On Generator of Network Arrivals with Self-Similar Nature

Kai-Lung Hua

Advisor: Prof. Po-Ning Chen

Department of Communications Engineering
National Chiao Tung University, Taiwan 30050, R.O.C.

June, 2002

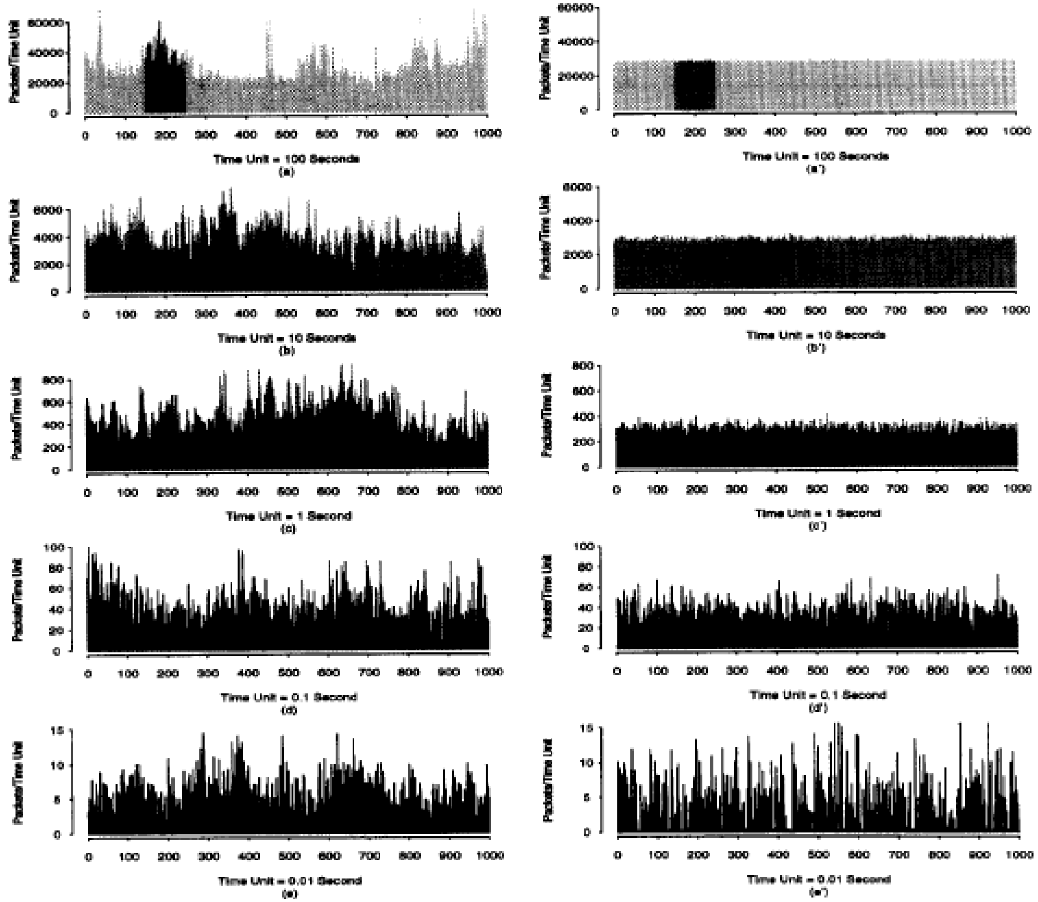
OUTLINE

- Introduction
- Self-Similar Process and Analysis Tests
- Scheme of Self-Similar Traffic Generator
- Simulation of Traffic Traces
- Conclusion and Future Work

Introduction

Background

- Recent empirical studies have shown that the modern computer network traffic is more appropriated modeled by *long range dependent (LRD) self-similar processes* than traditional short range dependent processes as Poisson process.
- Incorrect assessments of network system may be obtained, if LRD nature is not considered for the experimental synthetic network traffic.



Measured traffic data vs. Markovian-model-based sequences.
(Source: Leland 1994)

Self-Similar Process and Analysis Tests

Terminologies

- m-**averaged process** $\boldsymbol{X}^{(m)} = \left(X_1^{(m)}, X_2^{(m)}, \cdots\right)$ of $\boldsymbol{X} = (X_1, X_2, X_3, \ldots)$

$$X_i^{(m)} = \frac{1}{m}\sum_{j=1}^{m} X_{m(i-1)+j} = \frac{X_{m(i-1)+1} + X_{m(i-1)+2} + \cdots + X_{mi}}{m}$$

- **Autocovariance function** of the m-averaged process $\boldsymbol{X}^{(m)}$

$$C_m(k) = \mathrm{Cov}\left\{X_i^{(m)}, X_{i+k}^{(m)}\right\}$$

- **Autocorrelation coefficient function** of the m-averaged process $\boldsymbol{X}^{(m)}$

$$\rho_m(k) = C_m(k)/C_m(0)$$

Definition 1 *A second-order stationary process* $\boldsymbol{X}$ *is* **exactly second-order self-similar** *with parameter* H*, where* $0.5 < H < 1$*, if either of the following conditions holds:*

- $$\rho_1(k) = \frac{1}{2}[|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}] \text{ for } k \in \{0, 1, 2, \ldots\}$$

- $$C_m(k) = C_1(k) m^{2H-2} \text{ for } k \in \{0, 1, 2, \ldots\} \text{ and } m \in \{1, 2, 3, \ldots\}$$

Definition 2 *A second-order stationary process* $\boldsymbol{X}$ *is* **asymptotically second-order self-similar** *with parameter* H*, where* $0.5 < H < 1$*, if*

$$\lim_{k\to\infty} \rho_m(k) = \frac{1}{2}[|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}] \text{ for } k \in \{0, 1, 2, \ldots\} \text{ and } m \in \{1, 2, 3, \ldots\}$$

Statistics Test for Self-Similarity

How to determine H?

- Variance-Time Analysis
- R/S Analysis
- Periodogram Analysis

Variance-Time Analysis

Taking $k = 0$ in $C_m(k) = C_1(k)m^{2H-2}$ yields

$$\text{Var}\{X^{(m)}\} = \text{Var}\{X\}m^{2H-2},$$

or equivalently,

$$\log(\text{Var}\{X^{(m)}\}) = \log(\text{Var}\{X\}) + (2H-2)\log(m).$$

The self-similar parameter H can be determined by the slope $(2H-2)$ of straight line of $\log(\text{Var}\{X^{(m)}\})$ against $\log(m)$.

R/S Analysis

According to Hurst effect (cf. Section 2.2.5 in the Thesis),

$$E[R(n)/S(n)] \approx cn^{-H}$$

or equivalently,

$$\log\left([E[R(n)/S(n)]\right) \approx \log(c) - H\log(n),$$

where

$$\frac{R(n)}{S(n)} = \frac{\max\limits_{1\le k\le n}\left[\sum_{j=1}^{k} X_j - k\mu(n)\right] - \min\limits_{1\le k\le n}\left[\sum_{j=1}^{k} X_j - k\mu(n)\right]}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}[X_j - \mu(n)]^2}}.$$

Periodogram Analysis

According to $1/f$-noise (cf. Section 2.2.2 in the Thesis), the power spectral density of self-similar processes obeys

$$S(w) \approx L(1/w) w^{1-2H} \quad \text{at } w \approx 0.$$

where $L(\cdot)$ is a slowly varying function satisfying

$$\lim_{k \to \infty} L(kx)/L(k) = 1 \quad \text{for all } x > 0.$$

Hence,

$$\log[S(w)] \approx \log[L(1/w)] + (1 - 2H) \log(w).$$

Usually, only the lowest 10% of the entire frequency range, 2π, is used to determine the regression line slope.

