Quantifying Chaos: Metrics and the Internet

Kate Lance

connect.com.au pty. ltd.
clance@connect.com.au

September 1997

Abstract

Performance measurement is vitally important to both users and providers of Internet functionality, yet the available tools are limited, their interpretation is difficult, and as Internet complexity multiplies, whole new forms of traffic behaviour are being observed. One recent finding is that packet traffic has a self-similar, fractal structure. Self-similarity can degrade traffic throughput and increase packet loss, response time, and demand on bandwidth and buffer resources, and our understanding of its impact and relevance to more traditional metrics has only just begun.

Why Metrics?

Measurement of performance is important to everyone who uses the Internet. We all want to experience good response, reliable throughput and data integrity. At any single moment Internet performance is determined by complex inter-relationships between physical resources, protocol implementations and human activity. If we are to understand and improve Internet performance we need metrics for:

- **Immediate diagnostics**: to understand or trouble-shoot current problems
- **Baseline information**: to collect statistics for capacity planning and to understand long-term trends, usage and protocol changes
- **Performance prediction**: to evaluate and compare reliability of resource providers based on their past response and service
- **Research and development**: to understand the behaviour of Internet traffic for future implementations and problem-solving

Immediate diagnostics are achievable with a number of tools. One of the most complete collections available is at CAIDA, the Cooperative Association for Internet Data Analysis [1]. But the limitations of these tools are often not recognised, based as they are on assumptions that no longer apply to the Internet of today: for instance, almost two years ago a world-wide study [2] showed that around half of all routes studied included a major asymmetry, yet route symmetry is the underlying assumption of many common tools. But useful understanding may still be achieved as long as the limitations are taken into account.

Baseline information is something that all organisations, especially service providers, need to collect and keep for their own benefit. Data from immediate diagnostic tools (like *ping* or *traceroute*) collected over time can provide useful insights into average change.

Performance prediction might offer techniques for users to compare and predict the performance of service and network providers. Such techniques would need to be developed in conjunction with research efforts, as the range of possible influences is enormous.

Research and development metrics must have well-understood theoretical frameworks: they need to clearly define what they are measuring and why. They should be standardised so that they provide data that can be compared over different times and places, and they need to be non-intrusive and generate the minimum of traffic so that they don't affect the behaviour they are trying to measure.

One possible point at which to start measuring the Internet is to try to define the smallest building blocks of connectivity: for instance, what do we mean by transmission time on a link? How do we measure it? What time units, what size of data units? Where do we measure it from? What are the limitations, the context, the assumptions, the uncertainties in the definition? The IPPM (Internet Protocol Performance Metrics) group of the IETF has made a major contribution to this field, but its work has shown that the task is extremely difficult (see [3], [4]).

Traffic Pathologies

We tend to assume that the measurement of Internet characteristics is reasonably well-understood and as long as enough data is collected and analysed then understanding of conditions on the Internet will follow. However, understanding what to measure and what the measurements actually signify may not be as straightforward as it might appear. As Internet complexity increases, even well-understood protocols are behaving in pathological ways.

Vern Paxson from Lawrence Berkeley National Laboratory carried out a major study of Internet connectivity and found pathologies in routing behaviour that included loops of up to many hours duration, erroneous routing, rapid very-short-term route changes (fluttering), and infrastructure failures -- with the probablility of a user experiencing a routing pathology *doubling* between late 1994 and late 1995 to a level of 1 problem per 30 routing events [2].

Packet traffic pathologies over a large number of world-wide networks included out-of-order deliveries, packet replication and corruption, while packet loss also doubled during 1995 to around 5% of the traffic [5]. Some of the reasons for these problems are fairly well understood, such as poor implementations of protocols. Others, such as a number of simultaneous failures observed in widely-separated network components, do not have any obvious means of propagation.

However, not only are protocol pathologies increasing but as traffic is being analysed in real depth for the first time, quite unexpected and significant phenomena are being recognised, with interesting implications for metrics research.

Models, Assumptions and Surprises

The Internet is the offspring of decades of research and engineering in a number of disciplines, but ultimately much of its behaviour is defined by the models that have been used to constrain the behaviour of packet traffic. Many of the models derived from earlier telecommunications research. While the Internet was limited these models worked reasonably well, so much so that when traffic levels started exploding, it was some time before researchers noticed that the behaviour of the vast streams of traffic was surprisingly different from that predicted by the models.

