On the Variability of Internet Traffic

Georgios Y Lazarou

Information and Telecommunication Technology Center
Department of Electrical Engineering and Computer Science
The University of Kansas, Lawrence (KS)

August 1, 2000

Outline

- Introduction and Motivation
- Some Background Information on Long-Range Dependence and Self-Similarity
- Characterizing the Variability of Traffic
  - examples
- Simulation Study on the Variability of Internet Traffic
- Conclusions and Future Work

What Is The Problem?

- Many empirical studies on a variety of networks have shown that traffic exhibits high variability
  - traffic is bursty (variable) over a wide range of time scales
- High variability was shown to have a significant impact on network performance
- Several studies claim that the high variability in traffic is due to long-range dependence (LRD) property of traffic process
- The assertion that traffic has LRD triggered a large effort in
  - explaining the cause of LRD in traffic
  - studying the impact of LRD on network performance, and
  - creating new traffic models that have LRD

What Is This Research All About?

- Develop a new theoretical and practical framework to accurately characterize the variability and correlation structure of a typical network traffic process at each time scale
- Determine that conventional traffic models can capture the high variability of traffic empirically observed over a wide range of time scales
- Investigate the contribution of TCP's dynamics to LRD or high variability of traffic

Why New Measure of Variability?

- Most commonly used measures of traffic burstiness are
  - peak-to-mean ratio (= 1 for CBR Flows)
  - squared coefficient of variation of interarrival times: $\frac{Var[X]}{(E[X])^2}$
  - indices of dispersion for intervals and counts:
    $J_k = \frac{Var[X_1 + \cdots + X_k]}{k(E[X])^2}$ $\quad IDC(t) = \frac{Var[N(t)]}{E[N(t)]}$

    $\Rightarrow$ any value other than one $\rightarrow$ bursty traffic
  - Hurst parameter (H)

    $\Rightarrow$ do not capture the fluctuation of the degree of traffic burstiness across time scales

    $\Rightarrow$ our novel measure of variability based on the slope of the IDC curve at each time scale and directly related to H

Why Care About Conventional Traffic Models?

- Analytically simpler and tractable than models with LRD
- The exact degree of traffic variability over all time scales can analytically be obtained
- Considerable amount of work has already been done on analyzing network performance (i.e., queueing behavior) associated with these models
- Performance evaluation depends on traffic characteristics over a finite range of time scales specific to system under study (i.e., maximum buffer size)
  - any model can be used as long as it captures traffic behavior over this range of time scales

Does Network Traffic Have LRD?

- There are not strong evidence that real network traffic exhibits long-range dependence
  - definition of LRD applies only to infinite time sequences
  - need to check the tail of correlation structure for LRD
  - all empirically collected and analyzed traffic traces were one to about three hours long

    $\Rightarrow$ 360000 to 1000000 samples for sampling period of 10 ms

    $\Rightarrow$ long enough for capturing the variability and correlation structure over the time scales associated with network performance

    $\Rightarrow$ but not long enough to assert that traffic processes have LRD

What TCP Has to Do With the Variability of Traffic?

- TCP traffic was used in most studies to
  - detect the presence of LRD in network traffic, or
  - give a possible explanation of what causes the asserted LRD
- The results from these studies claim that
  - aggregate TCP traffic exhibits LRD behavior over a wide range of time scales
  - not surprising observation since TCP is a bursty protocol

    $\Rightarrow$ transmits packets as fast as it cans and then becomes idle waiting for acknowledgments
  - presence of LRD depends on whether a reliable, flow- and congestion-controlled protocol is employed at transport layer
  - natural to assume that the dynamics of TCP have a great impact on its traffic variability

Major Contributions

- A new measured of variability: index of variability $H_v(\tau)$
  - a plot of $H_v(\tau)$ describes the behavior of a traffic process in terms of its variability over a range of time scales
  - better measure for capturing the burstiness of traffic than H
- The results show that
  - traditional models can capture the high variability observed in network traffic over a wide range of time scales
  - the amount of correlation that a traffic process has at a particular time scale does not alone determine the degree of variability at that time scale
  - the dynamics of TCP alone can not cause considerable variability over a substantial range of time scales

Long-Range Dependence

Definition: A weakly stationary discrete-time real-valued stochastic process $Y = \{Y_t, t = 0, 1, 2, \ldots\}$ ($\mu = E[Y_t] = \text{constant}$, $\sigma^2 = E[(Y_t - \mu)^2] < \infty$) with an autocorrelation function $r(k)$ is called long-range dependent if

$$\sum_{k=1}^{\infty} r(k) = \sum_{k=1}^{\infty} \frac{E[(Y_t - \mu)(Y_{t+k} - \mu)]}{\sigma^2} = \infty$$

- $r(k)$ measures the correlation between elements of Y separated by k units of time
- correlations between observations that are separated in time decay to zero at a slower rate that one would expect from data following Markov-type (i.e., SRD) models
- Self-Similar Processes: most popular models with LRD
  - statistical properties remain same over all time scales
  - several definitions, asymptotically second-order

Asymptotically Second-Order Self-Similar Processes

Assume that

$$r(k) \sim k^{-\beta} L(k) \qquad as \;\; k \to \infty$$

where $0 < \beta < 1$ and L is slowly varying at infinity, that is,

$$\lim_{k \to \infty} \frac{L(kx)}{L(k)} = 1 \qquad \forall x > 0$$

i.e., $L(t) = const$, $L(t) = log(t)$.

For each $m = 1, 2, 3, \ldots$, let $Y^{(m)} = \{Y_k^{(m)}, k = 1, 2, 3, \ldots\}$, where

$$Y_k^{(m)} = \frac{Y_{km-m+1} + \cdots + Y_{km}}{m} \qquad k \geq 1$$

Definition: Y is called asymptotically second-order self-similar with self-similarity parameter H if $\frac{Y^{(m)}}{m^{H-1}}$ has the same variance and autocorrelation as Y as $m \to \infty$. That is, $\forall k$ large enough,

$$r^{(m)}(k) \to r(k) \qquad as \;\; m \to \infty$$

Hurst Parameter

- Most important parameter of self-similar processes
  - measures the degree of self-similarity
  - expresses the speed of decay of autocorrelation function
  - $0.5 < H < 1 : \Rightarrow$ LRD $\quad 0 < H \leq 0.5 : \Rightarrow$ SRD
- Claimed to be a good measure of variability
  - the higher the value of H, the burstier the traffic
- Popular belief: higher the H, poorer the queueing performance
  - but, there are examples showing otherwise
  - different processes with same H can generate vastly different queueing behavior
- Conclusion: the single value Hurst parameter does not capture the fluctuation of traffic burstiness across time scales

