Theoretical Network Load Limit when Self-similarity has no Adverse Effect on the Network

Hung V. Tran and Thomas M. Chen
Netrake Corporation and Southern Methodist University
Dallas, Texas
Phone: (214) 291-1111 and (214) 768-8541
Fax: (214) 291-1010 and (214) 768-3573
Email: hung@netrake.com and tchen@seas.smu.edu

Abstract:

It is well-known that the Multimedia traffic is statistically self-similar, with fractal-like behavior, that is, network traffic exhibits scale-invariance at a wide range of scales. Such scale-invariance is different from traditional models of network traffic. It is also well-known that self-similarity degrades the network performance by requiring large queueing buffers, causing delay and packet dropping problems, and that the traditional queueing theory is inadequate to predict network performance. The conventional wisdom is that the higher the load on the network, the higher the degree of self-similarity.

In this paper we first show a method of how to approximate a queueing buffer size with self-similar input process, then we will derive a theoretical network load limit when self-similarity has no adverse effect on the network. This load limit was found to be about the same load value published in literature, on average delay, with actual Ethernet traffic traces compared against curve obtained from the traditional Queueing Network Analyser.

I. Fractional Brownian Motion Storage Model

In [1], the authors plotted the logarithm of the probability of queue length exceeding a value of x, log (P[Queue length $> x$]), on the y-axis versus the Queue length x on the x-axis. Then the curve using real Internet trace and the curve obtained by simulation, using the Fractional Brownian Motion, are almost identical, showing that for large data, asymptotically the fluid model using Fractional Brownian Motion gives good estimates in the case of large traffic flows.

Materials in this section were taken from [2], [3] and [4]. Please refer to these published articles for more details.

Traditionally, a *normalized Fractal Brownian Motion* model $Z(t)$, $-\infty < t < \infty$, with self-similarity parameter $H \in [\frac{1}{2}, 1)$, is assumed to satisfy ([2]) following properties

- $Z(t)$ has stationary increments
- $Z(0) = 0$, $E[Z(t)] = 0, \forall t$
- $E[Z^2(t)] = \mid t \mid^{2H}, \forall t$
- $Z(t)$ has continuous paths
- $Z(t)$ is Gaussian, i.e. its finite-dimensional distributions are multivariate Gaussian distributions
- For $H = \frac{1}{2}$, $Z(t)$ is the standard Brownian motion.

Most results in [2], [3] and [4] do not depend on the Gaussian character of $Z(t)$, so they can be immediately generalized by replacing $Z(t)$ by a more general self-similar process.

Definition 1:

The *stationary storage model with fractional Brownian net input* is a stochastic process $V(t)$, where

$$V(t) = Sup_{s \leq t}[A(t) - A(s) - C(t - s)], \qquad (1)$$

where
$t \in -(\infty, \infty)$,
$A(t) = mt + \sqrt{am}Z(t)$,
$Z(t)$ is a normalized fractional Brownian motion,
m is the mean input rate, $m > 0$,
a is the variance coefficient, $a > 0$,
$H \in [\frac{1}{2}, 1)$ is the self-similarity parameter of $Z(t)$, and
$C > m$ is the service rate.

Although we have introduced a traffic model $A(t)$ and a constant leak rate C, it is in fact mathematically relevant for $V(t)$ only that the net input process

$$V(t) = A(t) - Ct$$

is of the form $c_1 Z - c_2 t$, with $c_2 > 0$.

Clearly, $V(t)$ is a stationary process by definition. The equation

$$V(t) = Sup_{s \leq t}[A(t) - A(s) - C(t - s)]$$

is similar to the well-known expression for the amount of work (or virtual waiting time) in a queueing system with service rate C, cumulative work arrival process $A(t)$. Benes called it 'Reich's formula'.