Schemes of Self-Similar Traffic Generator

Existing Techniques for the Generation of Self-Similar Traffics

- **Time Domain Approach**
  Random Midpoint Displacement (Lau, Erramili, Wang and Willinger 1995)
- **Frequency Domain Approach**
  Spectrum fitting and IFFT (Paxson 1995)

Random Midpoint Displacement

- **Input**
  Desired Hurst parameter H
  Length of self-similar trace $[0,T)$ and sampling coefficient n.
- **Output**
  An approximated Fractional Brownian Motion (FBM) sequence of duration T with sampling period $T/2^n$.

Note: FBM is a continuous self-similar zero-mean Gaussain process with autocorrelation function $R(s,t) = (1/2)(s^{2H} + t^{2H} - |s-t|^{2H})$.

RMD Principle behind

If $Z(t)$ is a FBM, Z_{RMD} is independent of $Z(b) - Z(a)$, and is zero-mean Gaussian distributed.

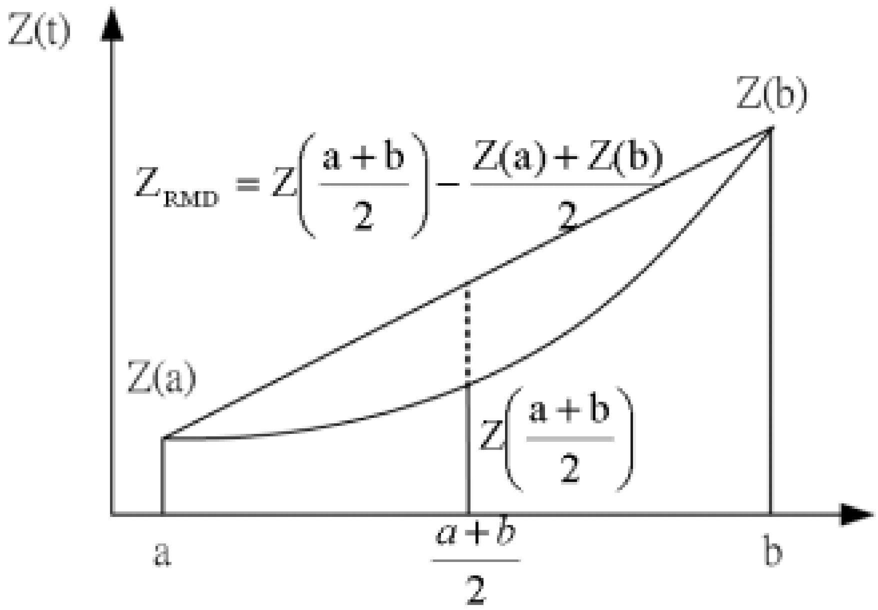
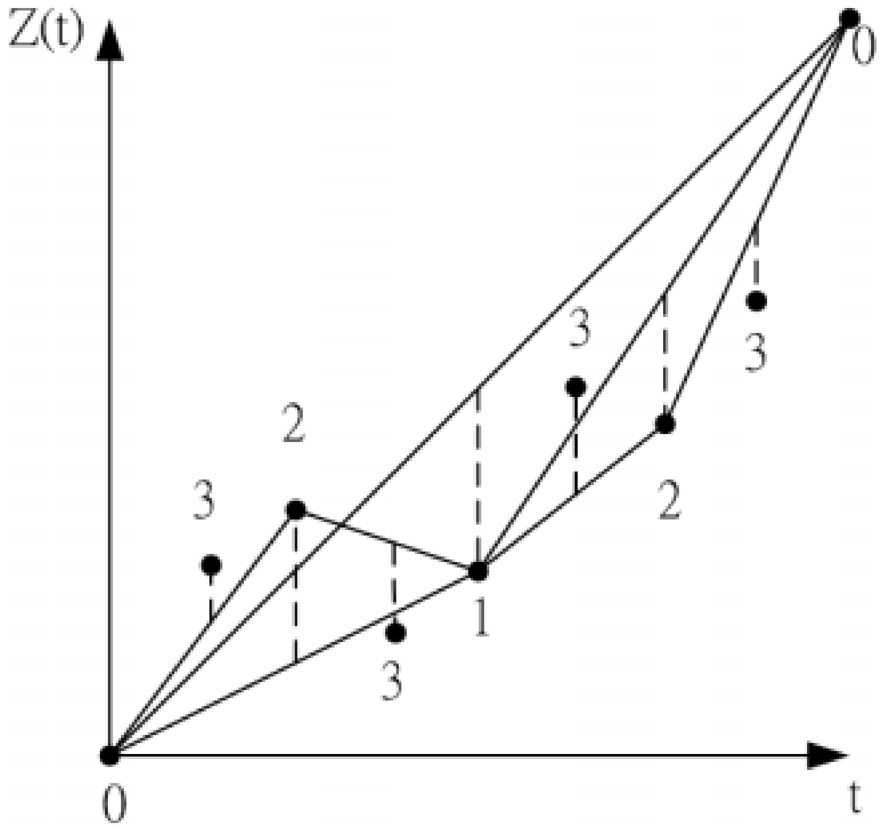


Illustration of the first three iterations of the RMD algorithm.



Random Midpoint Displacement

- After the generation of an FBM approximated sequence, the RMD algorithm approximates the accumulate arrival traffic $A(t)$ through

$$A(t) = M \cdot t + \sqrt{a \cdot M}\, Z(t).$$

- The number of packet per unit time is then approximated by:

$$A[t, t + \triangle t) = M \cdot \triangle t + \sqrt{aM}\left[Z(t + \triangle t) - Z(t)\right].$$

- Note:This is a three-parameter model depending on M (mean rate), a (peakness factor, ratio of variance to mean of the number of the packets per unit time) and H.

Random Midpoint Displacement

- **Advantages**
  - Simple, efficient and fast.
- **Drawbacks**
  - Not on the fly
  - May produce negative integer
  - The resultant H may be deviated from the target one.

Spectrum Fitting through IFFT

- The power spectrum of a (exactly second-order) self-similar process

$$S_y(w) = \sin(\pi H) \cdot \Gamma(2H+1) \cdot |1 - e^{-jw}|^2 \sum_{k=-\infty}^{\infty} |w + 2\pi k|^{-1-2H} \text{ for } -\pi \leq w < \pi.$$

- The *infinite* sum is then approximated by a *finite* sum.

$$\begin{aligned} \hat{S}_y(w) &= \sin(\pi H) \cdot \Gamma(2H+1) \cdot |1 - e^{-jw}|^2 \\ &\times \left(\sum_{k=-3}^{3} |w + 2\pi k|^{-1-2H} + \frac{1}{8H\pi} \sum_{k=\pm 3, \pm 4} |w + 2\pi k|^{-1-2H} \right). \end{aligned}$$

The algorithm is quoted as follows.

- **Sample** the positive-side (i.e., $w \in [0, \pi)$) of the approximated power spectrum with $n/2$ samples.
- **Multiply** each sample by an independent **exponentially distributed** random variable with mean 1.
  - It is known that the power spectrum for a frequency, when being estimated through the periodogram, is distributed asymptotically as an independent exponential random variable with mean equal to the actual power.
- **Square-root** all samples.
- **Multiply** each sample by an independent **phase** that is uniformly distributed over $[0, 2\pi)$.
- **Copy** the current samples to the other half of the spectrum (i.e., from $-\pi$ to 0) to ensure **symmetricity**.
- **Perform IFFT** on the resultant n samples.