Estimation of Hurst Parameter: Aggregated Variance Method

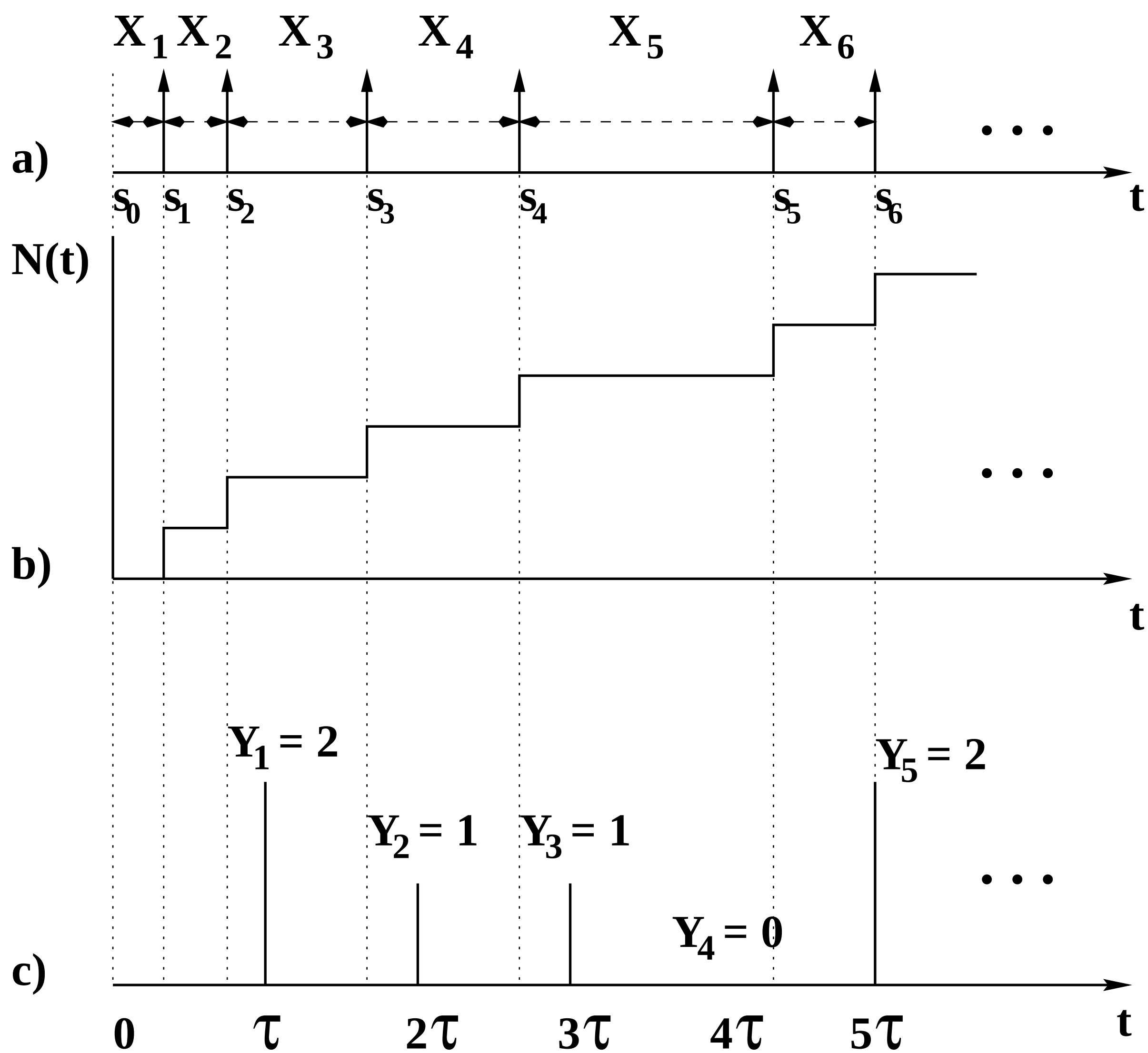
Assume a traffic sequence $\hat{Y}$ of length N. Construct $\hat{Y}^{(m)}$ by dividing $\hat{Y}$ into blocks of length m, and averaging the sequence over each block. Its sample variance is then given by:

$$\hat{Var}[Y^{(m)}] = \frac{\sum_{k=1}^{\frac{N}{m}} (Y^{(m)}(k) - \bar{Y})^2}{\frac{N}{m}} \qquad where \qquad \bar{Y} = \frac{\sum_{t=1}^{N} Y_t}{N}$$

For successive values of m that are equidistant on a log scale, the sample variance of the aggregated series is plotted versus m on a log-log plot. By fitting a least-squares line

$$\hat{H} = 1 - \frac{slope}{2}$$

Relating Packet Traffic With Point Processes (I)



Relating Packet Traffic With Point Processes (II)

- Counting process: $\{N(t), t \geq 0\}$, (weakly) stationary where

$$N(t) = sup\{n : n = 0, 1, 2, \ldots; S_n \leq t\}$$

- Traffic process: $Y = \{Y_n(\tau), \tau > 0, n = 1, 2, \ldots\}$ where

$$Y_n(\tau) = N[n\tau] - N[(n-1)\tau]$$

- Index of dispersion for counts:

$$IDC(t) \equiv \frac{Var[N(t)]}{E[N(t)]} = \frac{Var[N(t)]}{\lambda t}$$

  λ : mean event (packet) arrival rate

- $IDC(t = m\tau) = \frac{m}{\lambda\tau} Var[Y^{(m)}] \qquad m = 1, 2, 3, \ldots$
  $Y^{(m)}$: aggregated packet (byte) count process

Index of Variability for Traffic Processes

For a self-similar process, plotting $log(IDC(m\tau))$ versus $log(m)$ results in an asymptotic straight line with slope $2H - 1$

Definition: For a general stationary traffic process Y, we call

$$H_v(\tau) \equiv \frac{\frac{d(log(IDC(\tau)))}{d(log(\tau))} + 1}{2}$$

the index of variability of Y for the time scale τ

Suppose Y results from the superposition of M independent traffic streams:

$$H_v(\tau) = 0.5\tau \left(\frac{\sum_{i=1}^{M} \frac{dVar[N_i(\tau)]}{d\tau}}{\sum_{i=1}^{M} Var[N_i(\tau)]} \right) = \frac{1}{2} \left\{ 1 + \tau \left(\frac{\sum_{i=1}^{M} \frac{d(IDC_i(\tau))}{d\tau} \left(\frac{1}{\Lambda_i}\right)}{\sum_{i=1}^{M} \left(\frac{IDC_i(\tau)}{\Lambda_i}\right)} \right) \right\}$$

$\Lambda_i = \frac{\sum_{j=1}^{M} \lambda_j}{\lambda_i}$ Poisson: $\frac{d(IDC_i(\tau))}{d\tau} = 0 \ \ \forall \tau, i \Rightarrow H_v(\tau) = 0.5 \ \ \forall \tau$

If $\lim_{\tau \to \infty} \left(\sum_{i=1}^{M} \left(\frac{IDC_i(\tau)}{\Lambda_i} \right) \right) = c < \infty$, then $\lim_{\tau \to \infty} H_v(\tau) = 0.5$

Correlation Structure of Traffic Process Y

Autocovariance function:

$$C_k(\tau) = \begin{cases} \frac{1}{2}Var[N((k+1)\tau)] + \frac{1}{2}Var[N((k-1)\tau)] - Var[N(k\tau)] & k > 1, \\ \frac{1}{2}Var[N(2\tau)] - Var[N(\tau)] & k = 1. \end{cases}$$

Autocorrelation function: $r_k(\tau) = \frac{C_k(\tau)}{Var[N(\tau)]} \qquad k = 0, 1, 2, \ldots$

Correlation intensity:

$$R(\tau) \equiv \sum_{k=1}^{\infty} r_k(\tau) = \frac{1}{2}\left(\frac{\lim_{k\to\infty} IDC(k\tau)}{IDC(\tau)} - 1\right)$$

Suppose Y is a superposition of M independent renewal processes

$$\lim_{k\to\infty} IDC(k\tau) = \sum_{i=1}^{M}\left(\frac{\mathcal{C}_i^2(X)}{\Lambda_i}\right) \quad where \;\; \mathcal{C}^2(X) = \frac{Var[X]}{(E[X])^2}$$

If $\mathcal{C}_i^2(X) = \infty$ for at least one $i \Rightarrow R(\tau) = \infty \Rightarrow Y$ is LRD process

Example: Hyperexponential Distribution of Order Two (I)

Suppose the underlying point processes of Y is a stationary renewal process with interarrival times hyperexponentially distributed of order two