Traditionally, traffic models assumed that packet flows were essentially separate and uncorrelated with each other, so that when flows are accumulated on a network their levels over time should start to approximate smooth white noise (Poisson models).

However, the reality is that traffic flows are strangely inter-related over very long timescales, which means that they never settle down to a smooth level but remain "bursty" (extremely variable) in distinctive patterns over time. This burstiness is not restricted to small-scale events but is replicated over and over no matter what the timescale, from micro-seconds to hours -- in other words, it has a self-similar structure: *Internet traffic is a fractal system.*

Real Ethernet Traffic **Poisson-model Traffic**

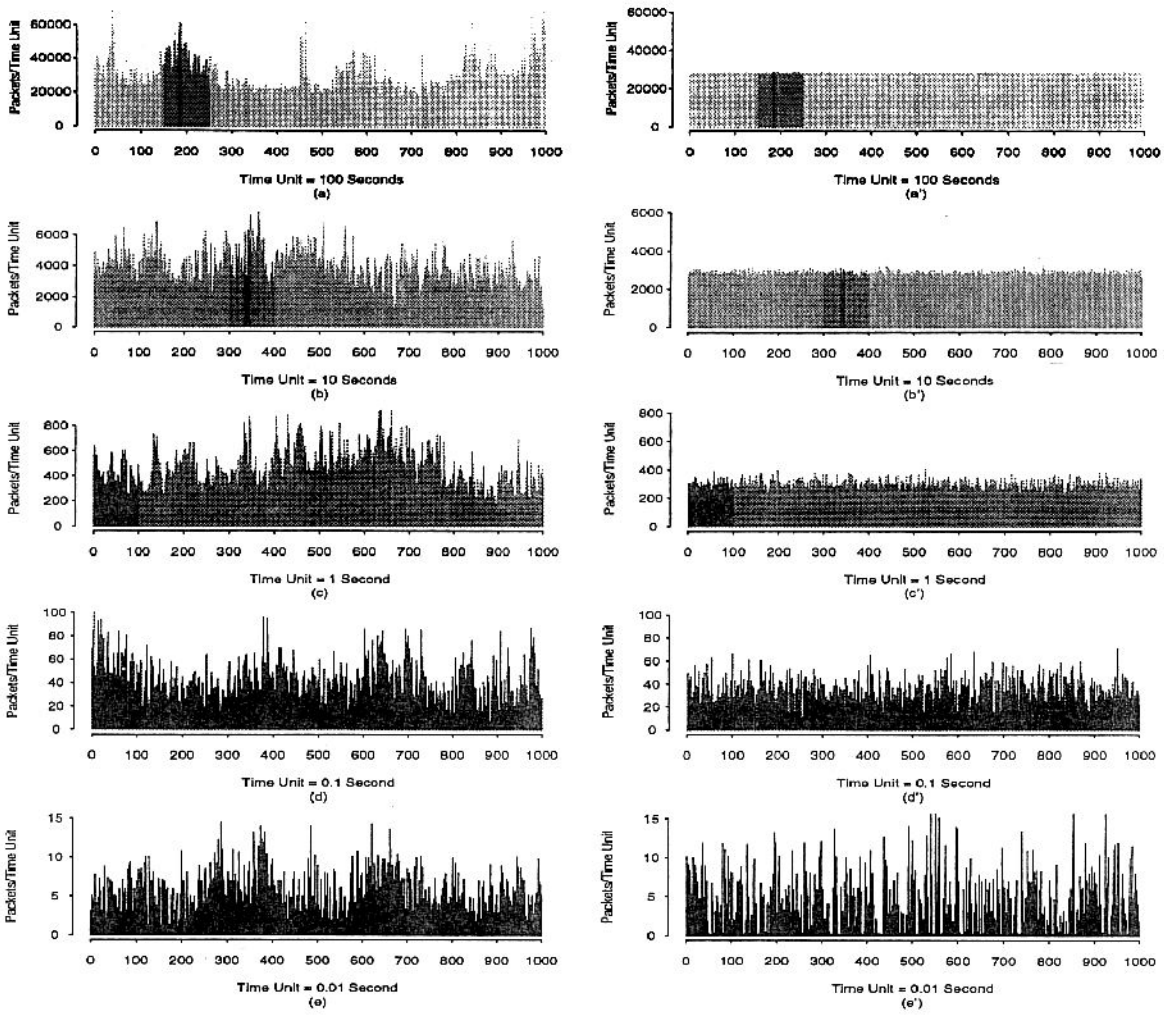

Figure 1: Right -- Synthetic traffic based on Poisson model, aggregated at different timescales. Left -- Real Ethernet traffic at the same timescales [6].