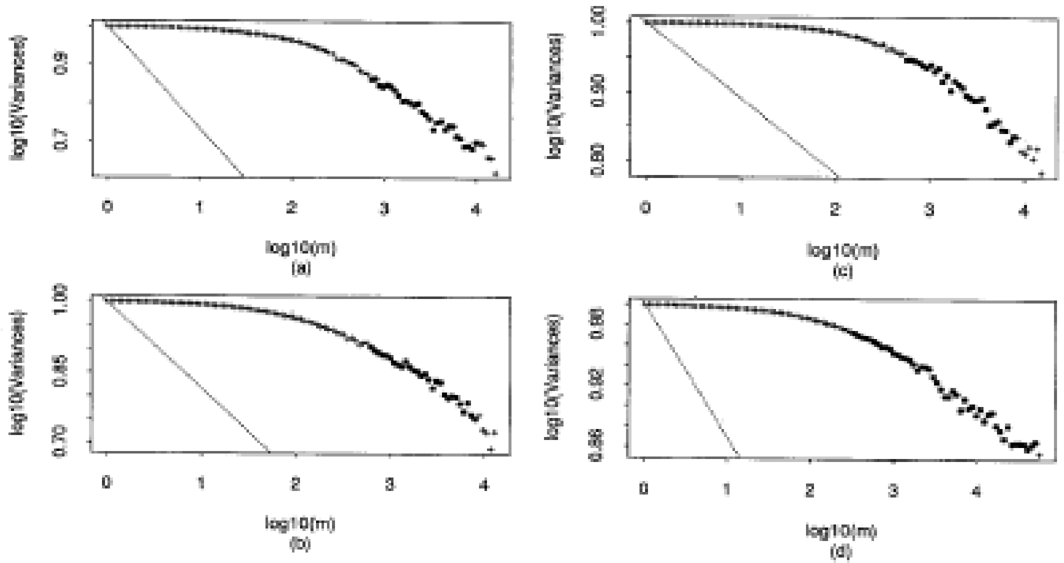
Spectrum Fitting through IFFT

- **Advantages**
  - Faster than, e.g., RMD algorithm (take about half time of the RMD algorithm)
- **Drawbacks**
  - Not on the fly
  - May produce negative integer
  - The resultant H may be slightly deviated from the target one.

Objective of the Research

A self-similar traffic generator that

- produces non-negative integers (representing packet train arrivals).
- generates self-similar traffic on the fly.
- fits the required self-similar Hurst parameter, H.
- its self-similar nature only lasts for a predetermined range, but disappears ultimately.
  - The self-similar nature of measured behavior of true network traffic only lasts beyond a practically manageable range, but disappears as the considered aggregated window is much further extended.

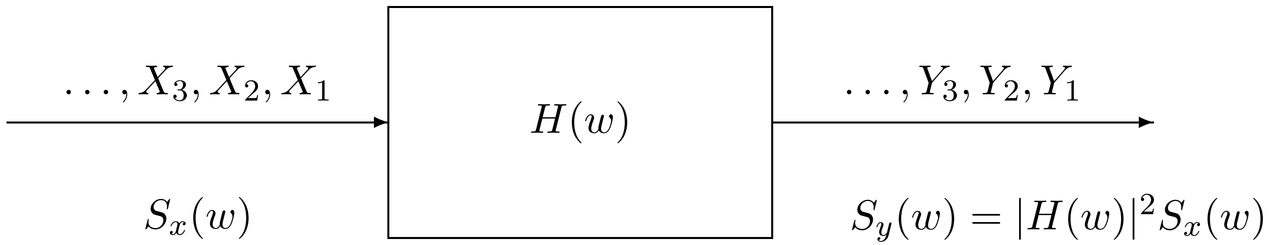


Measured behavior of true network traffic.
(Source: Beran 1995)

$$\mathrm{Var}\{X^{(m)}\} = \mathrm{Var}\{X\} m^{2H-2}$$

$$\log(\mathrm{Var}\{X^{(m)}\}) = \log(\mathrm{Var}\{X\}) + (2H-2)\log(m)$$

A Filter-based Self-Similar Traffic Generator



- Power spectrum of an exactly second-order self-similar process

$$S_y(w) = \sin(\pi H) \cdot \Gamma(2H+1) \cdot |1 - e^{-jw}|^2 \sum_{k=-\infty}^{\infty} |w + 2\pi k|^{-1-2H} \quad \text{for } -\pi \le w < \pi.$$

- There exists no close-form formulated filter $h[n]$ that gives the above filter output power spectrum due to i.i.d. input. Some simplification is necessary.

- Take the main term (k = 0), and approximate $|w|$ by $|1 - e^{-jw}| = 2|\sin(w/2)|$. $|w|$.

$$\tilde{S}_y(w) = |1 - e^{-jw}|^{1-2H} \quad \text{for } -\pi \le w < \pi,$$

where the coefficient, $\sin(\pi H) \cdot \Gamma(2H + 1)$, is removed for simplicity.

- It remains to design a filter whose output spectrum due to an i.i.d. input equals $\tilde{S}_y(w)$, or specifically, $|H(w)|^2 = |1 - e^{-jw}|^{1-2H}$.

- By Taylor's expansion,

$$(1-z)^{-(2H-1)/2} = \sum_{n=0}^{\infty} \frac{\Gamma(n + (2H-1)/2)}{\Gamma(n+1)\Gamma((2H-1)/2)} z^n,$$

where $\Gamma(\cdot)$ represents the Euler gamma function.

- Replacing z with e^{-jw} gives that

$$\left(1 - e^{-jw}\right)^{(1-2H)/2} = \sum_{n=0}^{\infty} \frac{\Gamma(n + H - 0.5)}{\Gamma(n+1)\Gamma(H-0.5)} e^{-jwn}.$$

- Term-wisely comparing with $H(w) = \sum_{n=0}^{\infty} h[n] e^{-jwn}$ concludes that

$$h[n] = \frac{\Gamma(n + H - 0.5)}{\Gamma(n+1)\Gamma(H-0.5)} \quad \text{for } n \geq 0.$$

Self-similarity of the approximated filter output

$$\tilde{S}_y(w) = |1 - e^{-jw}|^{1-2H} \quad \text{for } -\pi \le w < \pi,$$

- Does such an extensive spectrum simplification remove the self-similarity of the filter output process?
- It can be proved that

$$\lim_{m\to\infty} \frac{C_m(0)}{C(0)m^{2H-2}} = \frac{2\Gamma(2-2H)\sin(\pi H - \pi/2)}{H(2H-1)}.$$

  This implies that for m large, $\log[C_m(0)/C(0)]$ asymptotically behaves like $(2H-2)\log(m) + \log[2\Gamma(2-2H)\sin(\pi H - \pi/2)/(H(2H-1))]$.

- Hence, the resultant filter output process is **asymptotic** self-similar from the aspect of **variance-time analysis**.

$$\tilde{S}_y(w) = |1 - e^{-jw}|^{1-2H} \quad \text{for } -\pi \le w < \pi,$$

- Does such an extensive spectrum simplification remove the **short-term** (i.e., m-small) self-similarity of the filter output process?
- The short-term behavior can be examined by:

$$C_m(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \tilde{S}_y(w)\frac{\sin^2(mw/2)}{m^2\sin^2(w/2)}dw = \frac{2^{2-2H}}{\pi}\int_0^{\pi/2}\frac{\sin^2(mw)}{m^2\sin^{2H+1}(w)}dw,$$

