPEG2000
$\hat{a}$	36.35	25.06	38.32	34.51
σ	0.33	0.70	0.62	0.42
H	0.86	0.68	0.76	0.89
ε-$\alpha_{r,b}$ (ε=1E-4)	37t+10	26t+5	39t+6	35t+12
D	30	24.77	26.23	32.49
D_s	24	20	22	29
ϵ_D	0	0	0	0
B	17.4	9.98	14.09	19.59
B_s	20	13	17	24
ϵ_B	3.53E-6	4.47E-6	1.04E-6	2.61E-6

Table 1. Calculated and Simulated Results for MP3, MPEG2, JPEG and JPEG2000

arrival curves. We denote calculated delay bound and maximum simulated delay as D and D_s, respectively, and calculated backlog bound and maximum simulated backlog as B and B_s, respectively. The ϵ_D and ϵ_B represent the calculated exceeding ratio of the points beyond the delay and backlog bound, respectively. From this table, we can see that all the calculated delay bounds well constrain the simulated delay, i.e., $\epsilon_D = 0$. The calculated backlog bounds fail to constrain the maximum observed backlog in simulations. This results in $\epsilon_B > 0$, but we can observe $\epsilon_B << \varepsilon$. This means the proposed arrival models are good.

7. Conclusion

Performance analysis techniques must properly characterize traffic flows. In this chapter, we have presented a traffic arrival model for self-similar traffic, which is a very influential category of traffic observed in various networks. This model complies with the linear arrival model, and enhances it with an additional parameter, excess probability ε, to capture the probability of bursty traffic surpassing the linear arrival envelope. We develop such a model because of two reasons. One is that, as we have proved in the chapter, self-similar traffic cannot be bounded by any deterministic function. The other is that we hope to keep the elegance of the traffic abstraction in network calculus. With such an ε-enhanced arrival curve, we have shown how to apply network calculus theory for performance analysis of self-similar traffic flows. Assuming the latency-rate server model, we give closed-form equations for computing delay and backlog bounds for self-similar traffic traversing a tandem of network elements. We have also devised experiments to exemplify the performance analysis flow. Our simulations with real on-chip multimedia application traces have validated our model and results.

We have aimed our performance analysis of self-similar traffic for on-chip networks. However, the arrival-curve-compliant self-similar traffic model and its associated performance analysis method and formulas are equally applicable to off-chip networks, since we do not make any NoC-specific assumptions. Nevertheless, we believe our approach is most beneficial to the design of NoCs since NoC is a closed system focusing on specific application domains whereas traffic can be closely inspected, properly profiled and characterized.

8. References

Agharebparast, F. & Leung, V. C. M. (2005). Modeling wireless link layer by network for efficient evaluations of multimedia QoS, *Proceedings of IEEE International Conference on Communications*, Vol. 2, pp. 1256–1260.

Bjerregaard, T. & Mahadevan, S. (2006). A survey of research and practices of network-on-chip, *ACM Computing Survey* Vol. 33(No. 1): 1–51.

Chang, C.-S. (2000). *Performance Guarantees in Communication Networks*, Springer-Verlag.

Cheng, Y., Zhuang, W. & Ling, X. (2007). Towards an FBM model based network calculus framework with service differentiation, *Mobile Networks and Applications* Vol. 12(No. 5): 335–346.

Ciucu, F., Burchard, A. & Liebeherr, J. (2005). A network service curve approach for the stochastic analysis of networks, *Proceedings of the 2005 ACM SIGMETRICS*, pp. 279–290.

Cruz, R. L. (1991). A calculus for network delay, part I: Network elements in isolation and part II: Network analysis, *IEEE Transactions on Information Theory* Vol. 37(No. 1): 114–141.

Fonseca, N., Mayor, G. & Neto, C. (2000). On the equivalent bandwidth of self-similar sources, *ACM Transactions on Modeling and Computer Simulation* Vol. 10(No. 2): 104–124.

Jiang, Y. (2006). A basic stochastic network calculus, *Proceedings of the 2006 ACM SIGCOMM*.

Le Boudec, J.-Y. & Thiran, P. (2004). *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*, Number 2050 in LNCS, Springer-Verlag.

Leland, W. E., Taqqu, M. S., Willinger, W. & Wilson, D. V. (1994). On the self-similar nature of ethernet traffic (extended edition), *IEEE/ACM Transactions on Networking* Vol. 2(No. 1): 1–15.

Lu, Z. (2007). *Design and analysis of on-chip communication for network-on-chip platforms*, Ph.D. thesis, Royal Institute of Technology.

Mao, S. & Panwar, S. S. (2006). A survey of envelope processes and their applications in quality of service provisioning, *IEEE Communications Surveys and Tutorials* Vol. 8(No. 3): 2–20.

Norros, I. (1995). On the use of fractal brownian motion in the theory of connectionless networks, *IEEE Journal on Selected Areas in Communications* Vol. 13(No. 6): 953–962.

Park, K. & Willinger, W. (2000). *Self-similar Network Traffic and Performance Evaluation*, John Wiley and Sons.

Qian, Y., Lu, Z. & Dou, W. (2010). Analysis of worst-case delay bounds for on-chip packet-switching networks, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* Vol. 29(No. 5).

Scherrer, A., Fraboulet, A. & Risset, T. (2005). Analysis and synthesis of cycle-accurate on-chip traffic with long-range dependence, *Technical report 2005-53, LIP, ENS-Lyon* .

Schmitt, J. & Roedig, U. (2005). Sensor network calculus - a framework for worst case analysis, *Proceedings of the International Conference on Distributed Computing in Sensor Systems*, pp. 141–154.

SoCLib Simulation Environment (n.d.). On-line, available at http://www.soclib.fr/.

Soteriou, V., Wang, H. & Peh, L. (2006). A statistical traffic model for on-chip interconnection networks, *Proceedings of IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS'06)*.

Starobinski, D. & Sidi, M. (2000). Stochastically bounded burstiness for communication networks, *IEEE Transactions on Information Theory* Vol. 46(No. 1): 206–212.

Stiliadis, D. & Varma, A. (1998). Latency-rate servers: A general model for analysis of traffic scheduling algorithms, *IEEE/ACM Transactions on Netwroking* Vol. 6(No. 5): 611–624.

Varatkar, G. & Marculescu, R. (2004). On-chip traffic modeling and synthesis for mpeg-2 video applications, *IEEE Transactions of Very Large Scale Integration (VLSI) Systems* Vol. 12(No. 1).

Yin, Q., Jiang, Y., Jiang, S. & Kong, P. Y. (2002). Analysis on generalized stochastically bounded bursty traffic for communication networks, *Proceedings of the 27th IEEE Conference on Local Computer Networks (LCN'02)*.

Advanced Topics in Multimedia Research

Edited by Dr. Sagarmay Deb

ISBN 978-953-51-0078-2
Hard cover, 104 pages
Publisher InTech
Published online 17, February, 2012
Published in print edition February, 2012

As multimedia has become a very important technology, significantly improving people's lives, this book provides an up-to-date scenario of various fields of research being carried out in the area. The book covers topics including web-based co-operative learning, effective distance learning through multimedia, quality control of multimedia on the internet, recovery of damaged images, Network-on-Chip (NoC) as a global communication vehicle, and Network GPS for road conditions (such as traffic and checkpoints). We believe that the book will help researchers in the field to proceed further in their research on multimedia.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821