

Implications of long-range dependence

Damon Wischik*, Statistical Laboratory

February 26, 2001

1 What is long-range dependence?

Leland, Taqqu, Willinger, and Wilson [8] observed that, over long time-scales, Internet traffic is *self-similar*. To explain what this means, some notation:

Let $X(0, t]$ be the total amount of traffic passing through a particular point in the network in the time interval $(0, t]$. Write X for the overall traffic process. Define the speeded-up version $X^{\otimes L}$ by $X^{\otimes L}(0, t] = X(0, Lt]$.

They observed that the process looks similar over all (long enough) time-scales: that is, the rescaled process

$$\frac{1}{a^H} X^{\otimes a}$$

has approximately the same distribution whatever the value of a (for large a). The parameter H is the Hurst parameter; they measured H and found it typically lay in $(\frac{1}{2}, 1)$. A self-similar process has bursts over all time-scales; a nice term for this is *Hurstiness*.

There is one class of processes for which the rescaled versions have exactly the same distribution for all a —those of the form $X(0, t] = \mu t + \sigma Z_t$, where Z_t is a *fractional Brownian motion*, for which $Z_t \sim \text{Normal}(0, t^{2H})$. This has become a popular model for Internet traffic: it allows one to investigate self-similarity, and it is analytically tractable.

How does traffic come to be self-similar? It has been shown that when many on/off sources are aggregated, if the on duration has a *heavy-tailed distribution*, then the (suitably rescaled) aggregate converges to a fractional Brownian motion.

For queueing theory, an important characteristic of fractional Brownian motion is that its variance grows as

$$\text{Var } X(0, t] \sim t^{2H} \quad \text{for large } t.$$

When $H > \frac{1}{2}$, this is called *long-range dependence*. It is not seen in traditional teletraffic models (Poisson processes, Markov chains etc.) in which $\text{Var } X(0, t] \sim t$. The famous plot by Leland et al., poorly reproduced in Figure 1, gives an excellent demonstration.

*Some of these ideas come from discussions with Richard Gibbens, Frank Kelly, and James Martin. Errors and misconceptions are my own.

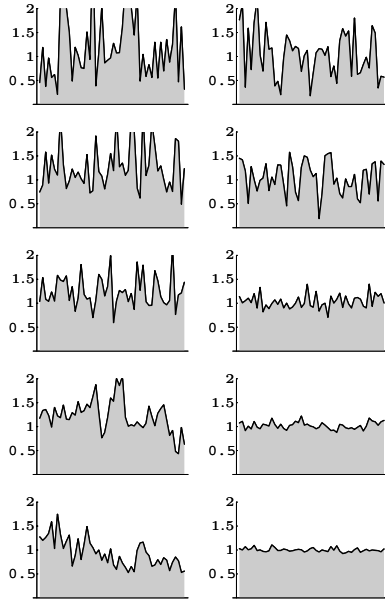


Figure 1: Internet traffic (left) is bursty at all timescales, while a traditional Gaussian traffic model (right), matched to have the same mean and variance, is not. The top plots have the finest time-scale; each subsequent plot shows a four-fold longer time-scale.

2 Multiplexing long-range dependent traffic

What happens when several long-range dependent flows are multiplexed together? Specifically, let $X^{\oplus L}$ be the aggregate of L independent copies of a traffic process X . What is the distribution of $X^{\oplus L}$?

When X is a fractional Brownian motion, say $X(0, t] = \mu t + \sigma Z_t$, the answer is very simple: $X^{\oplus L}$ is another fractional Brownian motion, with $L^{-1}X^{\oplus L}(0, t] = \mu t + L^{-1/2}\sigma Z_t$. In other words, the aggregate is smoother, by a factor of $L^{1/2}$. It is just as Hursty as before—it has exactly the same Hurst parameter H —but it *is* smoother.

This happens with just about any traffic process. By the central limit theorem, under very mild conditions on the process X ,

$$L^{-1}X^{\oplus L}(0, t] \sim N(\mu t, L^{-1/2}\sigma_t^2)$$

where μt and σ_t^2 are the mean and variance of $X(0, t]$.

To illustrate this, Figure 2 shows traffic traces for the Bellcore data, at different degrees of aggregation, and compares them to traffic traces for a matched Gaussian process. (The Bellcore data set is a traffic trace for a single network; I split the data up into several non-overlapping pieces, and looked at aggregates of two or more of those pieces.)