Figure 1 above is a diagram from Leland, Taqqu, Willinger and Wilson [6], who were the first to show the striking difference between Poisson-model traffic and real traffic. On the right is the Poisson-model traffic with counts of packets per hundredths of a second at the bottom, accumulated by tens up to counts per hundred second intervals at the top. Compared on the left are real traffic traces, measured on a LAN between 1989 and 1992.

The real traces, over 27 hour's worth at top left, clearly show no indication that they will settle down to the smooth Poisson-type traffic modelled on the right.   The burstiness of network traffic survives over an extremely wide range of timescales, with a self-similar structure no matter what the timescale: "*traffic spikes ride on longer-term ripples that in turn ride on still longer term swells*" [6].

What is Self-Similarity?

Over the last 15 years or so it has been recognised that the output of many systems, natural or artificial, can diverge wildly, unpredictably and unrepeatably, due to minute variations in their initial conditions -- these are the fractal, or chaotic dynamical systems. The weather is one such system: apparently insignificant differences in pressure, air-flow, heat, cold, or humidity may have such profound effects on developing conditions that even with the most powerful computers available, future weather may be only roughly predicted for a few days ahead.

Yet despite the enormous complexity of these dynamical systems, hidden within their behaviour are common characteristics that occur no matter what their variations may be. One of their most distinctive characteristics is that they show self-similarity over an enormous range of scales, so that their form over time, space, or mathematical description (for instance) has a similar pattern from the largest through to the smallest structure of the system.

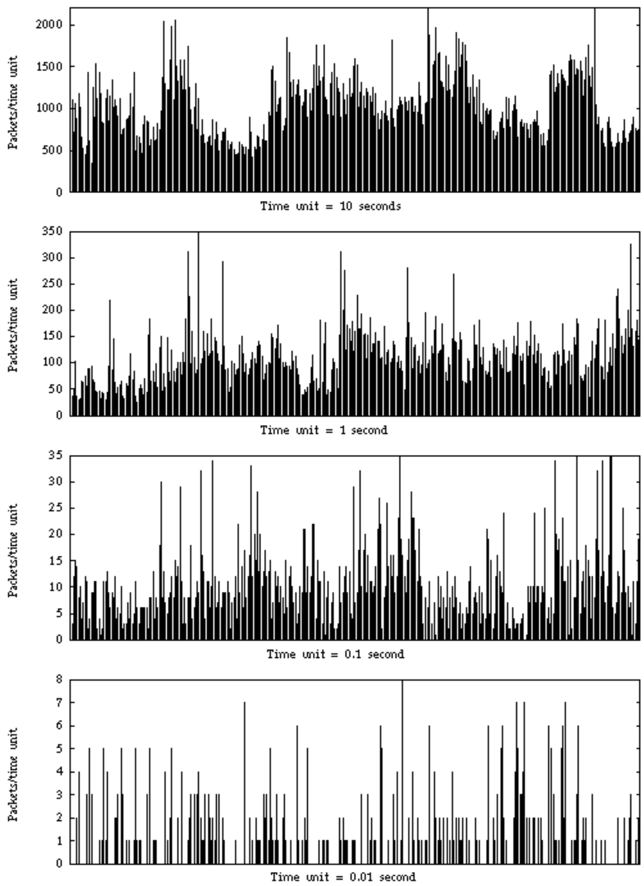

Figure 2: Packet arrivals at different timescales, from The University of Newcastle network, February 1995.

The relatively simple time-series of accumulated traffic, on the left of Figure 1 and in Figure 2, are examples of complex fractal systems occurring over time. Figure 2 is a diagram of Ethernet traffic collected on the external link of The University of Newcastle network in 1995 -- a network with a completely different mix of traffic measured several years after the study of Figure 1. Again, the self-similarity of the traffic "texture" is apparent, whether the counts are over hundredths of a second or at 10-second intervals.

Another kind of fractal is the famous Mandelbrot set (Figure 3) -- images generated by simple equations undergoing enormous numbers of computer iterations [7]. "Zooming in" on any point of the Mandelbrot set brings up very similar, but non-identical, replicas over a range of size scales limited only by the power of the computer doing the calculations. (See [8] for a fractal image generator.)

