Voice and Multi-fractal Data in the Internet

Markus Fiedler, Patrik Carlsson, Arne Nilsson
Blekinge Institute of Technology
{Markus.Fiedler,Patrik.Carlsson, Arne.Nilsson}@bth.se

Abstract

In the Internet era there is a need to understand how traditional Public Switched Telephone Network (PSTN) services and the new Internet services can coexist. The best-effort based IP network will have to maintain the level of service that customers from the PSTN world expect. This study contributes to the traffic engineering of such a system by presenting an analytical model for integrating Voice over IP (VoIP) traffic with multi-fractal data traffic. This model is based on the stochastic fluid flow model. Effects on delay and loss performance of adding VoIP traffic to a data link as well as capacity requirements to maintain a certain quality of service are discussed

1. Introduction

This is the era of convergence of two distinct technologies, the Public Switched Telephone Network (PSTN) and the IP network. For the users the most notable change will be the new service integration. To successfully merge these two networks the quality of service in the new network must be at least at the same level as the quality the users are used to. One important performance parameter in this context is the delay an application experiences. Long and heavily varying delays in the network as well as high packet loss will lead to low throughput and poor response times at the application level. In this paper we investigate analytically what will happen when data and voice traffic are merged. Recent analyses of real traffic, *e.g.* [5, 11], indicate that data traffic exhibits long-range dependence as well as self-similar or multi-fractal properties. It is a well-known fact that voice traffic is much well behaved. The main question then is to investigate how well behaved traffic interact with a not so well behaving traffic. Of interest is to determine which traffic type will suffer from the integration and to what extent. In our analysis a multi-fractal stochastic process [7] form the data process used in this study. The process exhibits long-range dependency on certain time scales. For the voice traffic, standard models acting on call and activity level are used [9]. The stochastic fluid flow model is the basis for the performance model [1, 4]. The flow of packets from a source is modelled by a fluid flow whose intensity varies over time. Recently, this model has successfully been applied to TCP and related scenarios by Misra, Gong and Towsley [6]. Fluid flow analysis for multi-fractal stochastic processes has been discussed in [2]. The model used in the analysis permits the integration of both traffic types as mentioned above. From the model we can calculate individual performance parameters in terms of loss ratios and delay quantiles, and this is presented in the paper. We focus on the best-effort scenario [10] shown in Figure 1. A full-duplex link connects an Intranet with the Internet. This link is assumed to carry both data traffic and VoIP traffic. Both types of traffic use the link in a best-effort way. Proper link dimensioning is of course needed in order to assure the desired quality of service. The remainder of the paper is organized as follows. Section 2 describes the creation of the multi-fractal data process, and Section 3 discusses the models for the voice streams. The fluid flow analysis and calculation of the performance parameters of interest (delay quantiles and loss ratios) are described in Section 4. Section 5 presents numerical results for a case study, and Section 6 concludes the paper.

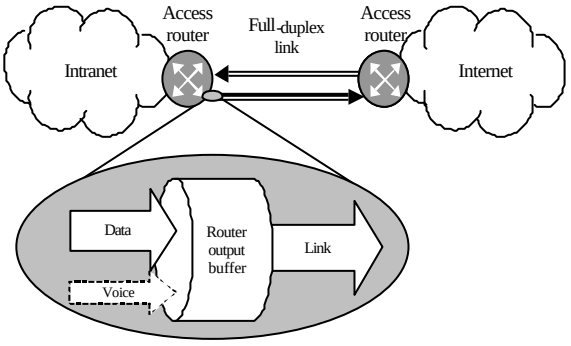

Figure 1. The Scenario

Table 1. Estimates of B for different data processes

N	Longest mean cycle time	γ for small time scales	γ for large time scales
4	10^3 s ~ 16.6 minutes	-1.2	-2.0
5	10^4 s ~ 2.7 hours	-1.1	-2.0
6	10^5 s ~ 27 hours	-1.1	-2.0
7	10^6 s ~ 11.6 days	-1.1	-1.9

2. Multi-fractal Data Process

The process that is used to emulate data traffic displays multi-fractal properties. N independent Markov Modulated Rate Processes (MMRP) [7] are multiplied together to form the desired multi-fractal process. In the form we use here it is assumed that each process is a two state process.