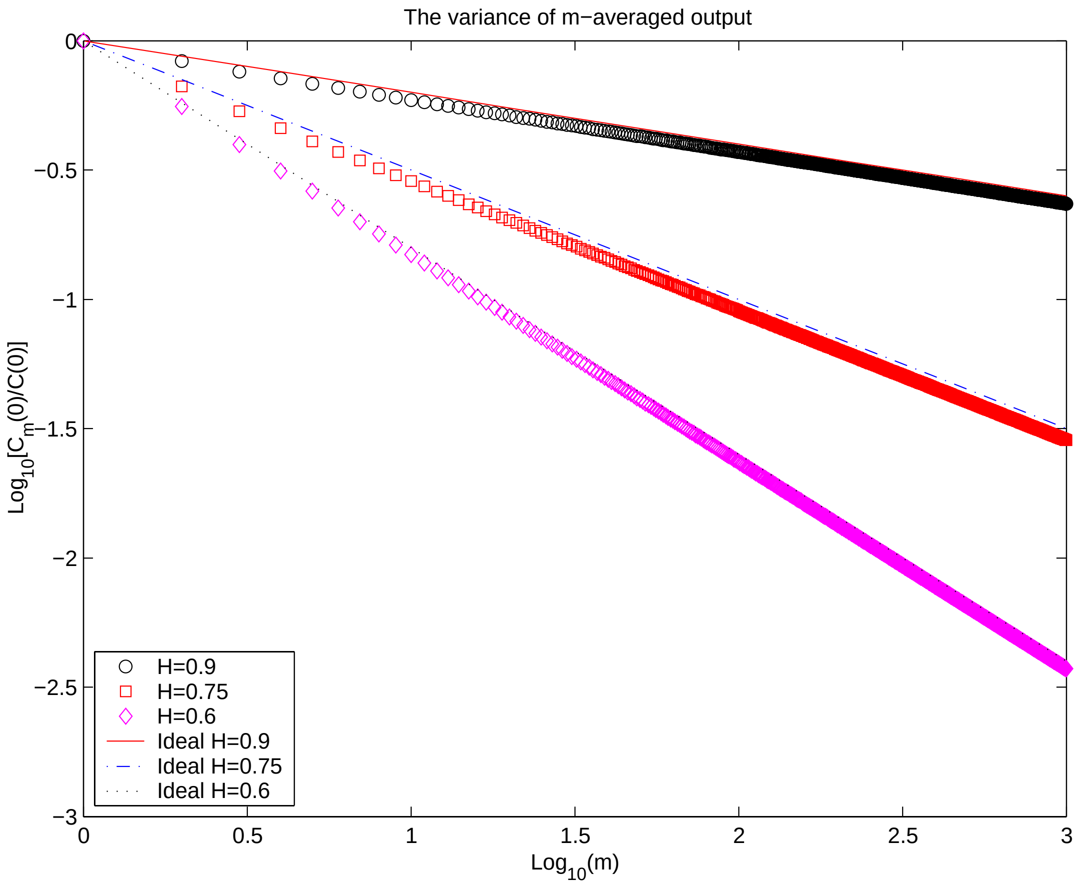
and

$$\log\frac{C_m(0)}{C(0)} = \log\frac{2\Gamma(1.5-H)}{\Gamma(1-H)\sqrt{\pi}} + \log\int_0^{\pi/2}\frac{\sin^2(mw)}{m^2\sin^{2H+1}(w)}dw.$$

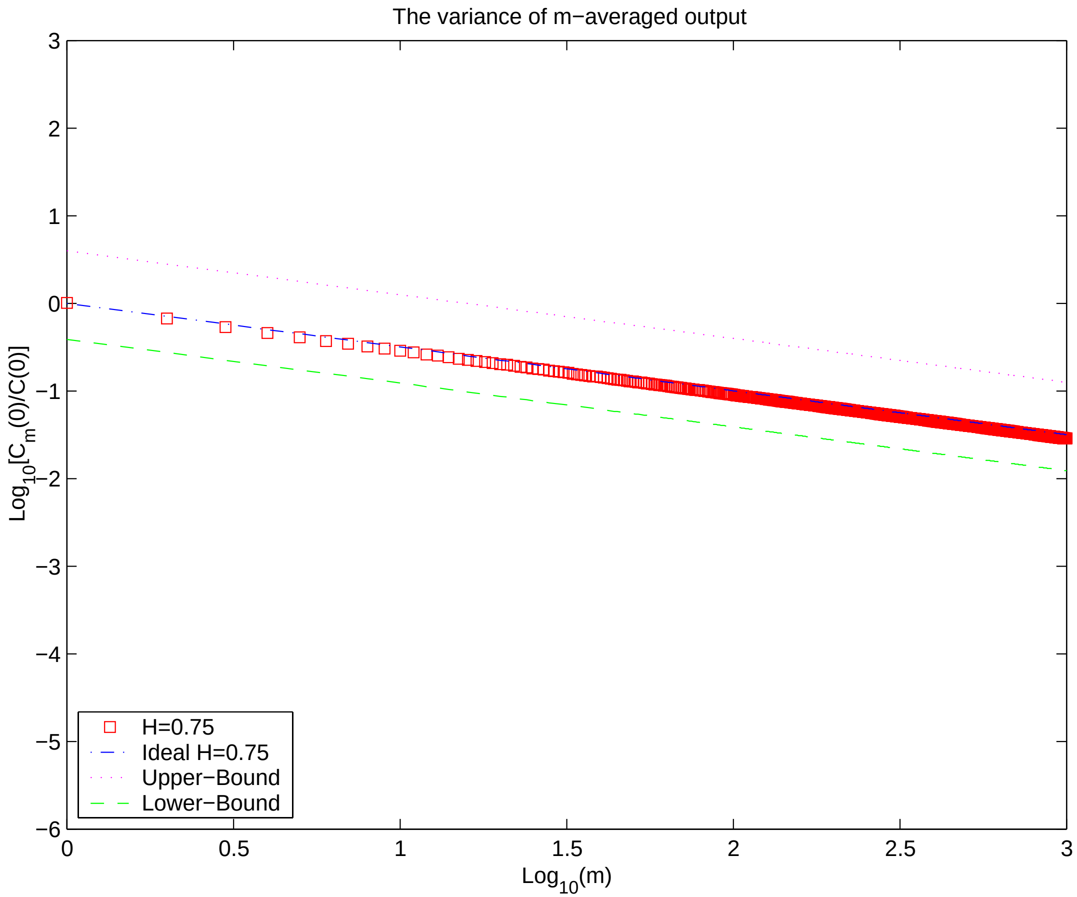
- It can be derived that

$$m^{2H-2}\frac{(2/\pi)^{2H}}{2(1-H)} \le \int_0^{\pi/2}\frac{\sin^2(mw)}{m^2\sin^{2H+1}(w)}dw \le m^{2H-2}\frac{\pi^3}{16H(1-H)(2H-1)}.$$

The exact variance-time plot of the filter output with simplified spectrum (drawn by Mathematica)

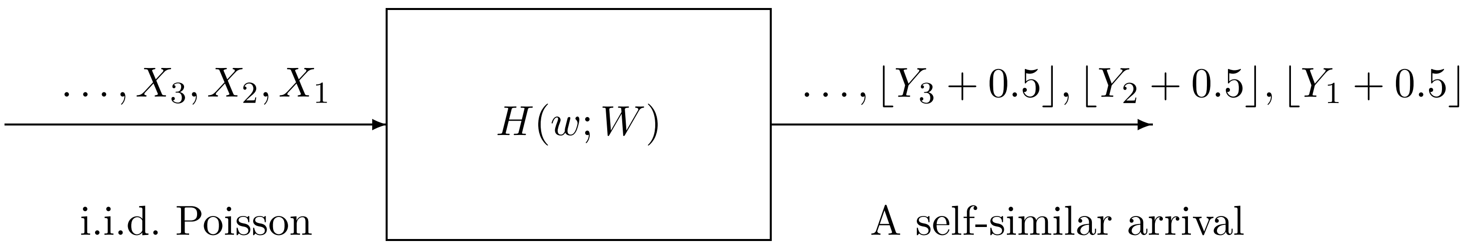


The upper and lower bounds of variance-time plot



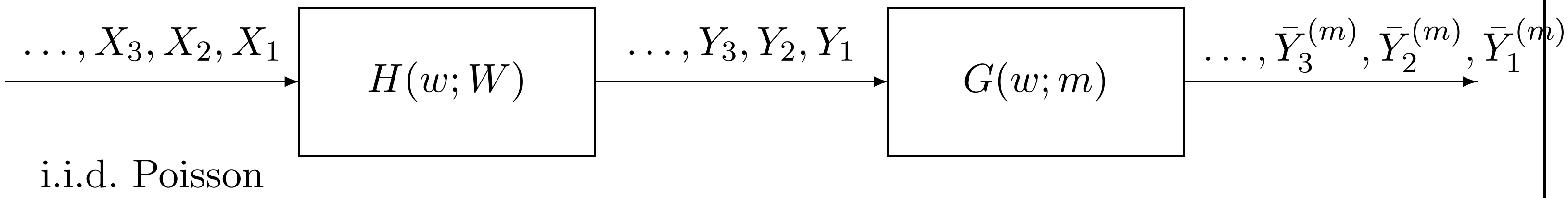
Truncation of the filter and rounding of the filter output

- The filter is of infeasibly infinite length, and must be **truncated** (with truncation window W).
- The filter output is real-valued, and hence must be **rounded** to the closest non-negative integers.