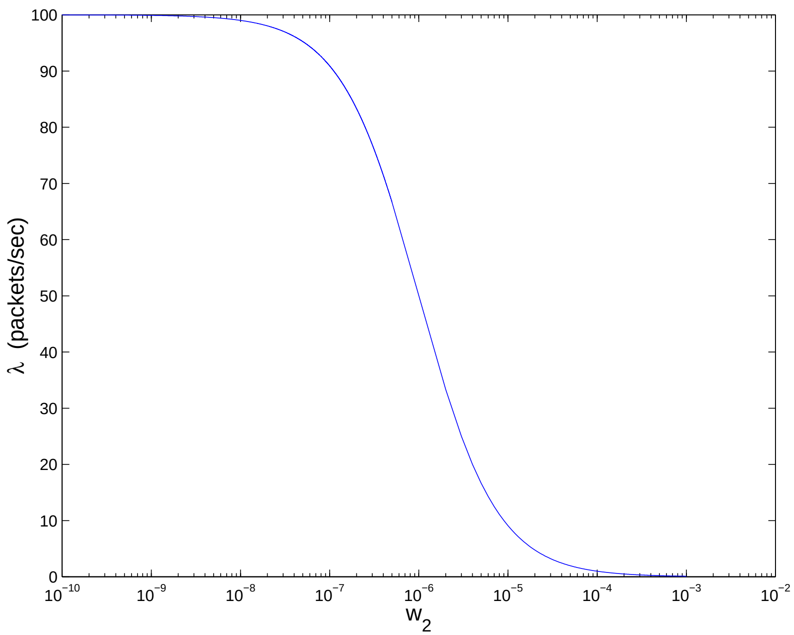
Pdf: $f_2(x) = w_1 a e^{-ax} + w_2 b e^{-bx}$ where $w_1 + w_2 = 1$

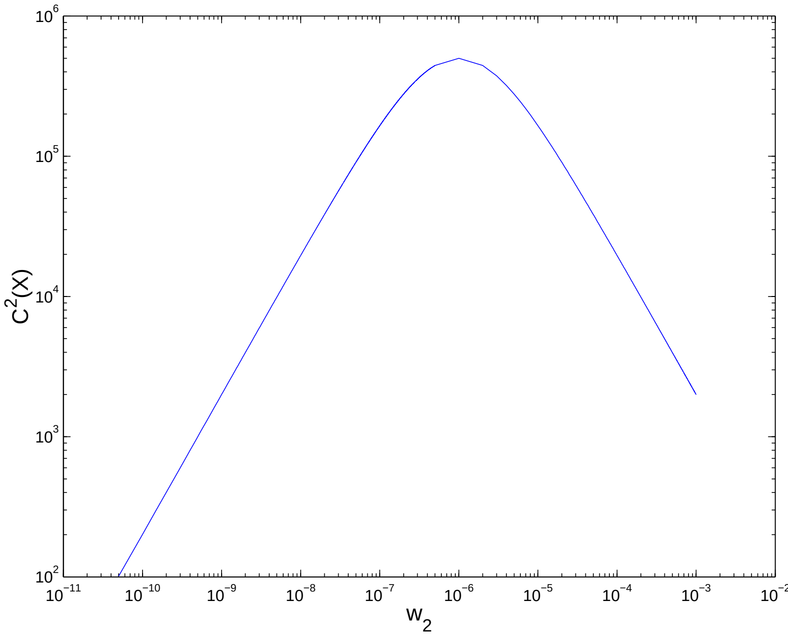
Then:

- $\lambda = \frac{1}{E[X]} = \frac{ab}{aw_2 + bw_1}$ $\quad \mathcal{C}^2(X) = 2\left[\frac{a^2 w_2 + b^2 w_1}{(aw_2+bw_1)^2}\right] - 1$
- $Var[N(\tau)] = \frac{2\lambda[(aw_1+bw_2)^2 - (a^2w_1+b^2w_2)]}{(aw_2+bw_1)^3}\left(1 - e^{-[aw_2+bw_1]\tau}\right) + \lambda\mathcal{C}^2(X)\tau$
- $IDC(\tau) = \frac{2[(aw_1+bw_2)^2 - (a^2w_1+b^2w_2)]}{(aw_2+bw_1)^3}\left(\frac{1-e^{-[aw_2+bw_1]\tau}}{\tau}\right) + \mathcal{C}^2(X)$
- $\lim_{\tau\to\infty} IDC(\tau) = \mathcal{C}^2(X) \Rightarrow R(\tau) < \infty \Rightarrow Y$ is SRD process
- If $a = b$ then $[(aw_1 + bw_2)^2 - (a^2w_1 + b^2w_2)] = 0$ and $\mathcal{C}^2(X) = 1$

  $\Rightarrow Var[N(t)] = \lambda t$ and $IDC(t) = 1$, i.e., Poisson process

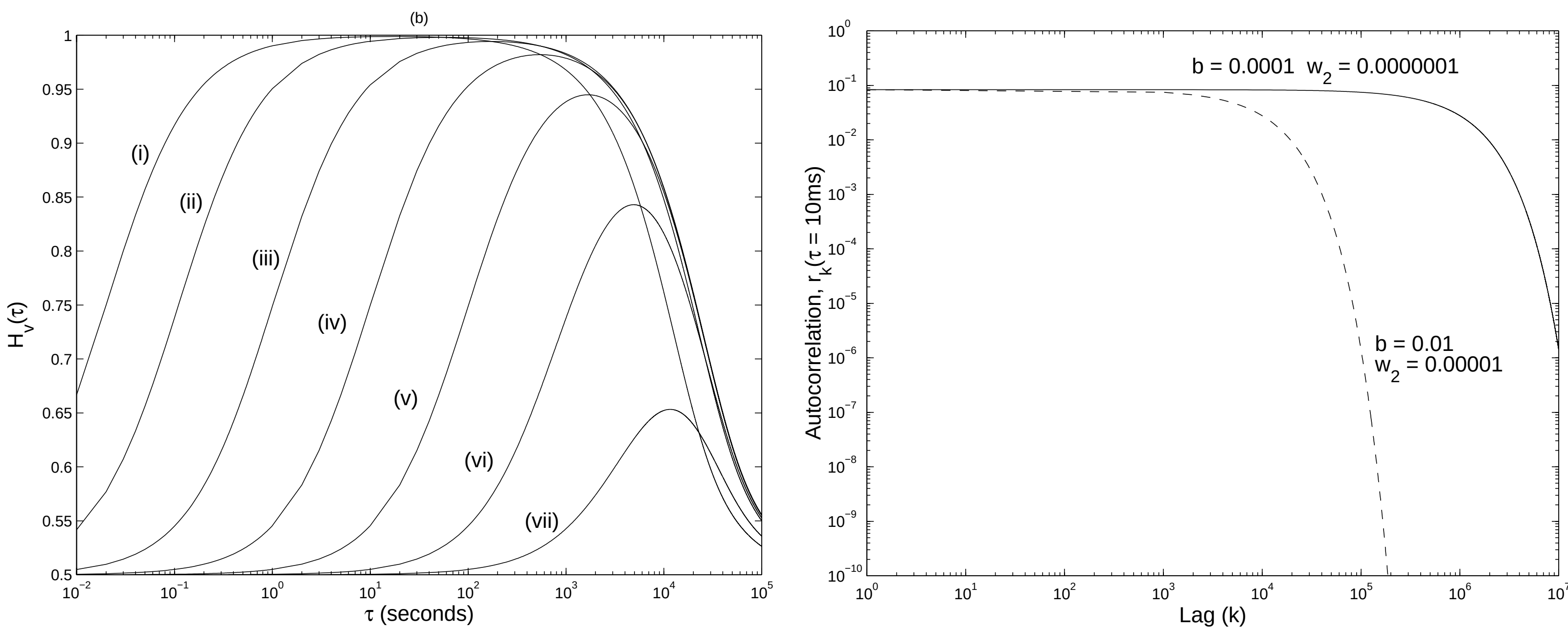
Example: Hyperexponential Distribution of Order Two (II)





a= 100 b= 0.0001

Example: Hyperexponential Distribution of Order Two (III)