$$R(t) = \prod_{i=0}^{N-1} R_i(t) \text{ Mbps} \quad R_i(t) = R_0(b^i t), b = 1 \tag{1}$$

The difference between the sub-processes is the timescale: Compared to sub-process 0, sub-process i is slowed down by a factor b^i. The sub-processes' transition rates are $\alpha_i = b^{-i}\alpha_0$ and $\beta_i = b^{-i}\beta_0$. The mean cycle time for sub-process i is given by $1/\alpha_i + 1/\beta_i = b^i$ s. A more detailed description of the processes, its properties and creation can be found in [2]. The transition rate matrix is formed using Kronecker addition, the diagonal rate matrix contains products of the contributions of the N sub-processes:

$$\mathbf{M}_i = \begin{pmatrix} -\alpha_i & \alpha_i \\ \beta_i & -\beta_i \end{pmatrix} \quad \mathbf{R}_i = \begin{pmatrix} \lambda & 0 \\ 0 & h \end{pmatrix} \tag{2}$$

$$\mathbf{M}_D = \mathbf{M}_0 \oplus \mathbf{M}_1 \oplus \cdots \oplus \mathbf{M}_{N-1}$$

$$\mathbf{R}_D = \text{diag}[\lambda^N, \lambda^{N-1}h, \lambda^{N-2}h\lambda, \lambda^{N-2}h^2, \cdots, h^N] \text{ Mbps} \tag{3}$$

The mean data rate for the process is

$$\mathbf{E}R_D = \left(\frac{\lambda}{\alpha_0 + \beta_0}\right)^N \text{ Mbps} \tag{4}$$

The process has $\gamma_D = 2^N$ states. The parameters $\alpha_0 = 2/\text{s}$ and $\beta_0 = 2/\text{s}$ were chosen such that the shortest cycle time is equal to one second. The parameter $b = 10$ and this implies that the mean cycle times of adjacent sub-processes differ with a factor of 10. The parameters $\lambda = 0.6h$ of the sub-processes are chosen dependent on $\mathbf{E}R_D$. The choice of $N = 5$ gives a longest mean cycle time around 2.7 hours, and small variations on a time scale of a day, see Figure 2. This choice was deemed to be of interest in the case study. Next we take the opportunity to present some of the characteristics of this multi-fractal process. Let $n(t,T)$ be the mean number of data units (Megabit) produced by $R(t)$ during the interval $]t-T, T]$. In Figure 2 we plot this function for the case when $\mathbf{E}R_D = 1$ (Mbps). For $T = 1$ s and 10^2 s the process variance is quite large, the tme-scales correspond to 200 s and 5.5 hours of data. As the time-scale increases the variations decrease, especially after we have passed $T = 10^4$ s which is the mean cycle time of the slowest sub-process. We notice that at $T = 10^6$ s, the variations have almost vanished. This behaviour becomes even clearer in Figure 3 where log-log plots of the normalized variance $\mathbf{Var}[n(t,T)/T]$ versus the time interval T of different multi-fractal processes are compared. All processes display the same characteristics: there is one rather flat section, followed by one section that seems to occur after the longest cycle time has been passed, in which the

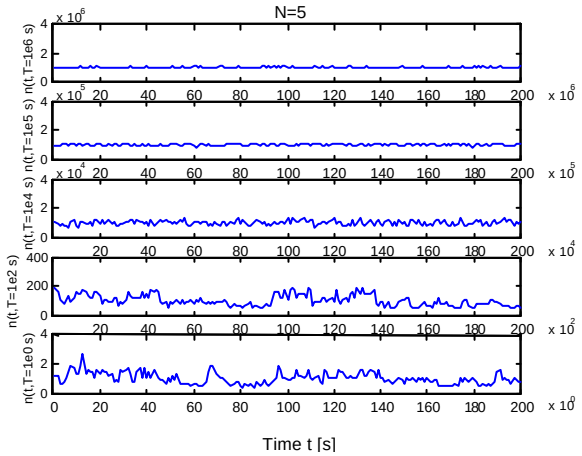

Figure 2. Time plot for the data process on different time scales.

gradient tends towards –1 indicating short-range dependence. In Table 1 we show estimates for ? on short and long time scales, which is a measure reflecting short-range (? ? ?2) or long-range (? ? ?1) dependence. We use $\mathbf{Var}[n(t,T)] ? (T/\mathrm{s})^{??3}$ as defined in [8].