On Truncation Window Size

- Define $h[n;W] = h[n]$ for $0 \le n < W$, and 0, otherwise.
- Then $C_m(0;W)$ can be obtained through the introduction of filter $G(w;m)$, where $g[n;m] = 1/m$ for $0 \le n < m$ and 0, otherwise.



The **variance-equivalent** m**-averaged process** of the filter output process.

- Let $L(w) = H(w;W)G(w;m)$. Then $\ell[n] = h[n;W] * g[n;m]$, where "$*$" is the convolution operation. We then distinguish between two cases: $m \le W$ and $m > W$.

A) $m \le W$**.**

By letting $\tilde{S}_y(w;W)$ be the truncated counterpart of $\tilde{S}_y(w)$, we obtain:

$$\begin{aligned}
C_m(0;W) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\tilde{S}_y(w;W)dw \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}|L(w)|^2 dw \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\sum_{n=0}^{\infty}\ell[n]e^{-jnw}\right]\left[\sum_{n=0}^{\infty}\ell[n]e^{jnw}\right]dw \\
&= \sum_{n=0}^{\infty}|\ell[n]|^2 \\
&= \frac{1}{m^2}\left\{\sum_{l=0}^{W-m}\left(\sum_{n=l}^{l+m-1}h[n]\right)^2+\sum_{l=0}^{m-2}\left[\left(\sum_{n=0}^{l}h[n]\right)^2+\left(\sum_{n=W-1-l}^{W-1}h[n]\right)^2\right]\right\}.
\end{aligned}$$

Based upon the above formula, we numerically calculate

$$\log_{10}[C_m(0;W)] \quad \text{for} \;\; \log_{10}(m)=0,1,2,\ldots,\lfloor\log_{10}(W)\rfloor,$$

and find the straight line that best-fits these points.

Results: The slope is very close to $2H-2$.

B) $m > W$**.**

$$\begin{aligned} C_m(0;W) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \tilde{S}_y(w;W)dw \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |L(w)|^2 dw \\ &= \sum_{n=0}^{\infty} |\ell[n]|^2 \\ &= \frac{1}{m}F_H^2(W-1) + \frac{2}{m^2}\left\{\sum_{l=0}^{W-2} F_H^2(\ell) - F_H(W-1)\sum_{l=0}^{W-2} F_H(\ell)\right\} \\ &= \frac{1}{m}A_H(W) - \frac{1}{m^2}B_H(W) \end{aligned}$$

where $F_H(u) = \sum_{n=0}^{u} h[n]$ is the cumulative sum of $h[n]$,

$$A_H(W) = F_H^2(W-1) = \frac{W^2}{(H-0.5)^2}h^2[W]$$

and

$$B_H(W) = \frac{W^2h^2[W]}{(H-0.5)H(H+0.5)}\left(W + \frac{(H-1.5)}{2}\right) - \frac{(H-0.5)}{2H}\sum_{n=0}^{W-1} h^2[n].$$

Hence,

$$\frac{\partial \log[C_m(0;W)]}{\partial \log(m)} = -1 + \frac{B_H(W)/A_H(W)}{m - B_H(W)/A_H(W)},$$

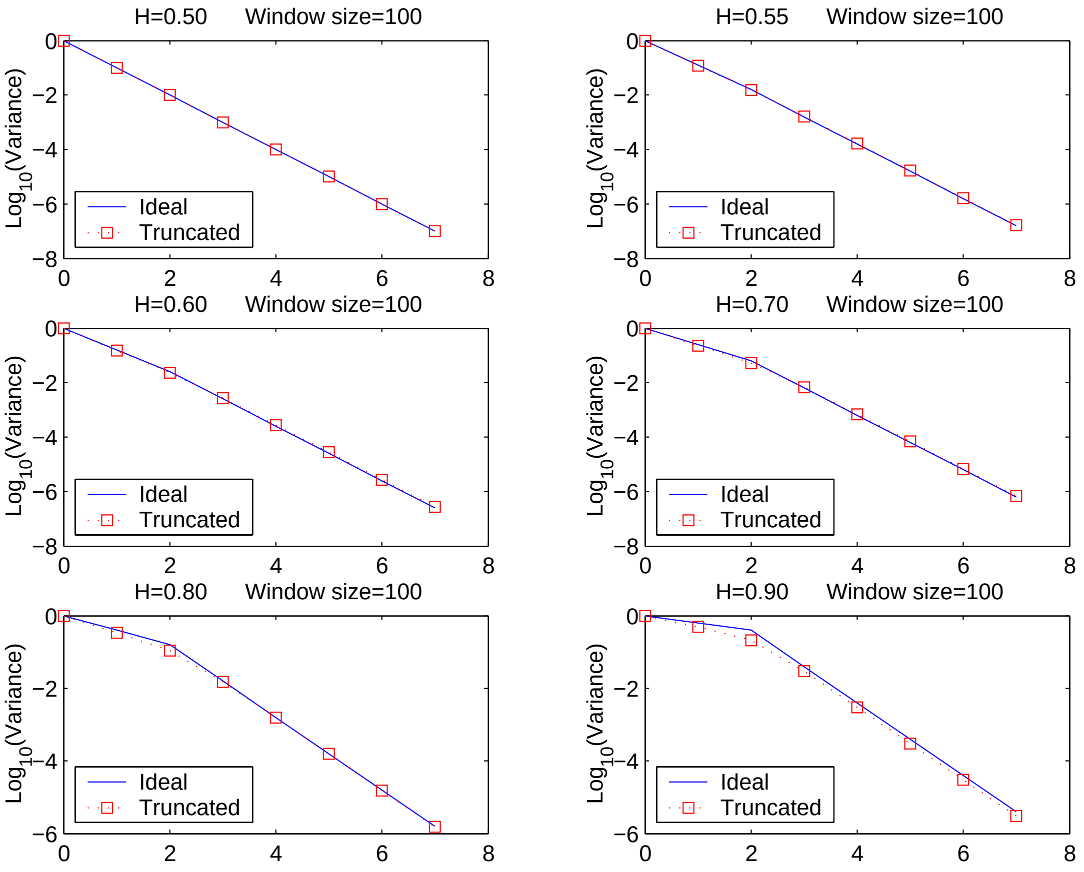
where

$$\begin{aligned}\frac{B_H(W)}{A_H(W)} &= \frac{(H-0.5)}{H(H+0.5)}\left(W + \frac{(H-1.5)}{2}\right) - \frac{(H-0.5)^3}{2HW^2 \cdot h^2[W]}\sum_{n=0}^{W-1} h^2[n] \\ &= \frac{(H-0.5)}{H(H+0.5)}\left(W + \frac{(H-1.5)}{2}\right) - O(W^{1-2H}).\end{aligned}$$

- A larger $B_H(W)/A_H(W)$ gives a larger $\dfrac{B_H(W)/A_H(W)}{m - B_H(W)/A_H(W)}$.
- $\dfrac{(H-0.5)}{H(H+0.5)}\left(W + \dfrac{(H-1.5)}{2}\right)$ peaks at $H = 1$.
- So taking $H = 1$ and $m = 10W$ yields $\dfrac{B_H(W)/A_H(W)}{m - B_H(W)/A_H(W)} \approx \dfrac{4W-1}{116W+1}$.