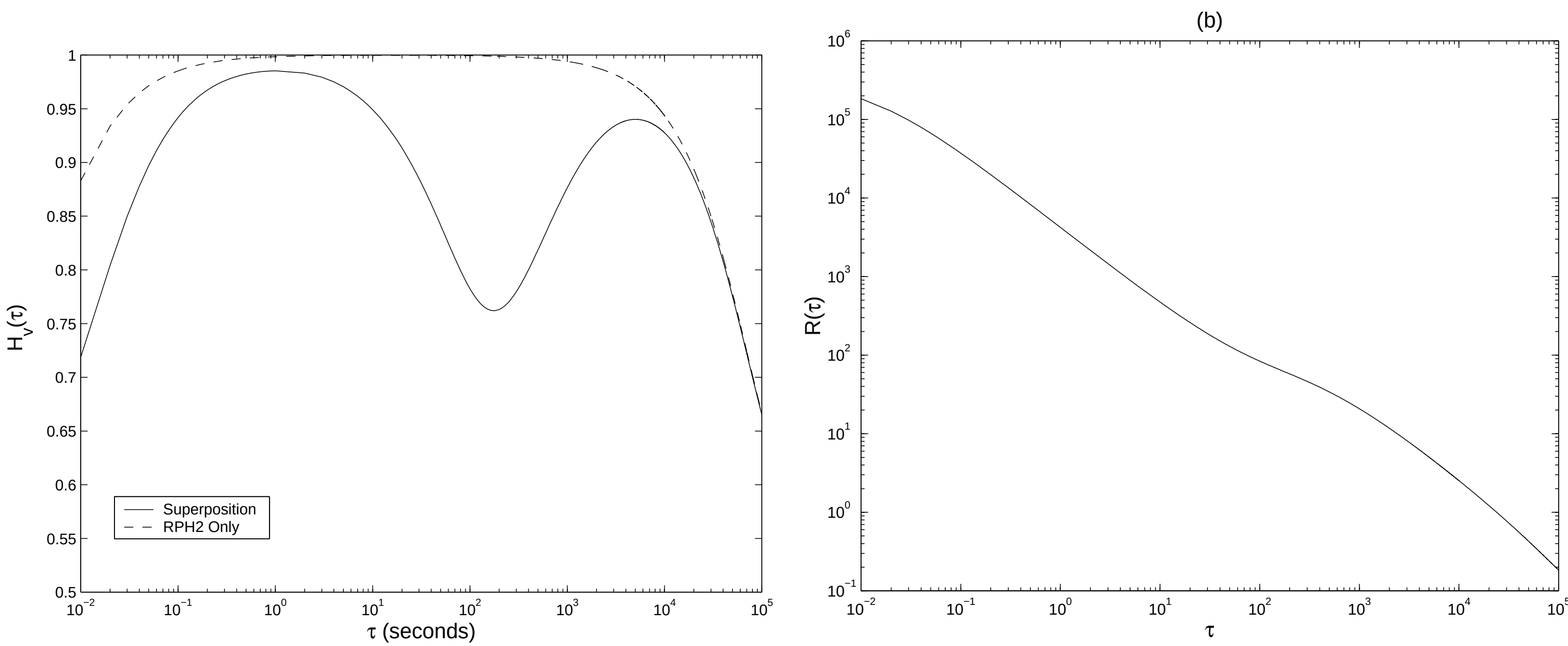
a= 100 b= 0.0001 w_2 = (i) 10^{-6} (ii) 10^{-7} (iii) 10^{-8} (iv) 10^{-9} (v) 10^{-10} (vi) 10^{-11} (vii) 10^{-12}

Example: Superposition of Heterogeneous Traffic Processes (I)

Suppose the underlying point-process of the packet (byte) count sequence Y is the superposition of:

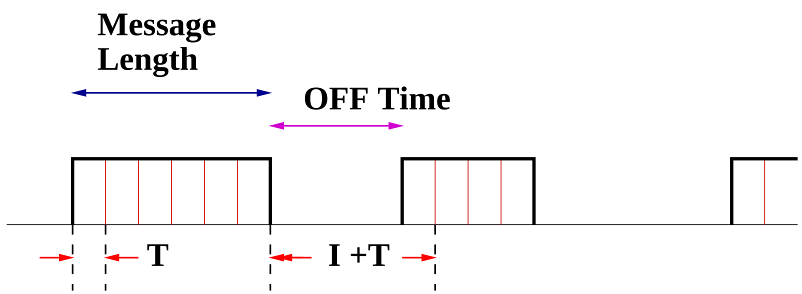
- 10 renewal processes with interarrival times hyperexponentially distributed of order two (RPH2)
- 20 two-state Markov Modulated Poisson processes (MMPP)
- 16 packetized voice streams
- 40 packet streams generated by ON/OFF traffic sources whose ON and OFF periods are both exponentially distributed
  - $\mathcal{C}^2(X) = 5.6323\text{x}10^5 \Rightarrow Y$ is not Poisson
  - Y is SRD

Example: Superposition of Heterogeneous Traffic Processes (II)



- Y has high variability over a range of time scales that spans 8 order of magnitude
- the amount of correlation that the process Y has at a particular time scale does not alone determine the degree of its variability at that time scale

ON/OFF/Exponential Model ⇒ Renewal Process



- W : number of packets during ON period: geometrically distributed
  - packet stream: renewal process
- Pdf: $f(x) = p\delta(x-T) + (1-p)\beta e^{-\beta(x-T)}u(x-T)$
  - β^{-1} : mean OFF period
  - T : packet transmission time
  - $p = \frac{E[W]-1}{E[W]}$: probability that the next interarrival time is T
  - $1-p$: probability that the next interarrival time is $I+T$
  - $\lambda = \frac{\beta}{(1-p)+\beta T}$: mean packet arrival rate

ON/OFF/Exponential Model : Exact Analysis

$$Var[N(\tau)] = 2\lambda \sum_{n=0}^{\infty} p^n (\tau - nT) u(\tau - nT) + 2\lambda \sum_{n=1}^{\infty} \sum_{z=1}^{n} \binom{n}{z} \frac{p^{n-z}(1-p)^z}{\beta} \{\beta(\tau - nT) G[\beta(\tau - nT), z] - zG[\beta(\tau - nT), z+1]\} u(\tau - nT) - \lambda\tau - (\lambda\tau)^2$$

$$\frac{d}{d\tau}(Var[N(\tau)]) = 2\lambda \sum_{n=0}^{\infty} p^n u(\tau - nT) + 2\lambda \sum_{n=1}^{\infty} \sum_{z=1}^{n} \binom{n}{z} p^{n-z} (1-p)^z G[\beta(\tau - nT), z] u(\tau - nT) - \lambda - 2\lambda^2 \tau$$

- $G(x,y) = \frac{1}{\Gamma(y)} \int_0^x t^{y-1} e^{-t}\, dt \qquad y > 0, \quad x > 0$: incomplete Gamma function
- $\lim_{k\to\infty} IDC(k\tau) = \mathcal{C}^2(X) = \lambda^2 \left(\frac{1-p^2}{\beta^2}\right)$
- $R(\tau) = \frac{1}{2}\left(\frac{\lambda^2(1-p^2)}{\beta^2 IDC(\tau)} - 1\right)$: Y is SRD