Recent measurement that we have preformed shows a process that has roughly the same shape as the data processes here. A preview of raw-data is presented in Figure 5. Please notice that this is almost raw data, and hence the y-scale will not stand to be compared with the simulated process. It is also worth noting that the last point at 10^5 is not a mean but formed from only one value, which might explain the gradient change.

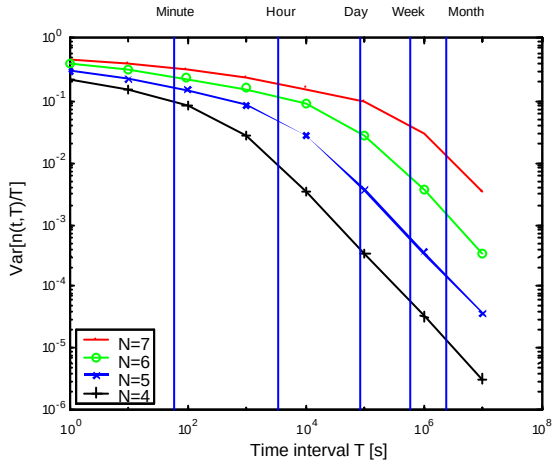

Figure 3. Normalized-variance-time plots for different data processes. The difference between the processes are the number of sub-processes, *N*, used in the construction.

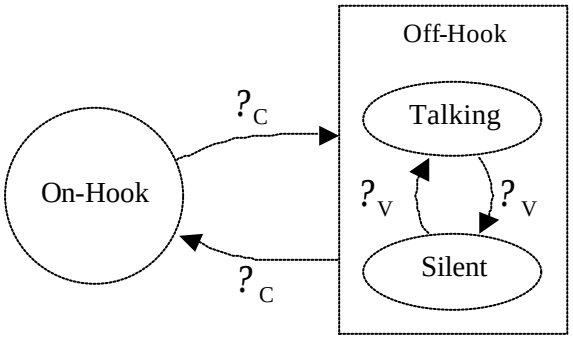

Figure 4. State description for three state voice model.

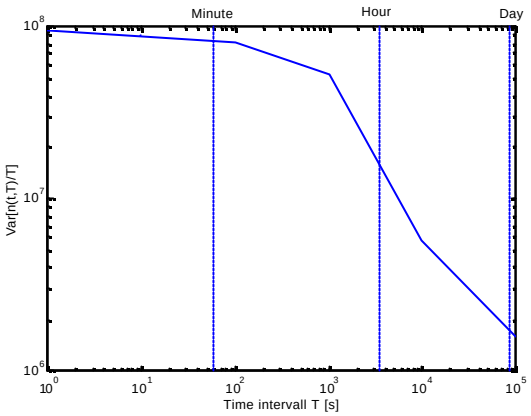

Figure 5. Preview of a Variance-time plot from a measurement preformed at BTH.

3. Voice Process

The VoIP stream that we emulate consists of two parts, a codec/transmission-model and a speech-model. We have chosen to use two different models for the speech-model. The codec/transmission-model is formed by a G.711 codec with voice activity detection (VAD). Using a VAD results in no data transfers from the codec to the lower layers in the transmission stack when there is no speech present. We are also assuming that once the codec does operate, it generates data packets at intervals corresponding to either 5 or 30 ms of speech. This means that 40 or 240 bytes of data leave the codec in each packet. The data is then transferred in to the payload field of the Real Time Protocol (RTP). This gives us the data rates for the different block lengths to be h_V ? 160 kbps or 80 kbps.