For mathematical details of the central limit theorem applied to aggregated traffic processes, see [1, 2]. For details of large deviations theory (a different sort of limit, which gives rise to the queueing estimates I describe next) applied

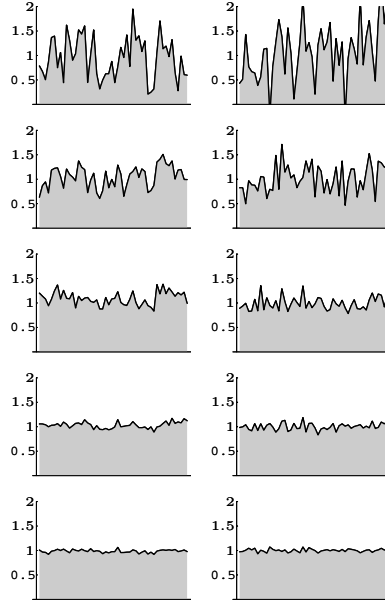


Figure 2: Internet traffic (left) becomes smoother as it is aggregated, as does a traditional Gaussian traffic model (right), matched to have the same mean and variance. The top plots show traffic from a single flow; each subsequent plot shows a four-fold increase in the level of aggregation.

to aggregated traffic processes (and yielding the same sort of conclusion), see [12].

3 Implications for queueing theory

For traditional teletraffic models, in which $\text{Var } X(0, t] \sim t$, it can be shown that the probability of overflow in a queue fed by traffic process X , served at rate C with buffer size B , is roughly

$$\log \mathbb{P}(\text{overflow}) \approx -BI(C) \quad \text{for large } B,$$

for some function $I(C)$ (which depends on the process X). This is *not true* for long-range dependent processes. In particular, for fractional Brownian motion $X(0, t] = \mu t + \sigma Z_t$,

$$\log \mathbb{P}(\text{overflow}) \approx -B^{2(1-H)}I(C) \quad \text{for large } B. \quad (1)$$

See [5, 10] for mathematical details.

This led to the interpretation: traditional teletraffic models provide bad estimates; since real Internet traffic is self-similar, we need much larger buffers than they suggest, if we are to provide quality of service.

This is not altogether fair. A better estimate for the probability of overflow is

$$\log \mathbb{P}(\text{overflow}) \approx -I(B, C)$$

where

$$I(B, C) = \inf_{t>0} \sup_{\theta \geq 0} \theta(B + Ct) - \theta t \alpha(\theta, t) \quad (2)$$

and $\alpha(\theta, t)$ is the *effective bandwidth* [7]

$$\alpha(\theta, t) = \frac{1}{\theta t} \log \mathbb{E} e^{\theta X(0,t]}.$$

This approximation *applies equally to traditional teletraffic models and to long-range dependent models*. It becomes more accurate as the level of aggregation increases.

In a system which serves an L-fold aggregate $X^{\oplus L}$, and has service rate C per flow and buffer size B per flow, $I(LB, LC) = LI(B, C)$. Thus, the probability of overflow decays exponentially in the degree of multiplexing.

In a heavily loaded queue, one can approximate

$$I(B, C) \approx \inf_t \frac{(B + (C - \mu)t)^2}{2\sigma_t^2}$$

where μt and σ_t^2 are the mean and variance of $X(0, t]$.

For mathematical details, see [3, 4, 12]. For the heavy-load case, see [11]. For improved estimates, see [9].

This result shifts the emphasis away from the Hurst parameter H , which is the governing term in (1), and towards the *critical time-scale*, the optimizing t^* in (2). Perhaps Hurstiness is not the right way to measure burstiness: perhaps it is better to look at $\sigma_{t^*}^2$, the variance over the critical time-scale.

4 Implications for optical networks

I do not have a mathematical model for an optical network, so I cannot be as concise in describing the implications of long-range dependence for optical networks.

A long-range dependent process like fractional Brownian motion has bursts at all time-scales: ‘peaks, riding on bursts, riding on swells’. This has certainly been observed in Internet traffic. It means that at any given point in the network, there will be sustained busy periods and sustained quiet periods.