Whole new fields of understanding have arisen from applying fractal decriptions in areas as diverse as the natural sciences -- fine-particle physics (complex surfaces), ecology (population dynamics), epidemiology (disease propagation), astronomy (galactic clustering), geophysics, meteorology and fluid dynamics -- and systems that involve *human* interaction, such as stockmarket fluctuations, music of all cultures and even the flow of cars on expressways.

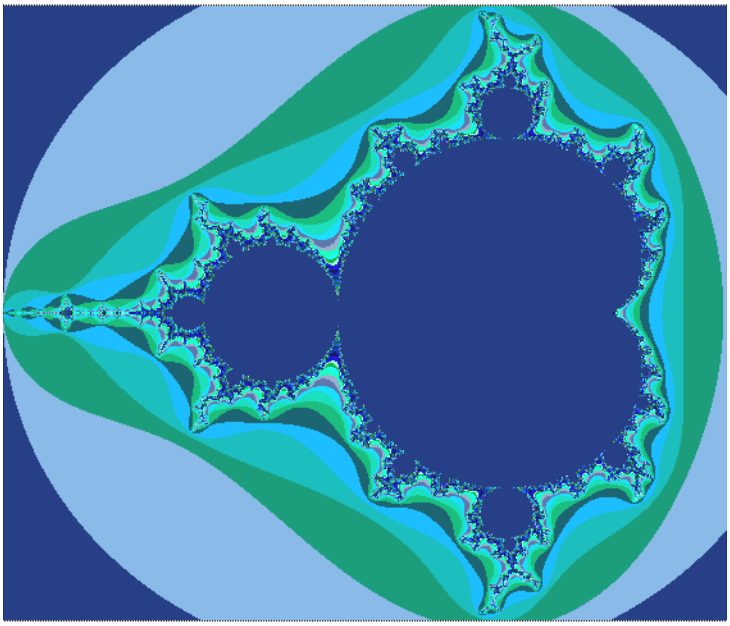

Figure 3: The Mandelbrot Set.

How Does Self-Similarity Affect Network Performance?

Under *current* traffic-handling algorithms in network hardware, which are optimised to handle traffic behaving in a Poisson-like manner, self-similarity degrades traffic throughput and increases packet loss and retransmissions, response time, and demand on buffer and bandwidth resources. UDP protocol performance deteriorates extremely quickly, with very high packet loss, as the degree of self-similarity increases. TCP protocols are more adaptable to poor conditions but still degrade smoothly in the presence of self-similarity [9]. However, queueing delay increases drastically in the presence of self-similarity under either protocol.

As self-similar models become the basis for future traffic-handling implementations, some of these drawbacks will disappear, but, due to the unpredictable nature of fractal traffic, extremes of behaviour will occur that may still exceed any possible resouce allocation. Fractal traffic will never be as well-behaved as (theoretical) Poisson traffic was expected to be.

What Causes Fractal Traffic?

So where do such strange long-term correlations between streams of traffic come from? How can randomly-generated packets, from vast numbers of different users working under wildly varying conditions, come to create patterns that can be seen at timescales as short as thousandths of a second yet still persist at timescales of hours?

The answer appears to be that packets are not randomly-generated but instead have long-timescale relationships, *long-range dependence*, between bursts of traffic. Long-range dependence causes self-similarity, which can be generated statistically by combining processes that are not normally distributed, but, instead show a long, well-populated tail of the distribution, called a "heavy tail". Figure 4 illustrates the heavy tail of the distribution of the sizes of 1.3 million files requested from a very large Web cache (myangup.connect.com.au) over an 18-hour period in August 1997.

