

When Stochastic Rate-Delay Services Meet Stochastic Network Calculus

Kai Wang and Chuang Lin

Dept. of Computer Science and Technology

Tsinghua University, Beijing 100084, China

{wang_kai, clin}@csnet1.cs.tsinghua.edu.cn

ABSTRACT

The delay-sensitive applications has been rapidly growing in recent years, while the current Internet is not well equipped to support such delay-sensitive traffic. The proposed SRD (Stochastic Rate-Delay) services enable a user to choose between a higher transmission rate or low queuing delay at a congested network link. Delay could be allowed to spike occasionally as long as average low delay remains guaranteed. We build a model using stochastic network calculus to analyze the proposal and shed insight on its fundamental characteristics quantitatively.

Categories and Subject Descriptors

C.2.1 [Computer-Communication Networks]: Network Architecture and Design; C.4 [Performance of system]: Modeling techniques

General Terms

Design, Performance

Keywords

Service Differentiation, Stochastic Network Calculus

1. INTRODUCTION

Delay-sensitive Internet traffic, such as live streaming video and VoIP require low end-to-end delay. The mismatch between the single best-effort service of the current Internet and diverse communication needs of different distributed applications has led to various proposals of alternative architectures, e.g. IntServ, DiffServ, ABE(Alternative Best Effort), BEDS(Best Effort Differentiated Services). Despite technically sound feasible, they both failed to deploy widely.

The proposal of RD (Rate-Delay) services [1] resolves this tension by offering two classes of service: an R service puts an emphasis on a high transmission rate, and a D service supports low queuing delay. The proposal modifies forwarding but not routing, and the user chooses the R or D service by marking a bit in the header of a transmitted packet. The router maintains two FIFO queues on an output link and achieves the intended rate-delay differentiation through link scheduling, tracking of the packets arrival times and dynamic buffer sizing.

Copyright is held by the author/owner(s).
CoNEXT Student Workshop '09, December 1, 2009, Rome, Italy.
ACM 978-1-60558-751-6/09/12.

It is interesting to investigate whether the strict enforcement of the queuing delay constraint is worth the overhead of tracking the arrival times of the enqueued D packets. This paper propose the SRD (Stochastic Rate-Delay) services, which allow the delay to spike occasionally while guarantees average low delay. The presented SRD implementation enforces the constraint through link scheduling and buffer sizing. Stochastic network calculus [2] can be employed in the design of computer networks to provide stochastic service guarantees. On briefly introducing its concepts, we build a model using it to analyze the proposed mechanisms which helps to estimate the achievable level of performance.

2. PRELIMINARIES

Network calculus is a theory using bounds to deal with queuing systems in computer networks. Consider a single flow at a network system, in time interval $(0, t]$, we call $A(t)$, $A^*(t)$, and $S(t)$ the arrival process, the departure process and the service process of the system, with $A(0) = A^*(0) = S(0) = 0$ (in bits). For service guarantee analysis, we are interested in two quantities: backlog and delay. At time t , the backlog $B(t)$ in the system is defined as: $B(t) = A(t) - A^*(t)$; The delay $D(t)$ is defined as: $D(t) = \inf\{\tau \geq 0 : A(t) \leq A^*(t + \tau)\}$. We introduce the following traffic model and the server model [2].

DEFINITION 1. A flow $A(t)$ is said to have a virtual-backlog-centric (v.b.c) stochastic arrival curve α with bounding function f , denoted by $A \sim_{vb} \langle f, \alpha \rangle$, if for all $t \geq 0$ and all $x \geq 0$, there holds

$$P\left\{ \sup_{0 \leq s \leq t} [A(s, t) - \alpha(t - s)] > x \right\} \leq f(x) \quad (1)$$

DEFINITION 2. A server $S(t)$ is said to provide a (weak) stochastic service curve β with bounding function g , denoted by $S \sim_{ws} \langle g, \beta \rangle$, if for all $t \geq 0$ and all $x \geq 0$, there holds

$$P\{A \otimes \beta(t) - A^*(t) > x\} \leq g(x) \quad (2)$$

With the definitions above, various properties are derived under the $(\min, +)$ algebra for analysis.