ON/OFF/Exponential Model : Fluid Analysis

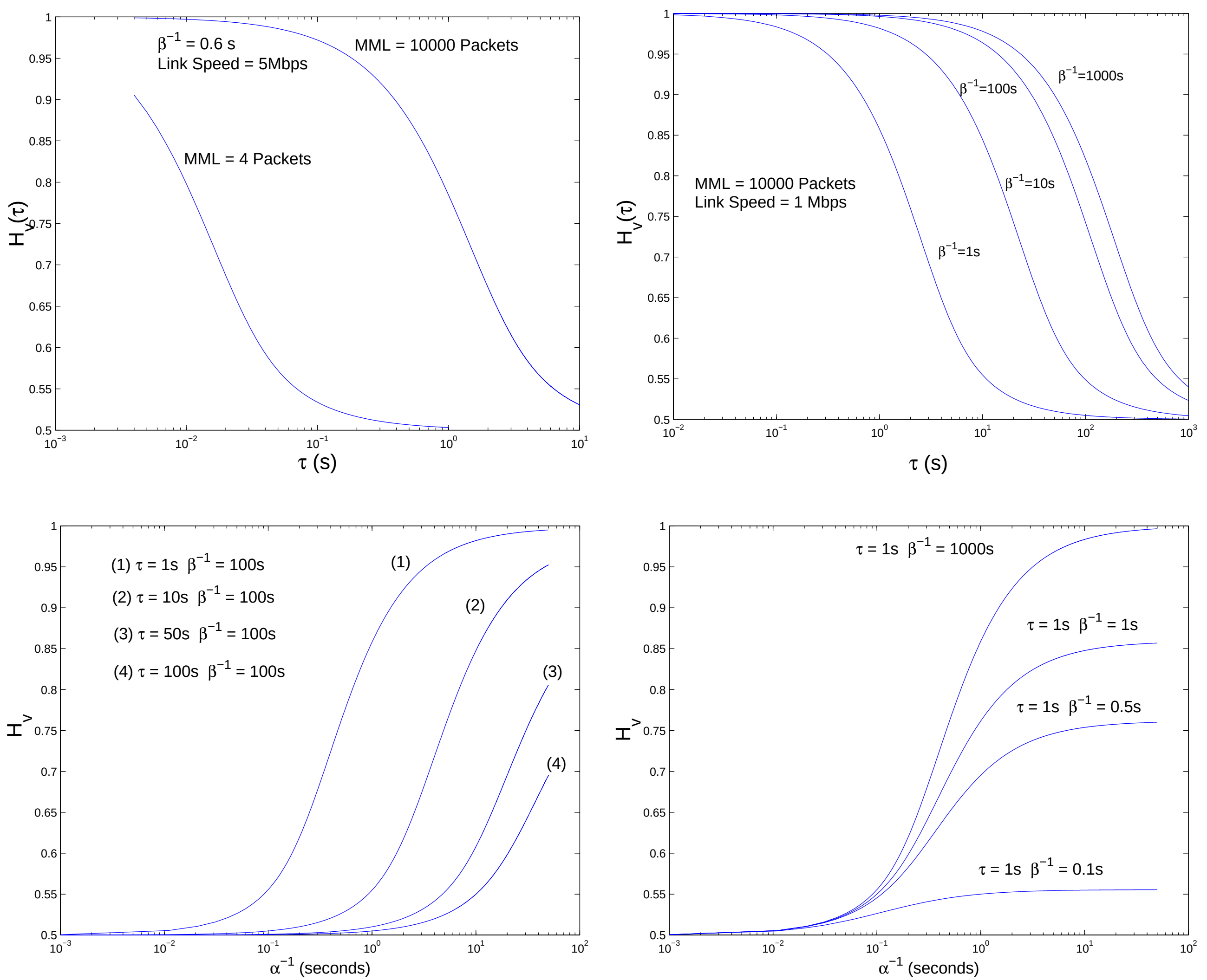
- $\alpha^{-1} = E[W]T$: mean ON period $\quad \rho = \alpha + \beta$

$$\tilde{V}ar[N(\tau)] = \frac{2(1-p)\lambda^3}{\beta^2}\left[\tau - \frac{1}{\rho}\left(1 - e^{-\rho\tau}\right)\right]$$

$$\frac{d}{d\tau}\left(\tilde{V}ar[N(\tau)]\right) = \frac{2(1-p)\lambda^3}{\beta^2}\left[1 - e^{-\rho\tau}\right]$$

- Enormous gain in computational speed
- $\lim_{\tau\to\infty} \frac{\tilde{V}ar[N(\tau)]}{Var[N(\tau)]} = \frac{2}{1+p}$
- $\lim_{\tau\to\infty} I\tilde{D}C(\tau) = \frac{2}{1+p}\,\mathcal{C}^2(X)$

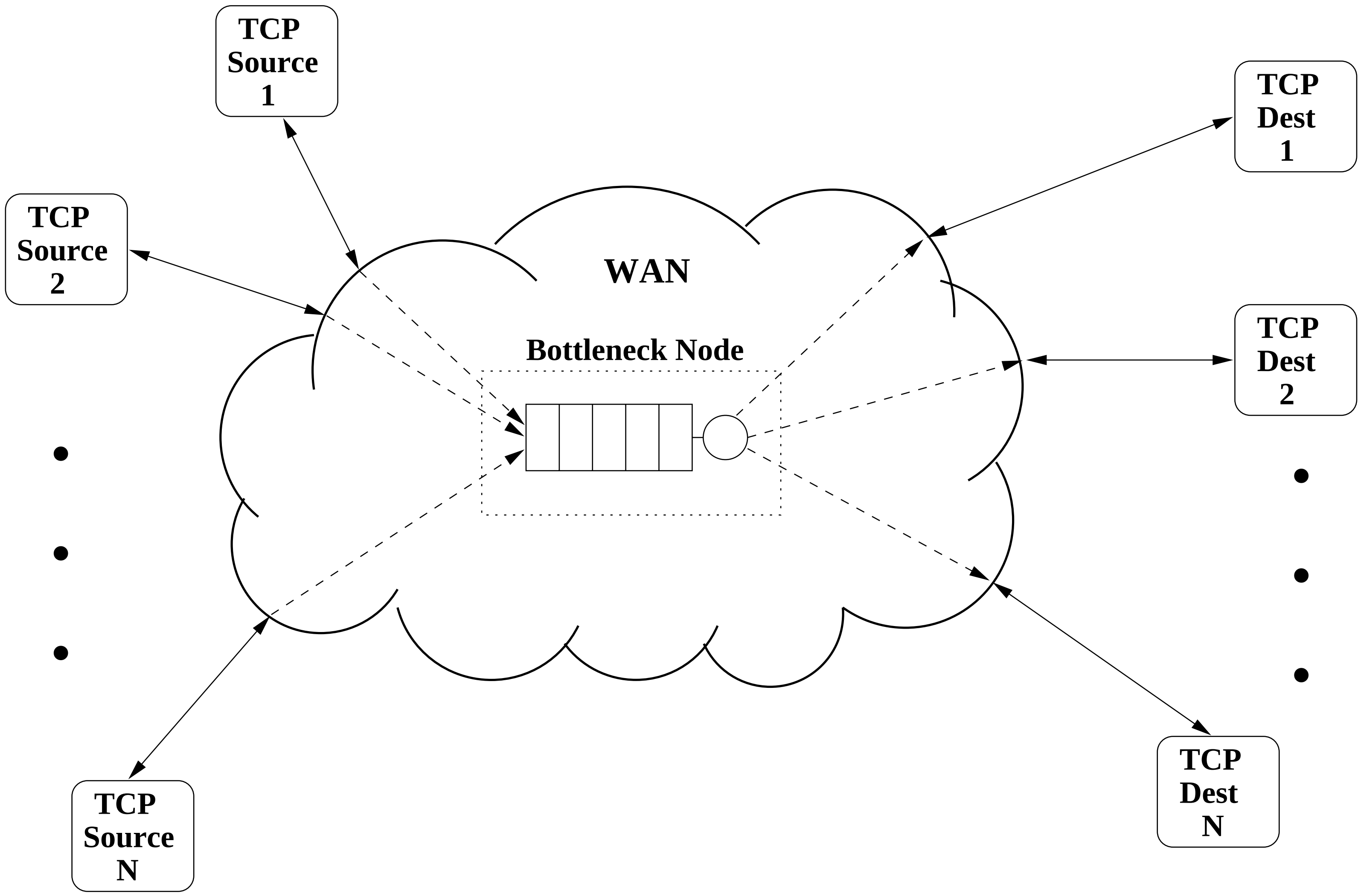
ON/OFF/Exponential Model : Index of Variability