We have

Proposition: ([2], Proposition 2.2)

Let $A_i(t)$ be processes

$$A_i(t) = m_i t + \sqrt{m_i a} Z_i(t), \forall t$$

where $a > 0$, $m_i > 0$, $i = 1,2, \ldots K$, and Z_i's are Fractional Brownian Motion processes with the common parameter H. Then

$$\begin{array}{rcl} A(t) & \equiv & \sum_{i=1}^{K} A_i(t) \\ & = & mt + \sqrt{ma} Z(t), \end{array}$$

where $m = \sum_{i=1}^{K} m_i$ and $Z(t)$ is a Fractional Brownian Motion process with the same parameter H.

□

Heuristically, a and H characterize the 'quality' of the traffic, while the long run mean rate m characterize its 'quantity'.

Since by (1)

$$V(t) \equiv Sup_{s \le t}[A(t) - A(s) - C(t-s)]$$

as defined above, and as $Z(\alpha t) =_{(d)} \alpha^H Z(t)$, where $=_{(d)}$ means 'equal in distribution', so we have

$$\begin{array}{rcl} V(\alpha t) & = & Sup_{s \le t}[A(\alpha t) - A(\alpha s) - C(\alpha t - \alpha s)] \\ & =_{(d)} & Sup_{s \le t}[m\alpha t + \sqrt{am} Z(\alpha t) \\ & & -m\alpha s - \sqrt{am} Z(\alpha s) - C(\alpha t - \alpha s)] \\ & =_{(d)} & Sup_{s \le t}[m\alpha t + \sqrt{am}\alpha^H Z(t) \\ & & -m\alpha s - \sqrt{am}\alpha^H Z(s) - C(\alpha t - \alpha s)] \\ & =_{(d)} & \alpha^H Sup_{s \le t}[m\alpha^{1-H} t + \sqrt{am} Z(t) \\ & & -m\alpha^{1-H} s - \sqrt{am} Z(s) \\ & & -\alpha^{1-H} C(t-s)] \\ & =_{(d)} & \alpha^H Sup_{s \le t}[\sqrt{am} A(t) - \sqrt{am} A(s) \\ & & -(m + \alpha^{1-H}(C-m)(t-s))]. \end{array}$$

Thus, we have

Theorem 1:

Define process $V(t)$ as above, with parameters m, H, a and C. Then for every $\alpha > 0$, the process $V(\alpha t)$ is distributed like α^H times the corresponding Fractional Brownian Motion process with parameters $m + \alpha^{1-H}(C-m)(t-s)$, H, a and C.

□

In telecommunication, a typical requirement in an application is for the probability of a certain amount of work in the system to exceed a certain level x to be upper-bounded by a 'Quality-of-Service' parameter ϵ.

$$P[V > x] = \epsilon.$$

This equation also defines the *storage requirement* x. This equation also defines a hypersurface in the space of system parameters, separating the acceptable parameters from the unacceptable ones.

We call x the storage requirement, C the service rate, and $\rho \equiv \frac{m}{C}$ the utilization, at the critical boundary (the hypersurface mentioned above).

We have the following theorem

Theorem 2:

Assume

$$P[V > x] = \epsilon.$$

then

$$\frac{1-\rho}{\rho^{1/(2H)}} C^{(H-\frac{1}{2})/H} x^{(1-H)/H} = constant (= a^{1/(2H)} f^{-1}(\epsilon)), \tag{2}$$

where the constant $a^{1/(2H)} f^{-1}(\epsilon)$ on the right-hand side depends only on a, H and ϵ.