3. METHODOLOGY

In these section, we use stochastic network calculus to evaluate the performance of the SRD services with a dumbbell topology. The core bottleneck and access links have capacities 100 Mbps and 200 Mbps respectively. The bottleneck link carries N_r R flows and N_d D flows from the senders to the receivers.

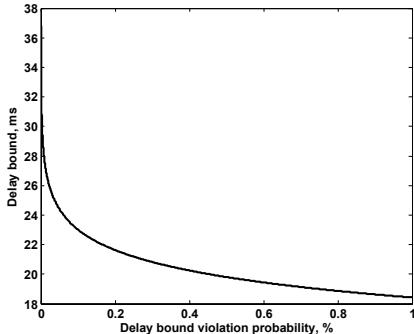


Figure 1: Queuing Delay bound under different violation probability

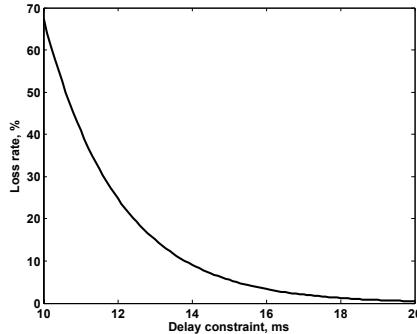


Figure 2: Loss rate under different delay constraints

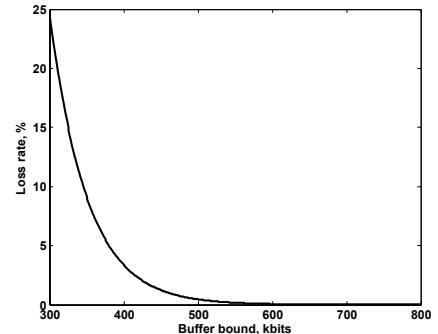


Figure 3: Loss rate under different buffer bound

For analytical purposes, we assume that both R and D queues are continuously backlogged and $R_r + R_d = C$, where R_r and R_d refer to the service rates for the R and D queues. We also assume that every flow within each class transmits at its respective fair rate, ρ_r or ρ_d : $R_r = N_r \rho_r$ and $R_d = N_d \rho_d$. We consider all flows are independent of each other, and assume the D flows have the same v.b.c stochastic arrival curve $A \sim_{vb} \langle f, \rho_d t \rangle$ with $f(x) = ae^{-bx}$, and $A \sim_{vb} \langle f, \rho_r t \rangle$ with $f(x) = ae^{-bx}$ for the R service. Applying the superposition property [2], we could get the aggregate arrival process for the D flows: $A_d \sim_{vb} \langle f_d, \alpha_d \rangle$, with $\alpha_d(t) = \sum_{i=1}^{N_d} \alpha_i(t) = N_d \rho_d t$, and $f_d(x) = f_1 \otimes \dots \otimes f_{N_d}(x)$. And it is similar to the R flows.

The router could be considered as a constant-rate server with link capacity C shared by N flows using the GPS(Generalized Processor Sharing) service discipline. We denote the weight assigned for each R flow and D flow as ϕ_r, ϕ_d respectively, which satisfies $\frac{\phi_r}{\phi_d} = k$, with $\sum_{i=1}^N \phi_i = 1$. The available service rates for the R and D queues should be respectively equal to $C_d = \frac{N_d C}{N_d + k N_r}$, and $C_r = \frac{N_r C}{N_d + k N_r}$. Thus the router provides a weak stochastic service curve $S \sim_{ws} \langle 0, C_d t \rangle$ for the D queue and $S \sim_{ws} \langle 0, C_r t \rangle$ for the R queue. The router supports the desired delay differentiation through buffer sizing for queues, with buffer allocations for the queues as B_r and B_d . If the corresponding buffer does not have enough free space for an arriving packet, the router discards the packet. The D buffer is configured to a much smaller dynamic size to ensure that queuing delay for each forwarded packet stays low on average but is allowed to spike occasionally: $P\{\text{Queuing delay for each forwarded packet} > d\} \leq \varepsilon$, where ε is the tolerable performance violation probability. The R buffer is chosen large enough to utilize the available link rate fully with $B_r = B - B_d$.