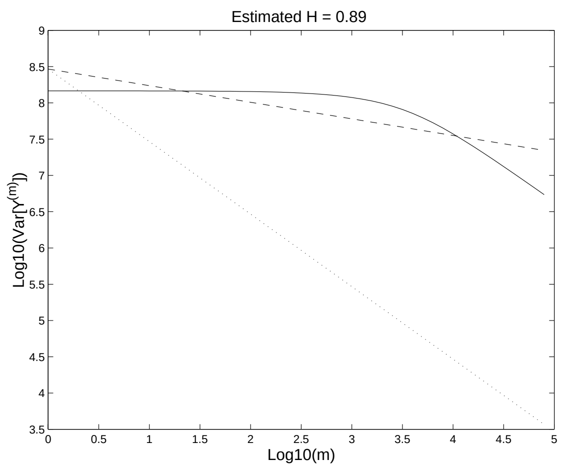
Simulation Study

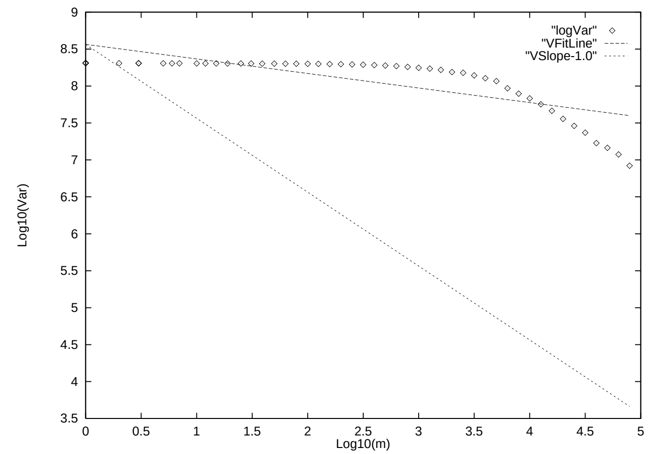
- Main Goals:
  - validate or invalidate our assumption that the primary factor contributing to high variability empirically observed in TCP traffic is the dynamics of TCP
  - validate the theory

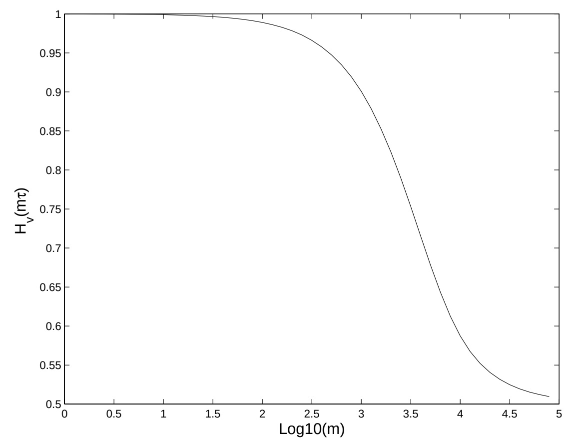
Simulation Study: Network Model



Simulation Study: ON/OFF/Exponential Model (I)

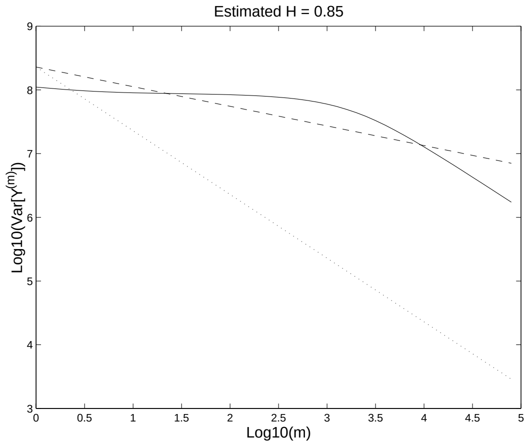


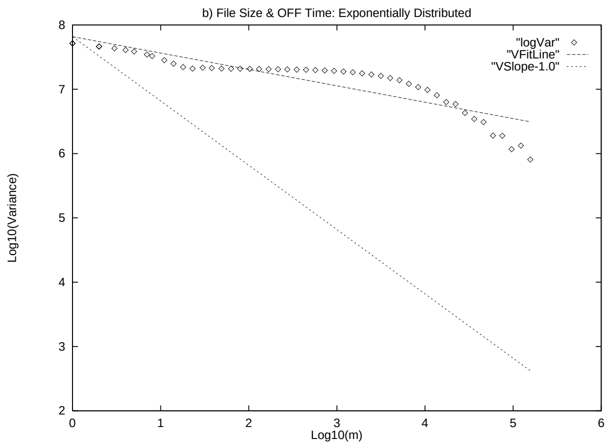


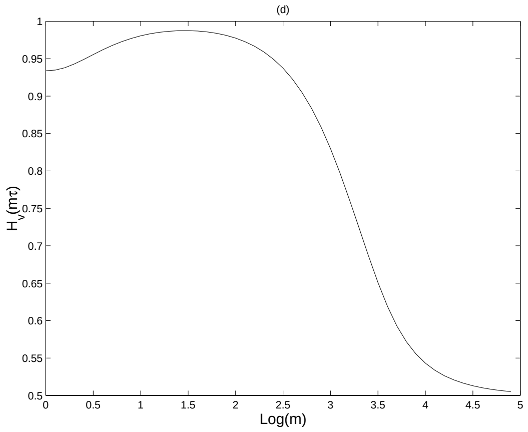


$E[W]$= 10000 pkts β^{-1}= 240s Link Speed=5Mbps Flows=64 TCP RTT=200-250ms

Simulation Study: ON/OFF/Exponential Model (II)