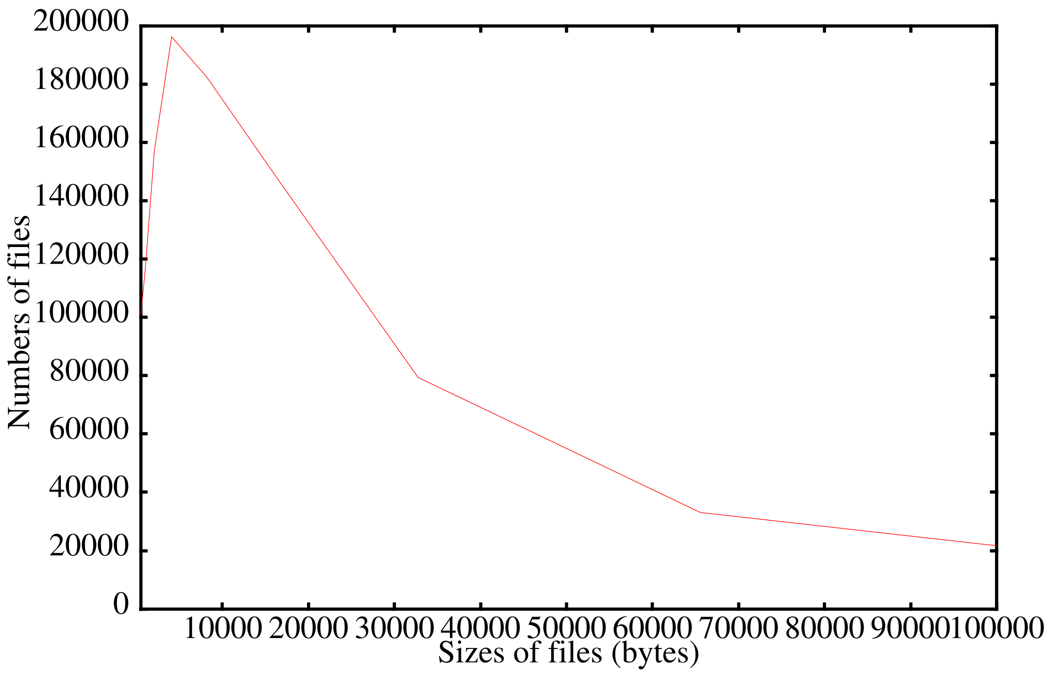

Figure 4: A heavy-tailed distribution -- the average sizes of files requested from a Web cache. The tail of the distribution actually reaches to more than 30,000,000 bytes.

Heavy-tailed activity occurs in many aspects of computing: the distribution of periods of human activity and inactivity at a computer has a heavy tail, as does the distribution of sizes of files on Unix computers [11], the sizes of files being transferred over networks via the Web, the inter-arrival times of files on a system [10], and the duration of periods of packet loss on networks [5]. All of these can set up self-similar patterns in traffic, which are highly persistent, even when aggregated with cross-traffic that has no self-similar characteristics.

There is probably no *single* cause of self-similarity in network traffic -- elements from physics to binary maths to human behaviour are able to generate it, and the greater the level of traffic, the more dominant the fractal behaviour becomes. It is an intrinsic element of all complex, chaotic systems, whether they arise from natural phenomena or from human manufacture. Perhaps it's not surprising that something that brings together human and electronic functionality as dramatically as the Internet does must result in unpredictable behaviour, but its significance is startling.

The Internet is Not Predictable

Self-similarity means that spikes or random changes in activity can not be predicted. Network traffic is a chaotic dynamic system and minute variations in one situation mean that its behaviour will diverge enormously from any other situation, no matter how similar. Chaotic, transitory elements of Internet activity are generated by the latest popular web sites, new software releases, Mars expeditions and Olympic games; faulty software and hardware, incorrect routing announcements, poor protocol implementations, inadequate bandwidth; working patterns, political conflicts, malicious attacks, honest mistakes and general stupidity.

A minor aspect of this turmoil may be determinable, such as provision of adequate bandwidth, but the vast majority of "initial conditions" are uncontrollable and their effects completely unpredictable. Ultimately, this means that the current behaviour of traffic may not be able to be used to predict the future behaviour of traffic beyond a few minutes or hours.

Vern Paxson's work [5] supports this: the best predictor he found of future network loss was whether or not the path studied had previously experienced *some* packet loss or *no* loss: this was a reasonably good indicator for several weeks into the future, probably based on the paths having roughly adequate capacity or not. But the actual observed *amount* of loss was a very poor predictor of the future loss rate beyond a few minutes or hours, and similarly the observed one-way transit time was not a predictor of the future one-way transit time.

How does this affect what we'd hoped to measure of Internet performance, if current problems or conditions do not predict future ones? At best it seems that we can only observe general changes over long baselines while nothing in the specific activity of any part of the structure can predict the future activity. This is a difficult thing to accept, we tend to assume as a matter of course that there is some connection between conditions today and those of the future. But, on the fractal Internet, this may not be so.