4. PERFORMANCE EVALUATIONS

Unless stated otherwise, the default setting for the parameters are $N_d = N_r = 100$, $k = 3$, $d = 20\text{ms}$, $B = 2\text{Mb}$, $\rho_d = 250\text{kbit/s}$, $a = 1$, $b = 0.002$. According to Theorem 3.5 in [2], the queuing delay for the D service, D_d :

$$P\{D_d > h(\alpha_d + x, \beta_d)\} \leq f_d \otimes g(x) \quad (3)$$

We have $P\{D_d > \frac{x}{C_d}\} \leq \inf_{x_1 + \dots + x_{N_d} = x} \sum_{k=1}^{N_d} ae^{-bx_k}$.

On the leverage of Lemma 3 in [3], the delay D_d is bounded by $P\{D_d > \frac{x}{C_d}\} \leq e^{-\frac{xb}{N_d}} (aN_d)$. Then we determine the delay bound D'_d such that $P\{D_d > D'_d\} \leq P_{delay}$, where P_{delay} is a small delay bound violation probability. We have for the

delay bound: $D'_d = \frac{N_d}{C_d b} \log \frac{aN_d}{P_{delay}}$. Figure 1 shows the delay constraints under different delay bound violation probability. With the violation probability of 0.5%, the delay can be constrained within 20 ms.

Also by Theorem 3.5 in [2], the backlog for the D flows:

$$P\{B_d > x\} \leq f_d \otimes g(x - \alpha_d \otimes \beta_d(0)) = f_1 \otimes \dots \otimes f_{N_d}(x) \quad (4)$$

The backlog B_d is bounded by $P\{B_d > x\} \leq e^{-\frac{xb}{N_d}} (aN_d)$. Similarly, we determine the backlog bound B'_d such that $P\{B_d > B'_d\} \leq P_{loss}$, where P_{loss} is the backlog bound violation probability. we have $B'_d = \frac{N_d}{b} \log \frac{aN_d}{P_{loss}}$. As $B_d = R_d t$, Figure 2 illustrates the loss rate for D flows under different delay constraints.

In the same way, the backlog B_r for the R service: $P\{B_r > x\} \leq e^{-\frac{xb}{N_r}} (aN_r)$. We determine the backlog bound B'_r with $P\{B_r > B'_r\} \leq P'_{loss}$. We have for the backlog bound B'_r : $B'_r = \frac{N_r}{b} \log \frac{aN_r}{P'_{loss}}$. Figure 3 depicts the buffer bound violation probability under different buffer bound. It is shown that larger buffer is needed for acceptable loss rate.

5. CONCLUSION

SRD services achieve the intended rate-delay differentiation through simple link scheduling and dynamic buffer sizing, and can tolerate some amount of excess delay while guarantees average low delay. Since many applications perform well with stochastic service guarantees, stochastic network calculus plays an important role in the analysis and provision of service guarantees in such networks. We would further explore new traffic models and server models that better suit the network analysis.

6. ACKNOWLEDGMENT

This research is supported by the 973 Program of China (No. 2009CB320504). And the authors wish to thank Prof. Yuming Jiang from NTNU for fruitful discussions.

7. REFERENCES

- [1] Maxim Podlesny and Sergey Gorinsky, RD Network Services: Differentiation through Performance Incentives, In Proceedings of ACM SIGCOMM'08.
- [2] Y. Jiang, A Basic Stochastic Network Calculus, In Proceedings of ACM SIGCOMM'06.
- [3] F. Ciucu, A. Burchard, and J. Liebeherr. A network service curve approach for the stochastic analysis of networks. ACM SIGMETRICS'05.