The three state voice processes, VoIP($?_V$? 3,h_V), consists of one “on-hook” state, and two “off-hook” states. ”Talking” and ”silent” identify the ”off-hook” states. The model is shown in Figure 4. The off-hook state model is the very old two state model of a voice source. A recent reference can be found in [9]. Using this source model the mean data rate will be

$$\mathbf{E}R_V ? \frac{?_C}{?_C ? ?_C} ? \frac{?_V}{?_V ? ?_V} ? h_V \qquad (5)$$

Typical values [9] are $?_C ? ?_C ? 1/480$ s, $?_V ? 1/0.6$ s and $?_V ? 1/0.4$ s. The transition matrix and the rate matrix used in the fluid flow analysis are given as

$$\mathbf{M}_V = \begin{pmatrix} -(\lambda_C + \frac{\lambda_V}{\lambda_V + \mu_V}\lambda_C) & \mu_C & \lambda_V \\ \frac{\lambda_V}{\lambda_V + \mu_V}\lambda_C & -(\mu_C + \lambda_V) & \mu_V \\ \frac{\lambda_V}{\lambda_V + \mu_V}\lambda_C & \lambda_V & -(\mu_C + \mu_V) \end{pmatrix}$$

$$\mathbf{R}_V = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & h_V \end{pmatrix} \quad (6)$$

The two state process, VoIP($\lambda_V = 2, h_V$), is a simplification of the previous process, where the "on-hook" state is ignored ($\lambda_C = 0$). This is a widely used worst-case model. The mean data rate becomes

$$\mathbf{E}R_V = \frac{\lambda_V}{\lambda_V + \mu_V} h_V \quad (7)$$

and the matrices have the following form:

$$\mathbf{M}_V = \begin{pmatrix} -\lambda_V & \lambda_V \\ \mu_V & -\mu_V \end{pmatrix} \quad \mathbf{R}_V = \begin{pmatrix} 0 & 0 \\ 0 & h_V \end{pmatrix} \quad (8)$$

This model allows us to simply model n_v independent sources using the basic birth-death model [1, 3].

4. Fluid Flow Analysis

The router output buffer, see Figure 1, is modelled as a fluid flow buffer, whose content and size are denoted by X and K, respectively $(0 \le X \le K)$. Fluid flow analysis [1, 4, 3] is used to derive the vector of the joint comple-mentary buffer content distribution $\vec{F}(x)$ with $F_s(x) = \Pr\{\text{buffercontent } X \le x, \text{state } S = s\}$ from which the performance parameters of interest may be derived. The system of $\lambda_D \mu_V$ differential equations to be solved is given by

$$(\mathbf{R}_D \mathbf{R}_V - C\mathbf{I})\frac{d}{dx}\vec{F}(x) = (\mathbf{M}_D \mathbf{M}_V)\vec{F}(x) \quad (9)$$

The matrices are defined in (3, 6, 8) and $\mathbf{I}$ is an identity matrix of size $\lambda_D \mu_V$.

The set of eigen-values and right eigen-vectors $\{\{z_q\},\{\vec{\psi}_q\}\}, q = 0,\ldots,\lambda_D \mu_V - 1$ can be obtained by solving $z_q(\mathbf{R}_d \mathbf{R}_v - C\mathbf{I})\vec{\psi}_q = (\mathbf{M}_d \mathbf{M}_v)\vec{\psi}_q$. Let $\vec{\pi}, \pi_s = \Pr\{S = s\}$, denote the vector of state probabilities and $\{a_q\}$ a set of coefficients determined by the boundary conditions. Then, the solution to (9) and its complement $\vec{G}(x)$ read

$$\vec{F}(x) = \sum_{\Re(z_q)<0} a_q \vec{\psi}_q e^{z_q x} \quad \vec{G}(x) = \vec{\pi} - \vec{F}(x) \quad (10)$$

A detailed description of fluid flow analysis and its numerical treatment can be found in [3, 2]. Let $r_{D,q}$ be the