However, researchers are already looking at the possibility of modelling Internet traffic in chaotic terms, and using the research techniques developed in other disciplines to *describe* the intrinsic structure, the "texture", the fractal dimension of Internet traffic: and perhaps a variation in this description can tell us that conditions are changing, though it may never be able to say "in this place", "on this network", or "at this time". I have no doubt that extraordinary tools will one day arise that will use fractal techniques, but they may not be able to predict, for instance, which is an uncongested network and which is not, based on past performance.

So Internet metrics may never be able to do some of the tasks it had been hoped they would. The IPPM group began life as the IP **Provider** Metrics group, one of their aims being to establish techniques for comparing Internet service providers based on their past performance. Such comparisons *can* be done at the level of society, economics, business -- in terms of infrastructure investment, customer service, technical staff, pricing -- but, against all common sense, this may not be able to be done at the detailed traffic performance level.

So, we may still gather useful information from Internet metrics to help with immediate diagnostics, baseline information and research and development, but predicting performance is a difficult, if not impossible, task in a chaotic system. However, to the ordinary user nothing much changes -- decisions about service providers can still be made upon economic grounds.

To the service providers themselves, however, rational understanding of resources, measurements and problem-resolution, is *more dependent than ever before* upon the findings of researchers -- how those findings are implemented in software and hardware, and how much data and support researchers can obtain to study Internet phenomena. For their own sakes, ISPs must cooperate with researchers and supply resources to extend this kind of work, to collect data, to store, share and analyse it.

Such cooperative efforts have already begun, with the establishment of CAIDA, the Cooperative Association for Internet Data Analysis [12], ISMA, Internet Statistics and Metrics Analysis workshops [13], and the Internet Traffic Archive [14], but the recognition in Australia of the necessity of cooperative data collection and analysis between ISPs and research organisations is lagging far behind the rest of the world.

The discovery of self-similarity in network traffic has had profound implications for Internet performance. Research has already uncovered other strange phenomena, such as periodic resonances and synchronisations between devices [15], which will have enormous implications for an Internet in which periodic activity and protocols are on the increase.

Who knows, or can even imagine, what remains to be discovered? The Internet *is* the interface of the extraordinary developments in modern science and the unpredictability of human response: we must measure it well.

References

[1] CAIDA Measurement Tools Taxonomy, **http://www.nlanr.net/Caidants/meastools.html**

[2] Paxson, V., End-to-End Routing Behavior in the Internet. ACM SIGCOMM '96, August 1996, Stanford, CA. **ftp://ftp.ee.lbl.gov/papers/routing.SIGCOMM.ps.Z**

[3] IPPM, the Internet Protocol Performance Metrics group, **http://io.advanced.org/IPPM/**

[4] Paxson, V., Towards a Framework for Defining Internet Performance Metrics. LBNL-38952, June 20, 1996. **ftp://ftp.ee.lbl.gov/papers/metrics-framework-INET96.ps.Z**

[5] Paxson, V., End-to-End Internet Packet Dynamics, ACM SIGCOMM '97, September 1997, Cannes, France. **ftp://ftp.ee.lbl.gov/papers/vp-pkt-dyn-sigcomm97.ps.Z**

[6] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," IEEE/ACM Transactions on Networking, 2(1), pp 1-15, February 1994.

[7] B. Mandelbrot, The Fractal Geometry of Nature, Freeman, New York, 1983.

[8] xfractint, a fractal image generator, **http://spanky.triumf.ca/www/fractint/getting.html**

[9] Kihong Park, Gi Tae Kim, and Mark E. Crovella, "The Effects of Traffic Self-Similarity on TCP Performance," Boston University Computer Science Department, April 1996, **http://cs-www.bu.edu/faculty/crovella/paper-archive/selfsim-tcp.ps**

[10] Kihong Park, Gi Tae Kim, and Mark E. Crovella, "On the Relationship Between File Sizes, Transport Protocols, and Self-Similar Network Traffic", **http://cs-www.bu.edu/faculty/crovella/paper-archive/files-protocols/TR-96-016.ps**

[11] Unix File Size Survey, **http://www.base.com/gordoni/ufs93.html**

[12] CAIDA, the Cooperative Association for Internet Data Analysis, **http://www.caida.org**

[13] ISMA, Internet Statistics and Metrics Analysis, **http://www.nlanr.net/ISMA/isma97_report.html**

[14] The Internet Traffic Archive, **http://www.acm.org/sigcomm/ITA/**

[15] S. Floyd and V. Jacobson, "The Synchronization of Periodic Routing Messages," IEEE/ACM Transactions on Networking, 2(2), pp 122-136, April 1994.