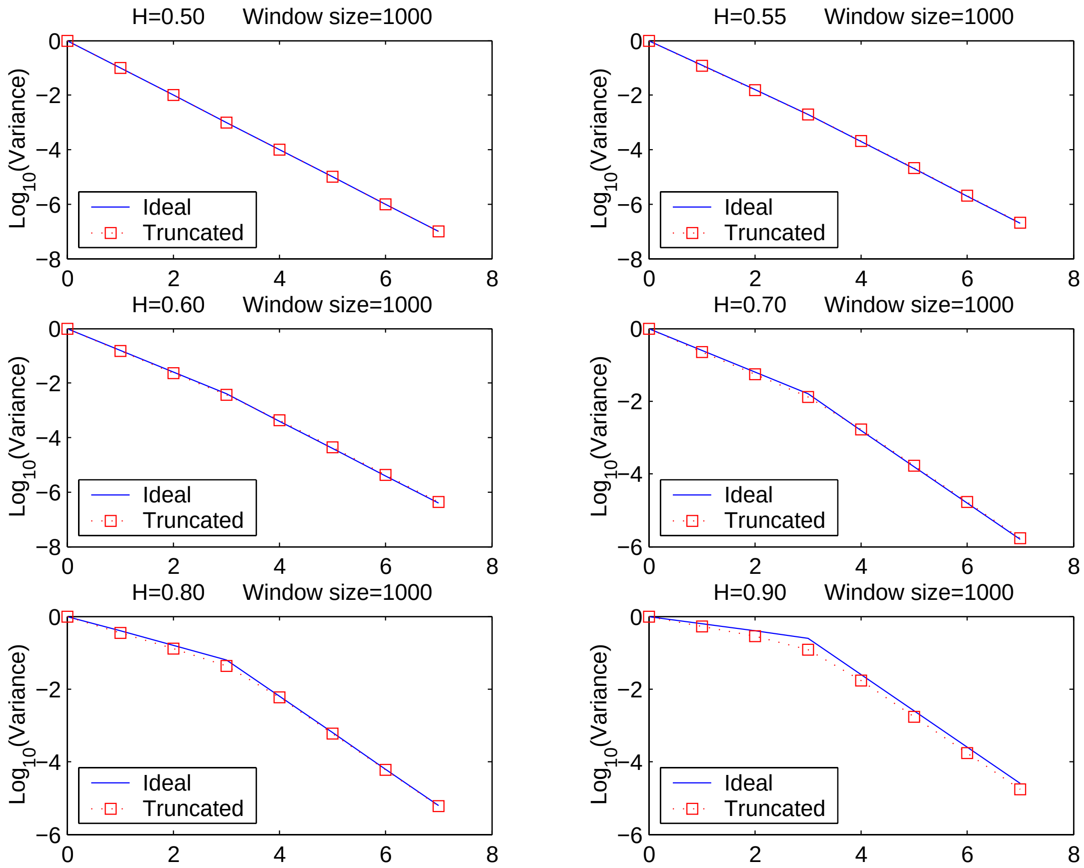
Result: The slope of the variance-log plot of the truncated filter output is close to -1 for m beyond $10W$.

- Variance-time analysis ($\log_{10}$ scale) for the rounded truncated-filter output with truncation window $W = 10^2$.
- The slope of the (blue) solid line is equal to $2H - 2$ for $m \leq W$, and -1 for $m > W$.



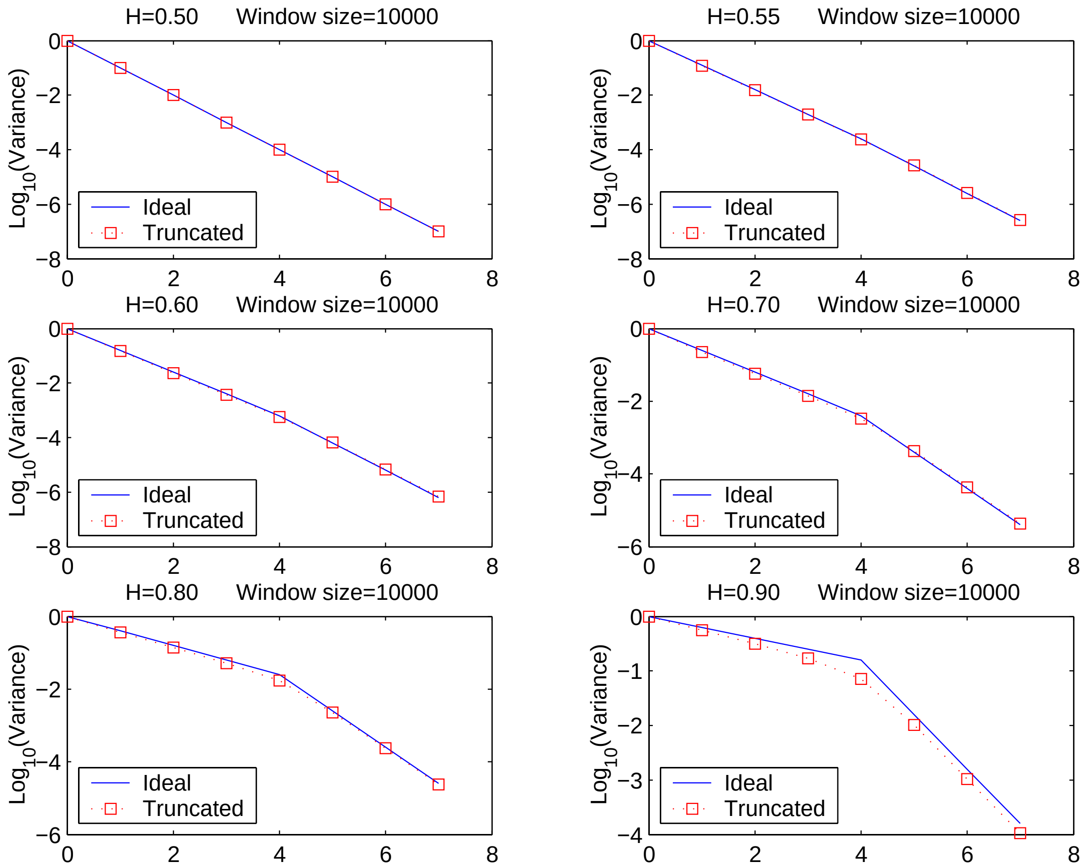
- Variance-time analysis ($\log_{10}$ scale) for the rounded truncated-filter output with truncation window $W = 10^3$.

- The slope of the (blue) solid line is equal to $2H - 2$ for $m \leq W$, and -1 for $m > W$.



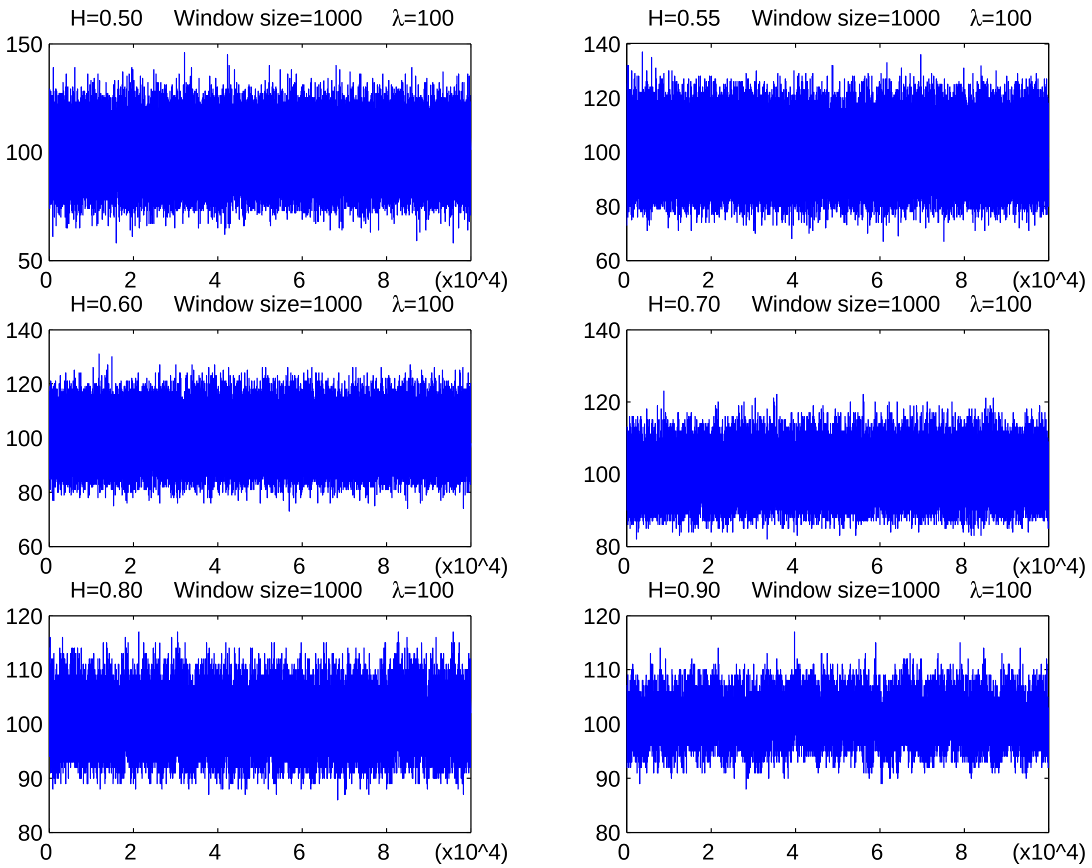
- Variance-time analysis ($\log_{10}$ scale) for the rounded truncated-filter output with truncation window $W = 10^4$.