Proof:

We have

$$\begin{array}{rcl} \epsilon & = & P[V > x] \\ & = & P[Sup_{t \ge 0}(\sqrt{am} Z(t) - (C-m)t) > x] \\ & = & P[Sup_{t \ge 0}(Z(t) - \frac{(C-m)}{\sqrt{am}} t) > \frac{x}{\sqrt{am}}] \\ & = & f(\left(\frac{x}{\sqrt{am}}\right)^{(1-H)/H} \frac{(C-m)}{\sqrt{am}}) \end{array}$$

Let $\rho \equiv \frac{m}{C}$, then

$$\begin{array}{rcl} f^{-1}(\epsilon) & = & \left(\frac{x}{\sqrt{am}}\right)^{(1-H)/H} \frac{(C-m)}{\sqrt{am}} \\ & = & x^{(1-H)/H} (a\rho C)^{-(1-H)/2H} (C - \rho C)(a\rho C)^{-1/2} \end{array}$$

so that

$$\frac{1-\rho}{\rho^{1/(2H)}} C^{(H-\frac{1}{2})/H} x^{(1-H)/H} = a^{1/(2H)} f^{-1}(\epsilon) = constant.$$

□

In the case of without self-similarity, when $H = \frac{1}{2}$, we have by (2)

$$\frac{1-\rho}{\rho} x = a f^{-1}(\epsilon) = constant.$$

This is a traditional case of a heavy traffic approximation of the M/D/1 queue. Solving for x, we have then for $H = \frac{1}{2}$

$$x_{1/2} = (a f^{-1}(\epsilon)) \frac{\rho}{1-\rho} = (constant) \frac{\rho}{1-\rho}. \tag{3}$$

For the general case, we have by (2)

$$\frac{1-\rho}{\rho^{1/(2H)}} C^{(H-\frac{1}{2})/H} x^{(1-H)/H} = constant = a^{\frac{1}{2H}} f^{-1}(\epsilon),$$

so solving for x in the general case for H, we have

$$x_H = (a^{\frac{1}{2H}} f^{-1}(\epsilon))^{\frac{H}{1-H}} \left(C^{\frac{2H-1}{2(1-H)}}\right) \left(\frac{\rho^{\frac{1}{2}}}{(1-\rho)^H}\right)^{\frac{1}{1-H}}, \quad (4)$$

which explicitly involves C also.

We have thus by (3) and (4)

$$\begin{aligned} \frac{x_H}{x_{1/2}} &= \left(a^{\frac{1}{2(1-H)}-1} f^{-1}(\epsilon)^{\frac{H}{1-H}-1} C^{\frac{2H-1}{2(1-H)}}\right) \\ &= \left(\frac{\rho^{\frac{1}{2}}}{(1-\rho)^H}\right)^{\frac{1}{1-H}} \left(\frac{1-\rho}{\rho}\right) \\ &= \left((aC)^{\frac{1}{2}} f^{-1}(\epsilon)\right)^{\frac{2H-1}{1-H}} \left(\frac{\rho^{\frac{1}{2}}}{1-\rho}\right)^{\frac{2H-1}{1-H}}. \end{aligned}$$

And so, we have

$$\frac{x_H}{x_{1/2}} = \left(\frac{\beta\rho^{\frac{1}{2}}}{1-\rho}\right)^{\frac{2H-1}{1-H}}, \quad (5)$$

where

$$\beta \equiv (aC)^{\frac{1}{2}} f^{-1}(\epsilon) \quad (6)$$

is a constant.

Theorem 3:

Let x_H be the storage size.

1. If H is fixed, when $\frac{1}{2} < H < 1$, then the ratio $r(\rho) \equiv \frac{x_H}{x_{1/2}}$ increases monotonically with respect to the utilization factor ρ, where $r(0) = 0$, $r((\frac{-\beta+\sqrt{\beta^2+4}}{2})^2) = 1$ and $r(1) = \infty$. Thus, specifically,

- If $0 \le \rho < \left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2$, then $x_H < x_{1/2}$.
- If $\left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2 < \rho \le 1$, then $x_H > x_{1/2}$.

2. Let ρ be fixed, when $0 \le \rho \le 1$, and let the ratio $r(H)$ be defined as $r(H) \equiv \frac{x_H}{x_{1/2}}$, then

- If $0 \le \rho < \left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2$, then $x_H < x_{1/2}$ and $r(H)$ approaches 0 when H approaches 1.
- If $\left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2 < \rho \le 1$, then $x_H > x_{1/2}$ and $r(H)$ becomes infinite when H approaches 1.