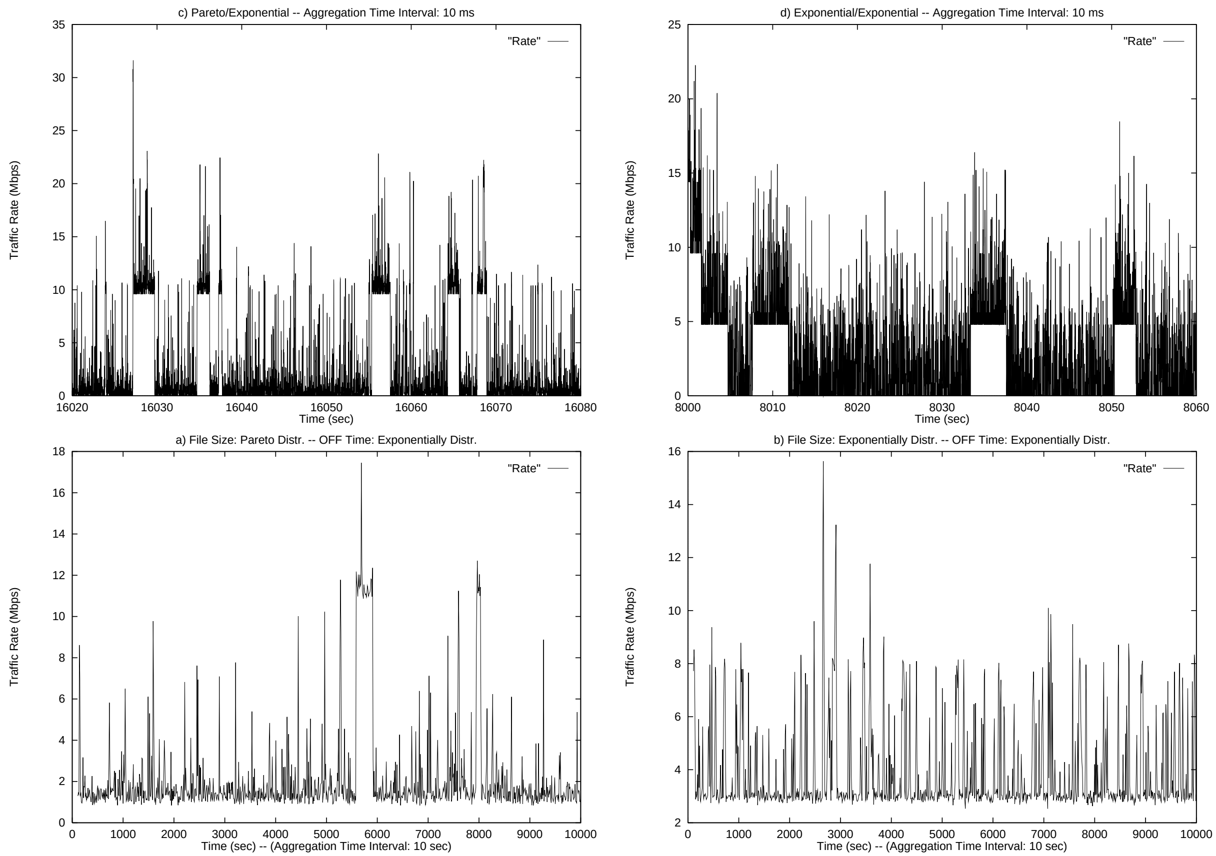




Flows=64 : $E[W]$= 4 pkts β^{-1}= 0.6s Link Speed=10Mbps TCP RTT=200-250ms

Flows=64 : $E[W]$= 10000 pkts β^{-1}= 900s

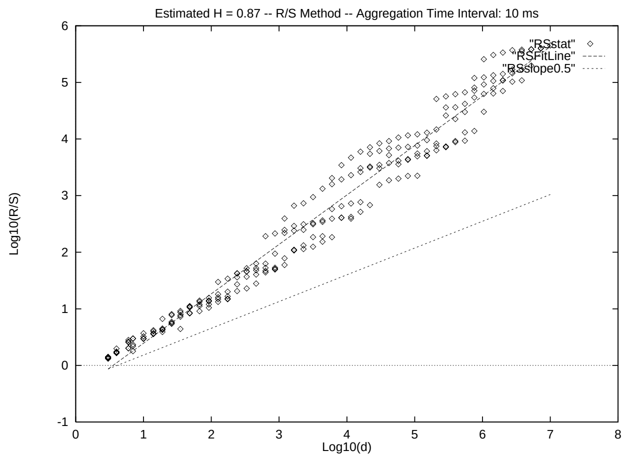
ON/OFF/Heavy-Tailed Model vs. ON/OFF/Exponential Model

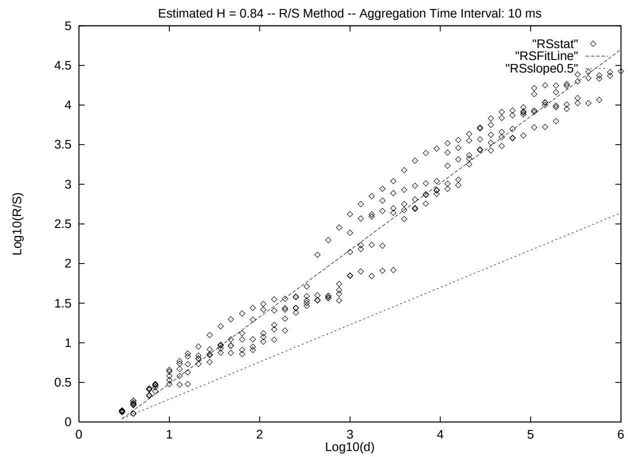


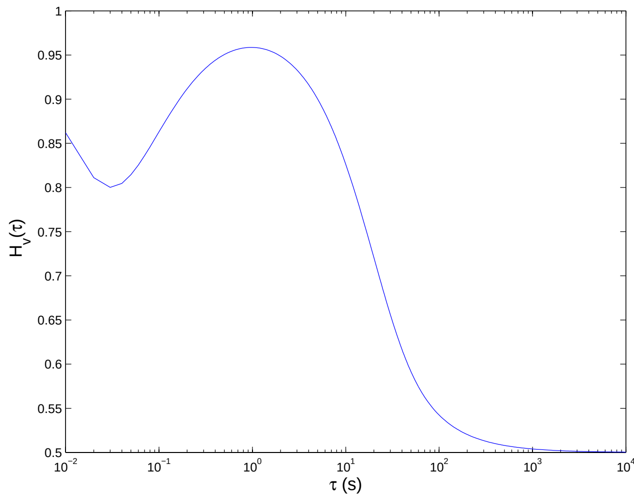
Top: Time Scale 10ms Bottom: Time Scale 10s – UDP case Flows=64