Table 2. Performance parameters for different VoIP streams

Voice traffic model	Delay quantiles		Loss ratio		Mean data rate, $\mathbf{E}R_V$ (kbps)	Capacity compensation $C(1)$ (kbps)
	$\lambda_{2,D}$ (ms)	$\lambda_{2,V}$ (ms)	l_D ($\times 10^{-5}$)	l_V ($\times 10^{-5}$)		
?	50.0	?	2.24	?	?	?
3,80	52.4	35.3	2.94	1.38	16	16.54
2,80	54.6	33.1	3.64	1.38	32	32.75
3,160	54.9	39.7	3.96	2.41	32	34.08
2,160	59.3	37.2	5.68	2.42	64	67.00

input rate of data process, $r_{V,q}$ the input rate of the voice process and $r_q = r_{D,q} + r_{V,q}$ the total input rate to the concentrator in system state q, respectively. The loss ratios for data and voice in a buffer of finite size K is then given by

$$l_D = \frac{1}{\mathbf{E}R_D} \sum_{q:r_q<C} G_q(K) r_{D,q} \left(1 - \frac{C}{r_q}\right)$$

$$l_V = \frac{1}{\mathbf{E}R_V} \sum_{q:r_q<C} G_q(K) r_{V,q} \left(1 - \frac{C}{r_q}\right) \quad (11)$$

The probability that data or voice experiences a delay longer than $w = x/C$ is obtained as [4]

$$G_D(w) = \frac{1}{\mathbf{E}R_D} \sum_{q} G_q(w)\, r_{D,q}$$

$$G_V(w) = \frac{1}{\mathbf{E}R_V} \sum_{q} G_q(w)\, r_{V,q} \quad (12)$$

Finally, the 10^{-k}-quantile for the delay is given by

$$\xi_{k,D} = \{w \mid G_D(w) = 10^{-k}\}$$

$$\xi_{k,V} = \{w \mid G_V(w) = 10^{-k}\} \quad (13)$$

This expression gives the delay that is exceeded on average by one out of 10^k data or voice packets. Note that because of the finite buffer size, a delay quantile ξ_k does not exist if $G(K/C) < 10^{-k}$.

5. Results

In our study we assumed a link capacity of 9.5 Mbps on the IP level. For instance, this speed is obtained on a 10 Mbps full-duplex Ethernet link for an average packet size of 500 bytes. We set the size of the output packet buffer to 128 Kbytes, which limits the maximal packet delay to 110.4 ms. For simple comparison, we use a delay quantile (13) of $\xi_{2,D} = 50\text{ms}$ for data traffic alone, which we approach by choosing a mean data rate of $\mathbf{E}R_D$ = 3.249 Mbps. Thus, in the absence of voice traffic, the link is loaded by 34.2 %. Then, n_v streams of VoIP(ξ_V, h_V) traffic are added. We begin with looking at how data traffic and one VoIP stream interact with each other. Table 2 reveals that the delay quantiles for data grow almost linearly with the mean data rate of the voice source. Both delay quantiles and loss ratios for the VoIP stream are lower than those for the data stream. In other words, VoIP streams suffer less from the integration with this kind of multi-fractal data traffic. The loss ratios for voice depend mostly on the data rate of the voice source when talking, while for data, the higher burstiness of the VoIP(3,160) stream — compared to the VoIP(2,80) stream — induces slightly higher loss on the data stream. The last column in Table 2 contains the capacity increase $C(1)$ needed to drive the delay quantile for data back to 50 ms, the value before one VoIP stream has been added. For any of the VoIP streams, this increase is slightly larger than its mean data rate. However, this type of capacity increase is difficult to obtain in the Ethernet case that scales by factors of ten, but other technologies (*e.g.* ATM) would allow a more precise scaling.

The impact of adding several VoIP(2,80) streams on the delay quantiles is investigated next. Additionally we look at a low-capacity link with merely 0.95 Mbps. This is of interest since such a link may appear in a DSL environment. A delay quantile of $\xi_{2,d} = 50\text{ms}$ is obtained when a mean data rate of 324.9 kbps is used. To lower the maximal delay from unrealistic 1104 ms to 276 ms we decrease the buffer size to 32 kbytes. Figure 6 shows the relationship between the delay quantiles for data traffic and the number of VoIP streams in the case of the high-capacity link. The delay quantiles for the voice traffic are smaller and grow more slowly than those of the data traffic; the growth becomes almost linear for more than five VoIP streams. Loss ratios also grow to their maximal values of 4.37e-4 for data and 1.57e-4 for voice, which are obtained for ten VoIP streams. This link manages to integrate several VoIP streams without any significant impact on performance. Table 3 shows the delay quantiles in case of the low-capacity link. The growth is approximately linear for voice traffic, but almost exponential for data traffic. No 10^{-2} quantile is obtained for the data traffic when more than two VoIP streams are integrated. The same goes for the voice traffic but here the quantile disappears after the fifth stream has been added. As expected, both data and voice suffer much more from each other in this low-capacity scenario, and re-dimensioning seems inevitable.