Proof:

1. If H is fixed, when $\frac{1}{2} < H < 1$, then $\frac{2H-1}{1-H} > 0$.

Let

$$f(\rho) \equiv \frac{\beta\rho^{\frac{1}{2}}}{1-\rho},$$

then by (5)

$$\frac{x_H}{x_{1/2}} = \left(\frac{\beta\rho^{\frac{1}{2}}}{1-\rho}\right)^{\frac{2H-1}{1-H}} = (f(\rho))^{\frac{2H-1}{1-H}}.$$

Since $\frac{2H-1}{1-H} > 0$, so that $\frac{x_H}{x_{1/2}}$ monotonically increases with respect to $f(\rho)$.

But the derivative of $f(\rho)$ is

$$f'(\rho) = \frac{\beta(1-\rho)}{2\rho^{1/2}(1-\rho)^2},$$

so $f'(\rho) > 0$ when $0 \le \rho \le 1$. Thus, $f(\rho)$ increases monotonically with respect to ρ, when $0 \le \rho \le 1$.

Thus, $\frac{x_H}{x_{1/2}}$ monotonically increases with respect to ρ, when $0 \le \rho \le 1$.

We notice that $f(0) = 0$, $f(1) = \infty$. Also, when $\rho = \left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2$, then $f(\rho) = 1$. Thus, when H is fixed and $\frac{1}{2} < H < 1$, when $0 \le \rho < \left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2$, then the storage requirement x_H is smaller than the traditional storage $x_{1/2}$, but it will increase faster as ρ increases, and reach $x_{1/2}$ when the utilization is about $\left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2$. After the utilization ρ passes the value $\left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2$, then the storage requirement x_H gets larger than the traditional storage requirement $x_{1/2}$, when ρ gets larger, and x_H becomes infinitely larger than $x_{1/2}$, when the utilization is about 1.

2. If we fix ρ and let H vary, where $\frac{1}{2} < H < 1$, consider the function

$$f(H) \equiv \frac{2H-1}{1-H},$$

then its derivative is

$$f'(H) = \frac{1}{(1-H)^2} > 0.$$

Thus the function $f(H)$ increases monotonically with respect to H.

We notice that $f(\frac{1}{2}) = 0$ and $f(1) = \infty$.

Since by (5)

$$\frac{x_H}{x_{1/2}} = \left(\frac{\beta\rho^{\frac{1}{2}}}{1-\rho}\right)^{\frac{2H-1}{1-H}}$$

and as ρ is fixed, so that

- $\frac{x_H}{x_{1/2}}$ will be less than 1 when $\frac{\beta\rho^{\frac{1}{2}}}{1-\rho} < 1$, in which case we have

$$0 \le \rho < \left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2,$$

and $\frac{x_H}{x_{1/2}}$ approaches 0 when H approaches 1

- $\frac{x_H}{x_{1/2}}$ will be greater than 1 when $1 < \frac{\beta\rho^{\frac{1}{2}}}{1-\rho}$, in which case we have
$$\left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2 < \rho \leq 1,$$
and $\frac{x_H}{x_{1/2}}$ becomes infinite when H approaches 1.

□

We notice a surprising result in Theorem 3 above: the higher the degree of self-similarity H, the lower the storage size requirement x_H, comparing with the traditional storage size when no self-similarity is detected, when the utilization factor ρ is less than $\left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2$. Thus, in case of high degree of self-similarity in Internet traffic, we want to keep the utilization factor as low as or lower than $\left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2$ as possible.

Intuitively, when the network utilization is low, the service rate is high enough to process data, thus less storage area is required for data to wait to be processed. Since long-range dependence tends to emphasize the trend, if the trend is for the data to be processed quickly and thus requiring less storage area, it continues to do so. On the other hand, when the network utilization is high, the service rate is low comparing to the rate of incoming data, more storage area is required to contain unprocessed data. Again, since long-range dependence tends to emphasize the trend, if the trend is to put data in storage to wait to be processed, long-range dependence will tend to amplify this trend, thus requiring even more storage area than its counter-part of almost purely random data.