Left: ON/OFF/Heavy-Tailed Right: ON/OFF/Exponential

ON/OFF/Heavy-Tailed Model vs. ON/OFF/Exponential Model

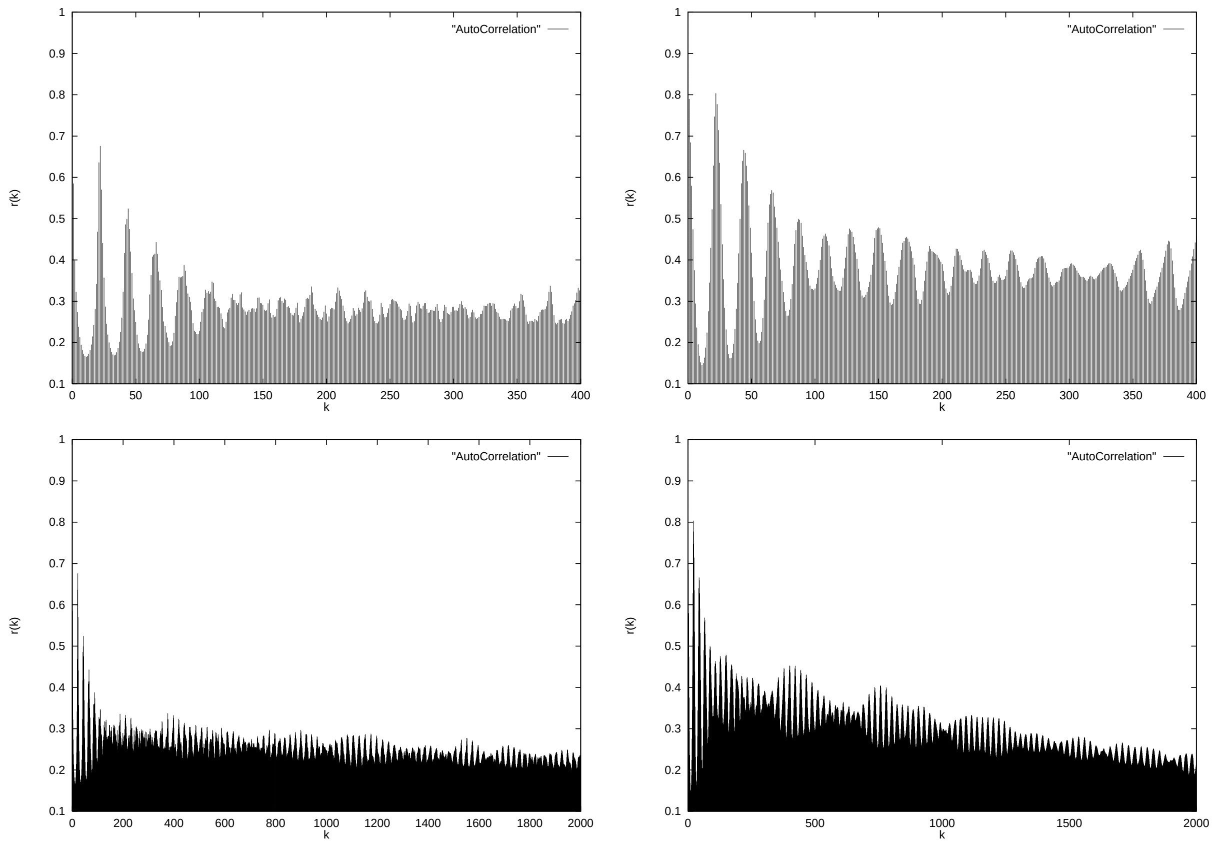






$\hat{H}$ = slope Heavy-Tailed: $\hat{H}$= 0.87 Exponential: $\hat{H}$= 0.84 $H_v(10ms)$=$H_v(10s)$= 0.86

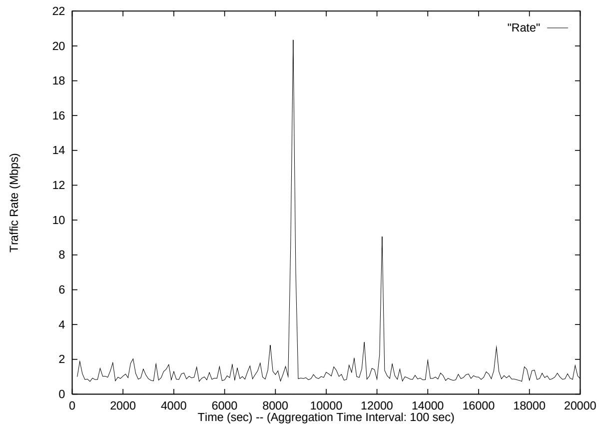
ON/OFF/Heavy-Tailed Model vs. ON/OFF/Exponential Model

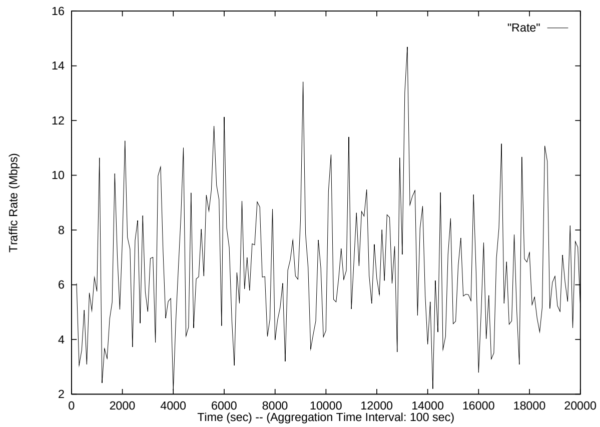


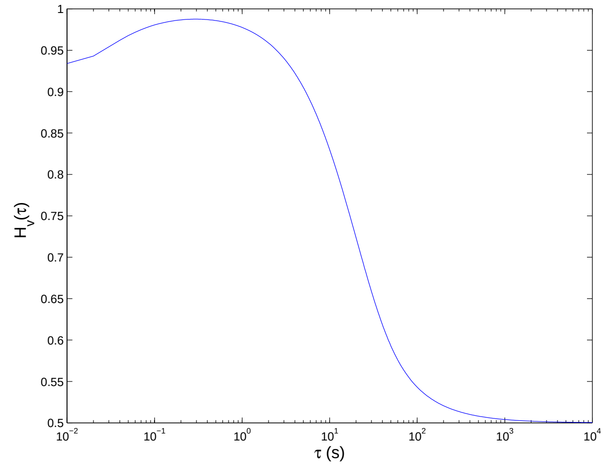
Left: ON/OFF/Heavy-Tailed TCP Win$= 1MB$ BDP$= 3.2MB$

Right: ON/OFF/Exponential Model TCP Win$= 64KB$ BDP$= 244KB$ τ=10ms

ON/OFF/Heavy-Tailed Model vs. ON/OFF/Exponential Model







Time Scale = 100s $H_v(100s) = 0.55$ TCP traffic

Simulation Study: Connections With Greedy Sources

- Goal: Determine if the dynamics of TCP alone can cause high variability over a wide range of time scales in traffic
  - all application-level factors that might contribute to the variability of traffic were eliminated
- Simulation experiments were conducted for the cases of
  - no packet losses
    $\Rightarrow$ several with uniformly distributed RTT (300ms to 600ms)
  - packet losses due to queue overflows
  - random packet losses
- Resulting aggregate TCP traffic
  - did not have LRD
  - had considerable variability only at short time scales (< 1s)

Conclusions (I)

- Constructed a new and theoretical practical framework for characterizing network traffic at all time scales based on the statistical properties of the underlying point processes
  - novel measure of variability: index of variability $H_v(\tau)$
    - $\Rightarrow$ captures degree of burstiness at each time scale
    - $\Rightarrow$ completely characterized by $Var[N(\tau)]$ or $IDC(\tau)$
  - new and straightforward way of calculating the autocovariance for all lags and all time scales
  - new and practical way for computing the infinity sum of the autocorrelation function for each time scale

Conclusions (II)

- Results from analyzing several traffic models show that
  - conventional traffic models can capture the high variability empirically observed in network traffic over a considerable range of time scales
  - $H_v(\tau)$ is a better measure for capturing the burstiness of network traffic than the Hurst parameter
  - the amount of correlation that a traffic process has at a particular time scale does not alone determine the degree of variability at that time scale
  - the mean file size, the mean OFF period, and the source link speed have a great impact on the variability of traffic generated by ON/OFF/Exponential sources

Conclusions (III)

- Results from analyzing TCP/UDP traffic streams suggest that
  - the dynamics of TCP alone can not cause high variability over a considerable range of time scales
  - the presence of high variability over a wide range of time scales does not necessarily depend on whether a reliable, flow- and congestion-controlled protocol is employed at transport layer
  - without prior knowledge about the traffic process

  $\Rightarrow$ we can not conclude based alone on the estimated value of the Hurst parameter that an empirically collected finite traffic sequence exhibits LRD

  $\Rightarrow$ the number of samples that might be required to get a good estimated of the Hurst parameter can be extremely large

Future Work

- Find a relation that associates $H_v(\tau)$ with queueing performance metrics (packet loss rate and delay)
  - expect different queueing behavior for each different $H_v(\tau)$ curve over all performance relevant time scales
- Construct a methodology of how to estimate $H_v(\tau)$ from empirically measured traffic traces
  - will help to develop accurate traffic models and traffic control mechanisms
- Analyze the ON/OFF/Hyperexponential traffic model
  - compromise between the ON/OFF/Heavy-Tailed and ON/OFF/Exponential models
- Obtain $H_v(\tau)$ for several stochastic processes thad have LRD
  - support our claim that $H_v(\tau)$ is a better measure than H