- The slope of the (blue) solid line is equal to $2H - 2$ for $m \leq W$, and -1 for $m > W$.

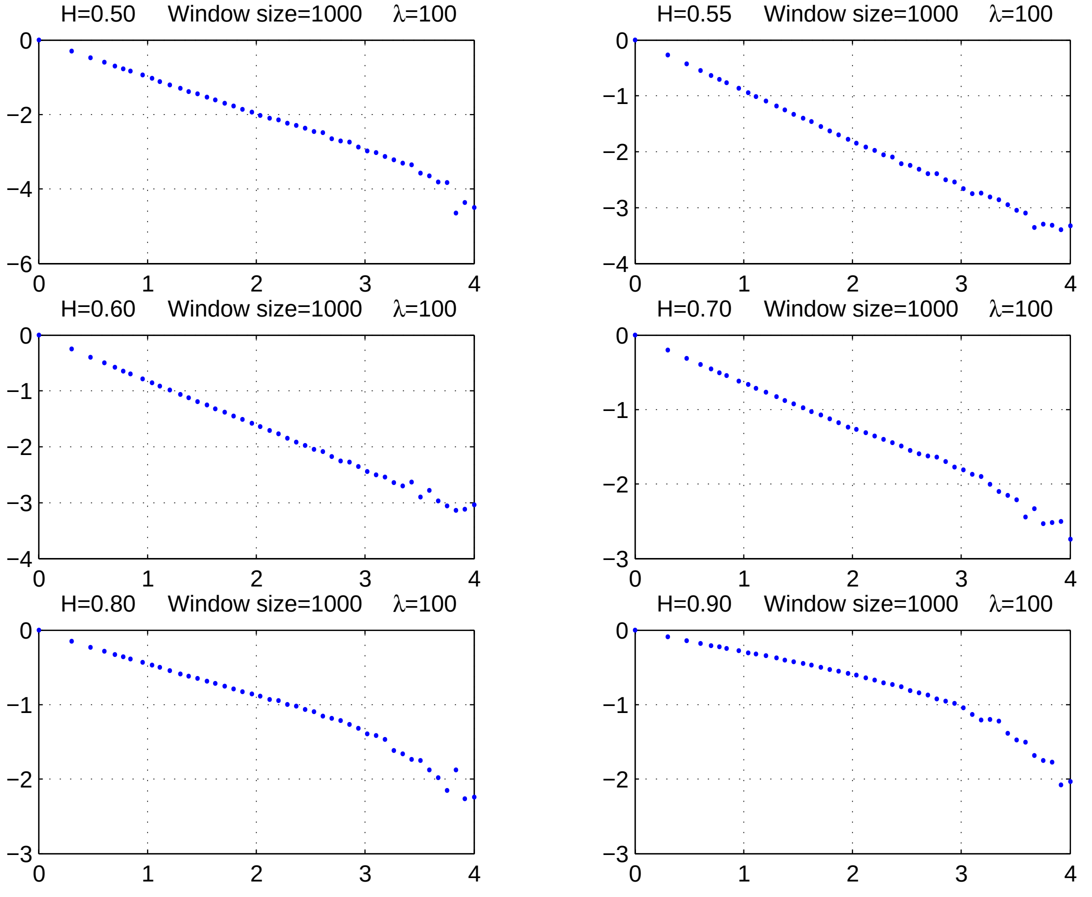


Test of generated traces

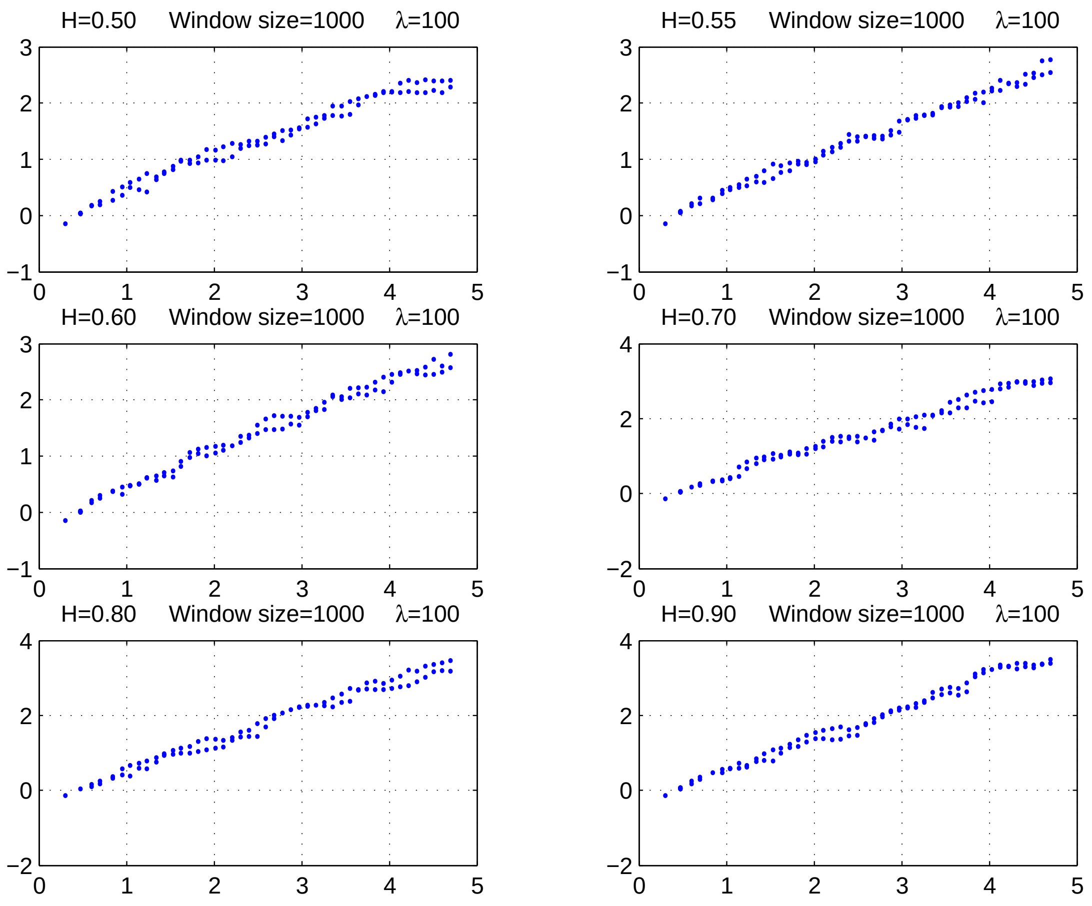
Filter-based synthesized arrival traffic with truncation window 10^3.