We notice that the function
$$g(x) \equiv \frac{-\beta+\sqrt{\beta^2+4}}{2}$$
has the derivative
$$g'(x) = \frac{1}{2}\left(\frac{\beta-\sqrt{\beta^2+4}}{\sqrt{\beta^2+4}}\right),$$
which is less than zero when $\beta \geq 0$. Thus $g(x)$ decreases when $\beta \geq 0$. Since $g(0) = 1$ and $g(\infty) = 0$, so $0 \leq \left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2 \leq 1$, when $\beta \geq 0$.

The constant $\beta \equiv (aC)^{\frac{1}{2}} f^{-1}(\epsilon)$ in (6) contains parameters a, C and $f^{-1}(\epsilon)$, where

- a is the variance coefficient
- C is the service rate
- $f^{-1}(\epsilon)$ is related to the desired QoS.

Thus, normally the constant β does not depend too heavily on the degree of self-similarity H, and we can then write
$$\frac{x_H}{x_{1/2}} = (constant)\left(\frac{\rho^{\frac{1}{2}}}{1-\rho}\right)^{\frac{2H-1}{1-H}}.$$
Since $\frac{x_H}{x_{1/2}} = 1$ when $H = \frac{1}{2}$, so that the *constant* in the above equation is actually 1, in which case $\beta = 1$.

Also, even if the constant β does depend heavily on the degree of self-similarity H, but if in the network, H varies much slower than the utilization factor ρ, then we have
$$\begin{aligned}\frac{x_H}{x_{1/2}} &= \left(\frac{\beta\rho^{\frac{1}{2}}}{1-\rho}\right)^{\frac{2H-1}{1-H}} \\ &= (\beta)^{\frac{2H-1}{1-H}}\left(\frac{\rho^{\frac{1}{2}}}{1-\rho}\right)^{\frac{2H-1}{1-H}} \\ &\approx (constant)\left(\frac{\rho^{\frac{1}{2}}}{1-\rho}\right)^{\frac{2H-1}{1-H}}.\end{aligned}$$
Again, in this case, the *constant* in the above equation is actually 1, and so $\beta = 1$.

Thus, in the two situations above, $\left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2 = (\frac{-1+\sqrt{5}}{2})^2 \approx 0.38196601$. So that in this case, when the degree of self-similarity H starts to get too high, and the packets start dropping, we want to keep the utilization factor as low as or lower than $(\frac{-1+\sqrt{5}}{2})^2 \approx 0.38196601$ as possible.

In [1], as described in section 'Fractal Queueing' above, the authors plotted average delay on the y-axis versus utilization on the x-axis, for time series of interarrival times. In Figure 1 (page 254), the curve (A) with the actual Ethernet traffic traces was plotted against curve (B) obtained from the traditional Queueing Network Analyzer. We notice that these two curves intersect around the utilization value of 0.38, matching our value predicted above.

Comparing between the calculations of $\frac{x_H}{x_{1/2}}$ in (5) and that of x_H in (4), we note that

- x_H in (4) involves explicitly variables a, $f^{-1}(\epsilon)$, which might be hard or impossible to obtain.
- Approximately, when the degree of self-similarity H of the traffic does not vary excessively fast with time (experimentally, it varies between 0.8 and 0.9 in Bellcore data), then the constant part β of $\frac{x_H}{x_{1/2}}$ in (5) is approximately 1, allowing easy calculations of $\frac{x_H}{x_{1/2}}$.
- Since both $x_{1/2}$ and x_H tend to increase or decrease together, under various network conditions, so $\frac{x_H}{x_{1/2}}$ should be more invariant in time and network conditions than x_H alone.