Another undoubted characteristic of Internet traffic is that it is not stationary. Sites become popular, then unpopular; traffic demand shifts around the network. This means that at any given point in the network, there will be sustained busy periods and sustained quiet periods.

Indeed, given a traffic trace, it is hard to distinguish long-range dependence from non-stationarity. Both will tend to exhibit burstiness at the time-scale of the length of the trace. Which model we choose will depend on *a priori* beliefs. For example, if we know that a particular news site was popular at a certain time, we could fit a non-stationary non-long-range-dependent model to the traffic trace. If we have no such knowledge, we could fit a stationary long-range dependent model.

Putting philosophy aside, an optical network architecture must surely take into account non-stationarity. It must surely be able to adapt, to allocate

more resources to parts of the network that are momentarily popular. Suppose (loosely) that it can adapt to changes in demand over some time-scale T . Then, who cares if traffic bursts over time-scales T and larger come about through non-stationarity or through long-range dependence? The network should adapt to the change in demand, whatever its cause.

What about bursts over time-scales shorter than T ? The network may not be able to shift its resources quickly enough to cope with these bursts. This is a classic problem, with a classic solution: make sure the network provides sufficient capacity (eg number of wavelengths, buffer size at edge routers) to cope with bursts over these time-scales.

Note that the provisioning problem depends only on bursts over time-scales shorter than T . Perhaps an appropriate way to measure burstiness is by the variances $(\sigma_t^2)_{0 < t < T}$. Hurstiness is not an issue, since it describes σ_t^2 for large t ; and a process with no long-range dependence at all may have very large short-timescale variances $(\sigma_t^2)_{0 < t < T}$.

5 Conclusion

Long-range dependent traffic processes become smoother when aggregated, in exactly the same way as traffic processes from traditional teletraffic models. The aggregate is still just as Hursty; nonetheless, over any fixed timescale, it is smoother.

Any optical network architecture should adapt to changing demand by re-locating its resources. Suppose it can adapt over a certain time-scale T . Then it will be useful to understand how bursty the traffic is over time-scales less than T ; over longer time-scales, classic notions of non-stationarity should be sufficient.

Since self-similarity and Hurst parameters describe burstiness over very long time-scales, they do not seem to be relevant here.

References

- [1] R. G. Addie. On weak convergence of long-range dependent traffic processes. Technical Report SC-MC-9816, University of Southern Queensland, 1998.
- [2] Ronald G. Addie, Moshe Zukerman, and Timothy D. Neame. Application of the central limit theorem to communication networks. Technical Report SC-MC-9819, University of Southern Queensland, 1998.
- [3] D.D. Botvich and N.G. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems*, 20: 293–320, 1995.
- [4] Costas Courcoubetis and Richard Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *Journal of Applied Probability*, 33:886–903, 1996. URL <http://www.statslab.cam.ac.uk/~rrw1/research/nsource2.ps>.

- [5] N. G. Duffield and N. O’Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *Mathematical Proceedings of the Cambridge Philosophical Society*, 118:363–374, 1995.
- [6] F. P. Kelly, S. Zachary, and I. Ziedins, editors. *Stochastic Networks: Theory and Applications*. Royal Statistical Society Lecture Note Series. Oxford, 1996.
- [7] Frank Kelly. Notes on effective bandwidths. In Kelly et al. [6], chapter 8, pages 141–168. URL <http://www.statslab.cam.ac.uk/~frank/eb.html>.
- [8] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, 1994.
- [9] N. Likhanov and R. R. Mazumdar. Cell loss asymptotics for buffers fed with a large number of independent stationary sources. *Journal of Applied Probability*, 36:86–96, 1999.
- [10] Neil O’Connell. Queue lengths and departures at single-server resources. In Kelly et al. [6], chapter 5. URL <ftp://hplose.hpl.hp.com/pub/noc/papers/9604.ps>.
- [11] Damon Wischik. Moderate deviations in queueing theory. URL <http://www.statslab.cam.ac.uk/~djw1005/Stats/Research/moddev.ps>. Work in progress, 2001.
- [12] Damon Wischik. Sample path large deviations for queues with many inputs. URL <http://www.statslab.cam.ac.uk/~djw1005/Stats/Research/sampleldp.html>. To appear in *Annals of Applied Probability*, 2001.