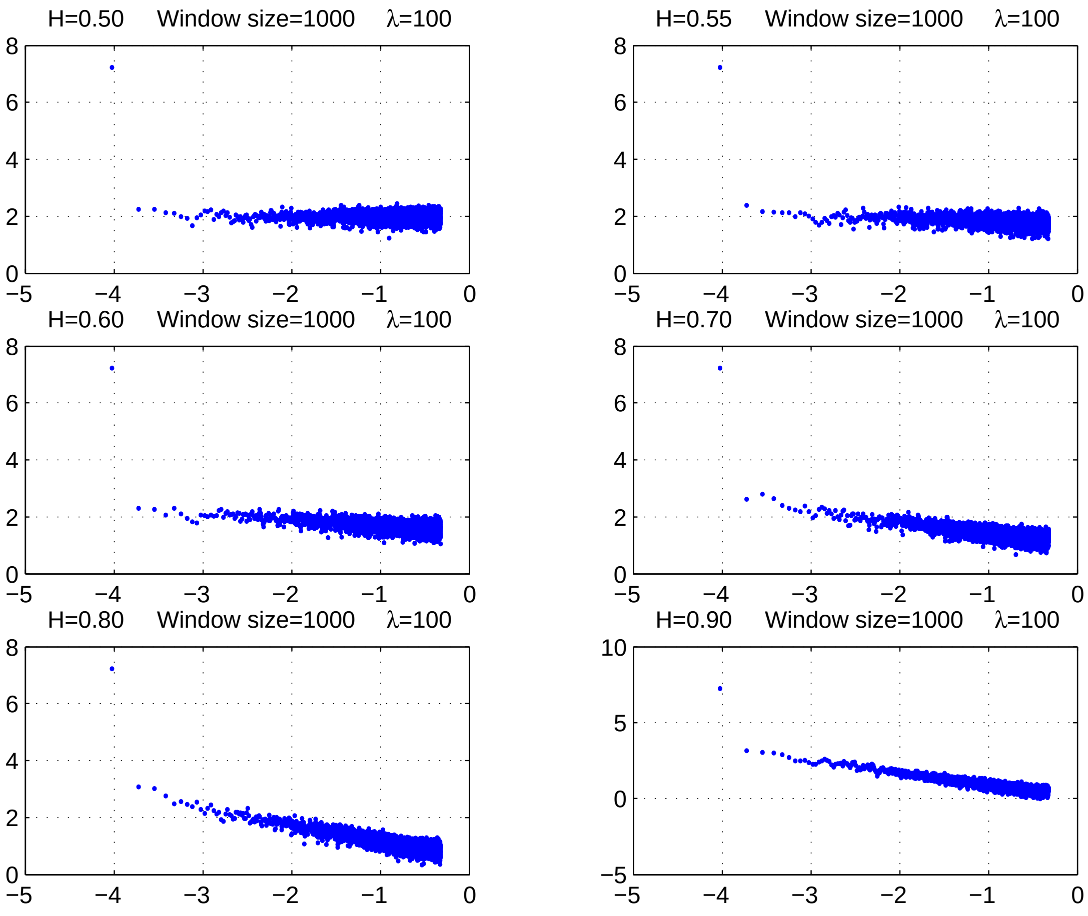
Variance-time plots ($\log_{10}$ scale) for the synthetic arrivals.



R/S plots for the synthetic arrivals.



Periodogram plots for the synthetic arrivals.



Window Length=10^3					
Ideal H	V-T ($m \leq W$)	V-T ($m > W$)	V-T	R/S	Periodogram
0.5001	0.4960561	0.4433182	0.4874457	0.5449268	0.5000444
	-0.81%	-11.34%	-2.53%	8.96%	-0.01%
0.55	0.5379669	0.4950074	0.5309530	0.5854389	0.5433692
	-2.19%	-1.00%	-3.46%	6.44%	-1.21%
0.6	0.6001428	0.4443881	0.5747131	0.6448041	0.6007081
	0.02%	-11.12%	-4.21%	7.47%	0.12%
0.7	0.6786277	0.5509871	0.6577881	0.6966293	0.6997149
	-3.05%	10.20%	-6.03%	-0.48%	-0.04%
0.8	0.7701778	0.6073597	0.7435949	0.7652807	0.8176999
	-3.73%	21.47%	-7.05%	-4.34%	2.21%
0.9	0.8486900	0.6015196	0.8083351	0.8443129	0.8826819
	-5.70%	20.30%	-10.19%	-6.19%	-1.92%

Impact of synthetic self-similarity by rounding

$\lambda = 1$; window size=10^4			
Ideal H	V-T	R/S	Periodogram
0.5001	0.4898783	0.5419452	0.5081486
	-2.04%	8.36%	1.61%
0.55	0.5504289	0.6089139	0.5424733
	0.08%	10.71%	-1.37%
0.6	0.6413529	0.6844096	0.6475869
	6.89%	14.07%	7.93%
0.7	0.4775099	0.5169073	0.5137266
	-31.78%	-26.16%	-26.61%
0.8	0.5399816	0.5730367	0.5276853
	-32.50%	-28.37%	-34.04%
0.9	0.5958403	0.6445508	0.5898216
	-33.80%	-28.38%	-34.46%

$\lambda = 10$; window size=10^4			
Ideal H	V-T	R/S	Periodogram
0.5001	0.5064982	0.5707382	0.5008728
	1.28%	14.12%	0.15%
0.55	0.5344366	0.5841011	0.5102031
	-2.83%	6.20%	-7.23%
0.6	0.5641452	0.5926723	0.6080758
	-5.96%	-1.22%	1.35%
0.7	0.7013537	0.6895256	0.7089697
	0.19%	-1.50%	1.28%
0.8	0.7799114	0.7876325	0.8004025
	-2.51%	-1.55%	0.05%
0.9	0.8716414	0.8550187	0.8922225
	-3.15%	-5.00%	-0.86%

- When the ratio of the maximal rounding error (here, 0.5) against λ (e.g., 10) is made less than 5%, the rounding influence on the degree of self-similarity of the filter output process can be neglected.

Conclusion and Further Work

- In this thesis, a new model is proposed for the synthesization of self-similar traffics based on the filter technique. This model can synthesize long range dependent with adjustable levels of bustiness and correlation.
- The model is **parsimonious** in the number of model **parameters**. Specifically, it only depends on three parameters:
  - H (self-similar parameter) : control the bustiness and autocorrelation of the synthesized traffic
  - λ (better ≥ 10) : define the mean of the synthesized traffic
  - W : determine not only the length of the filter but also the valid aggregation size of self-similar nature from the aspect of variance-time analysis.
- Compared to the RMD and the Paxson IFFT, our model provides additional advantages that the synthetic traffic can be generated **on the fly**, and is always **non-negative**.

- **Complexity**
  - The known fastest algorithm, Paxson IFFT, requires $(n/2)(\log_2(n) + 2)$ complex multiplications, where n is the length of the generated trace.
  - Our filter-based approach requires $n \times W$ complex multiplications, where W represents the truncation window size.
- After analytically analyzing our approach based on variance-time test, together with simulations from R/S and periodogram tests, we conclude that our model guarantees the generation of a traffic with desired degree of self-similarity for the intended range.

Future work

- **Higher order statistics**
  - The self-similarity of a network trace is simply defined based on the second-order statistics such as autocorrelation. A question that naturally follows is that what the impact of the **higher order statistics** of the traffic on the system performance is. Few publications have touched this issue, which may be of some interest to specific applications.
- **Time-variant filter technique**
  - Another issue that can be of interest is that our approach now only employs a **time-invariant linear** filter. It should be possible to extend the self-similar range of the synthetic trace by incorporating **time-variant** or **input-dependent** or **non-linear** filter technique. Exploration along this line may result in further improvement of our system.