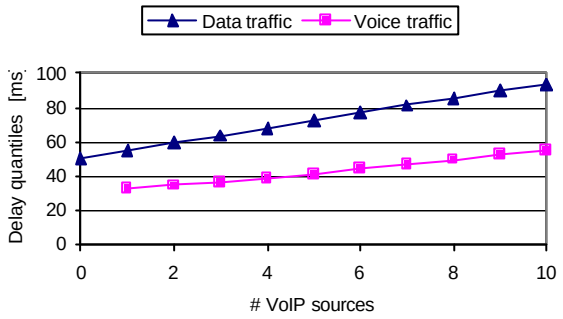

Figure 5. Delay quantiles experienced by the traffic when VoIP streams are added without increasing the capacity.

Table 3. Delay quantiles for data and voice traffic when C= 0.95 Mbps

n_V	0	1	2	3	4	5
$?_{2,D}$	50	108	176	?	?	?
$?_{2,V}$	?	79	123	173	222	258

6. Conclusions

We presented an analytical study of the expected behaviour when data traffic with multi-fractal properties and Voice over IP traffic are integrated on a best-effort link. The synthesis of data and voice traffic and the calculation of individual performance parameters were described. Our results show that a VoIP connection experiences smaller loss and delay than data, which means that VoIP traffic does not suffer extensively from the integration. On a high-capacity link, delay quantiles grow roughly linearly with the number of VoIP streams. Some issues are left for further study. The parameters of the multi-fractal data traffic model should be estimated from real traffic. The large parameter space should give opportunities to adapt the process to specific traffic types. This has to be preceded by a thorough study of the impact of parameters on the variance-time relation. Once parameters have been matched, advanced dimensioning rules may be generated.

References

[1] D. Anick, D. Mitra, and M.M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal*, 61(8):1871–1894 (1982).

[2] P. Carlsson, M. Fiedler: Multifractal products of stochastic processes: Fluid flow analysis. *Proceedings of 15th Nordic Teletraffic Seminar (NTS-15)*, Lund, Aug. 22–24, 2000.

[3] M. Fiedler, H. Voos. New results on the numerical stability of the stochastic fluid flow model analysis. *Proceedings of Networking 2000 Conference*, Paris, May 2000.

[4] K. Kontovasilis, N. Mitrou. Bursty traffic modelling and efficient analysis algorithms via fluid-flow models for ATM IBCN. *Annals of Operations Research*, 49:279–323 (1994).

[5] W. Leland, W. Willinger, M. Taqqu, D. Wilson. On the Self similar Nature of Ethernet Traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15 (1994).

[6] V. Misra, W. Gong, and D. Towsley. Fluid-based analysis of a network of AQM routers supporting TCP flows with an application to RED. *Proceedings of ACM SIGCOMM '00*, Stockholm, Sweden, August 2000.

[7] P. Mannersalo, I. Norros, and R. Riedi. Multifractal products of stochastic processes: A preview. *COST-257 Technical Document 257TD(99)31*. http://nero.informatik.uni-wuerzburg.de/cost/Final/

[8] I. Norros. On the use of fractional Brownian motion in the theory of connectionless networks. *IEEE Journal on Selected Areas in Communications*, 13(6):953–962 (1995).

[9] M. Schwartz. *Broadband Integrated Networks*. Prentice Hall 1996.

[10] D. Verma. *Supporting Service Level Agreements on IP Networks*. Macmillan 1999.

[11] A. Veres, Zs. Kenesi, S. Molnár, and G. Vattay. On the propagation of long-range dependence in the Internet. *Proceedings of ACM SIGCOMM '00*, Stockholm, Sweden, August 2000.