The discovery of self-similar traffic has posed a considerable challenge on effort to approximate queue length, as the traditional methods are based on such assumptions like independent increments, renewal process, Markovian assumptions, etc. The explicit form for exact queue length distribution seems to be unknown still for any queueing system with a long-range dependent input process ([21]). Thus, any good approximation for x_H is welcome not only for practical applications in Telecommunication, but also is important theoretically as a check against further developments.

Under the general setting in Theorem 2, it is a good assumption that in the heavy traffic condition, the ratio $\frac{x_H}{x_{1/2}}$ should give a good approximation under various network conditions. A simple algorithm to approximate x_H under diversed network conditions could be described as follows:

1. Use the traditional methods to calculate $x_{1/2}$, in which case, there is no self-singularity.

2. Calculate $\frac{x_H}{x_{1/2}}$ in (5). If constant β is not known, and if the degree of self-similarity H does not vary much, we may assume that $\beta = 1$ in (5).

3. Approximate x_H when the data traffic has self-similarity by

$$x_H(\Phi, \rho, H) \approx (x_{1/2}(\Phi)) \left(\frac{x_H}{x_{1/2}}(\rho, H) \right),$$

where Φ is a set of parameters for $x_{1/2}$. We note that Φ might contain ρ.

4. If simulation is available, verify the above approximation expression for x_H.

To approximate the constant β in Theorem 3, one procedure is to collect the delays under different utilization factor values as done in [1], then these values are to be plotted out against those values such as from the traditional Queueing Network Analyzer ([1]). The intersecting value C^2 between these two curves would give an approximate value for $\left(\frac{-\beta+\sqrt{\beta^2+4}}{2}\right)^2$. Solving for

$$\left(\frac{-\beta + \sqrt{\beta^2 + 4}}{2} \right)^2 = C^2,$$

we have

$$\beta = \frac{1 - C^2}{C}.$$

We notice that its derivative is

$$\beta' = \frac{-1 - C^2}{C^2} < 0,$$

so when C varies from 0 to 1, β decreases from ∞ to 0.

II. Conclusion

As the theory was developed for a single server case, in the case of multiple servers situation, if it could be partitioned equivalently into a combination of queues with single servers, obviously the above algorithm applies immediately. In the case that the queue cannot be logically partitioned into combinations of queues of single servers, if the ratio $\frac{x_H}{x_{1/2}}$ does not vary much over the network environment, it might still be possible to use the above algorithm, although simulations ought to be used to verify that it is acceptable to do so.

In the multimedia traffic case, where each stream has different H, assuming independence of these streams of data, different x_H's could be calculated and we could take the maximum in case of multiplexed data.

For QoS and Performance control applications, we thus have

1. An algorithm to approximate the buffer size in self-similar traffic situation, which would tell us if buffer overflow conditions might have occurred in the router/switch or in the network.

2. Also, as mentioned above, when due to self-similarity of traffic data, the packet dropping probability starts to increase and exceed a tolerable level, or when the delay starts to be too excessive in this case, methods like Call Admission Control could be employed to admit only selected new applications, to reduce down the utilization factor to about perhaps 0.38196601 to eliminate the system degradation due to high traffic self-similarity degree.

(This work was supported in part by National Science Foundation grant 9980561. This work does not represent the beliefs or policies of the NSF.)

III. References:

[1] Fractal Queueing Models, A. Erramilli, O. Narayan, W. Willinger, pp. 245-270, 'Frontiers in Queueing', ed. J.H. Dshalalow, CRC Press, 1997.

[2] A Storage model with self-similar input, I. Norros, Queueing Systems, 16, 1994, 387-396.

[3] Time Scale Analysis of an ATM Queueing System with Long-range dependent Traffic, G. Mayor, J. Silvester, INFOCOM 97, Vol. 1, pages 205-212.

[4] On the use of Fractional Brownian Motion in the Theory of Connectionless Networks, I. Norros, IEEE Journal on Selected Areas in Communications, Vol 13, No. 6, August 1995, Pages 953-962.