LBNL-40319
UCB//CSD-97-945

Measurements and Analysis of End-to-End Internet Dynamics

Vern Paxson
Ph.D. Thesis

Computer Science Division
University of California, Berkeley

and

Information and Computing Sciences Division
Lawrence Berkeley National Laboratory
University of California
Berkeley, CA 94720

April, 1997

This work was supported by the Director, Office of Energy Research, Scientific Computing Staff, of the United States Department of Energy under Contract No. DE-AC03-76SF00098.

Measurements and Analysis of End-to-End Internet Dynamics

by

Vern Edward Paxson

B.S. (Stanford University) 1985
M.S. (University of California, Berkeley) 1991

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

GRADUATE DIVISION
of the
UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Prof. Domenico Ferrari, Chair
Prof. Michael Luby
Prof. John Rice

1997

Measurements and Analysis of End-to-End Internet Dynamics

Abstract

Measurements and Analysis of End-to-End Internet Dynamics

by

Vern Edward Paxson
Doctor of Philosophy in Computer Science

University of California at Berkeley

Prof. Domenico Ferrari, Chair

Accurately characterizing end-to-end Internet dynamics—the performance that a user actually obtains from the lengthy series of network links that comprise a path through the Internet—is exceptionally difficult, due to the network's immense heterogeneity. It can be impossible to gauge the generality of findings based on measurements of a handful of paths, yet logistically it has proven very difficult to obtain end-to-end measurements on larger scales.

At the heart of our work is a "measurement framework" we devised in which a number of sites around the Internet host a specialized measurement service. By coordinating "probes" between pairs of these sites we can measure end-to-end behavior along $O(N^2)$ paths for a framework consisting of N sites. Consequently, we obtain a superlinear scaling that allows us to measure a rich cross-section of Internet behavior without requiring huge numbers of observation points. 37 sites participated in our study, allowing us to measure more than 1,000 distinct Internet paths.

The first part of our work looks at the behavior of end-to-end routing: the series of routers over which a connection's packets travel. Based on 40,000 measurements made using our framework, we analyze: routing "pathologies" such as loops, outages, and flutter; the stability of routes over time; and the symmetry of routing along the two directions of an end-to-end path. We find that pathologies increased significantly over the course of 1995, indicating that, by one metric, routing *degraded* over the year; that Internet paths are heavily dominated by a single route, but that routing lifetimes range from seconds to many days, with most lasting for days; and that, at the end of 1995, about half of all Internet paths included a major routing asymmetry.

The second part of our work studies end-to-end Internet packet dynamics. We analyze 20,000 TCP transfers of 100 Kbyte each to investigate the performance of both the TCP endpoints and the Internet paths. The measurements used for this part of our study are much richer than those for the first part, but require a great degree of attention to issues of *calibration*, which we address by applying *self-consistency checks* to the measurements whenever possible. We find that packet filters are capable of a wide range of measurement errors, some of which, if undetected, can significantly taint subsequent analysis. We further find that network clocks exhibit adjustments and skews relative to other clocks frequently enough that a failure to detect and remove these effects will likewise pollute subsequent packet timing analysis.

Using TCP transfers for our network path "measurement probes" gains a number of advantages, the chief of which is the ability to probe fine time scales without unduly loading the network. However, using TCP also requires us to accurately distinguish between connection dy-

namics due to the behavior of the TCP endpoints, and dynamics due to the behavior of the network path between them. To address this problem, we develop an analysis program, `tcpanaly`, that has coded into it knowledge of how the different TCP implementations in our study function. In the process of developing `tcpanaly`, we thus in tandem develop detailed descriptions of the performance and congestion-avoidance behavior of the different implementations. We find that some of the implementations suffer from gross problems, the most serious of which would devastate overall Internet performance, were the implementations ubiquitously deployed.

With the measurements calibrated and the TCP behavior understood, we then can turn to analyzing the dynamics of Internet paths. We first need to determine a path's *bottleneck bandwidth*, meaning the fastest transfer rate the path can sustain. Knowing the bottleneck bandwidth then lets us determine which packets a sender transmits must necessarily *queue* behind their predecessors, due to the load the sender imposes on the path. This in turn allows us to determine which of our probes are perforce *correlated*. We identify several problems with the existing bottleneck estimation technique, "packet pair," and devise a robust estimation algorithm, PBM ("packet bunch modes"), that addresses these difficulties. We calibrate PBM by gauging the degree to which the bottleneck rates it estimates accord with known link speeds, and find good agreement. We then characterize the scope of Internet path bottleneck rates, finding wide variation, not infrequent asymmetries, but considerable stability over time.

We next turn to an analysis of packet loss along Internet paths. To do so, we distinguish between losses of "loaded" data packets, meaning those which necessarily queued behind a predecessor at the bottleneck; "unloaded" data packets, which did not do so; and the small "acknowledgement" packets returned to a TCP sender by the TCP receiver. We find that network paths are well characterized by two general states, "quiescent," in which no loss occurs, and "busy," in which one or more packets of a connection are lost. The prevalence of quiescent connections remained about 50% in both our datasets, but for busy connections, packet loss rates increased significantly over the course of 1995. We further find that loss rates vary dramatically between different regions of the network, with European and especially trans-Atlantic connections faring much worse than those confined to the United States.

We also characterize: loss symmetry, finding that loss rates along the two directions of an Internet path are nearly uncorrelated; loss "outages," finding that outage durations exhibit clear Pareto distributions, indicating they span a large range of time scales; the degree to which a connection's loss patterns predict those of future connections, finding that observing quiescence is a good predictor of observing quiescence in the future, and likewise for observing a busy network path, but that the proportion of lost packets does not well predict the future proportion; and the efficacy of TCP implementations in dealing with loss efficiently, by retransmitting only when necessary. We find that most TCPs retransmit fairly efficiently, and that deploying the proposed "selective acknowledgement" option would eliminate almost all of their remaining unnecessary retransmissions. However, some TCPs incorrectly determine how long to wait before retransmitting, and these can suffer large numbers of unnecessary retransmissions.

We finish our study with a look at variations in packet transit delays. We find great "peak-to-peak" variation, meaning that maximum delays far exceed minimum delays. Delay variations along the two directions of an Internet path are only lightly correlated, but correlate well with loss rates observed in the same direction along the path. We identify three types of "timing compression," in which packets arrive at their receiver spaced more closely together than when originally transmitted. The prevalence of none of the three is such as to significantly perturb network performance, but all three occur frequently enough to require judicious filtering by network measurement procedures to avoid deriving false timing conclusions.

We then look at the question of the time scales on which most of a path's queueing variations occur. We find that, overall, most variation occurs on time scales of 100–1000 msec, which means that transport connections might effectively adapt their transmission to the variations, but only if they act quickly. However, as with many Internet path properties, we find wide ranges of behavior, with not insignificant queueing variations occurring on time scales as small as 10 msec and as large as one minute.

The last aspect of packet delay variations we investigate is the degree to which it reflects an Internet path's *available bandwidth*. We show that the ratio between the delay variations packets incur due to their connection's own loading of the network, versus the total delay variations incurred, correlates well with the connection's overall throughput. We further find that Internet paths exhibit wide variation in available bandwidth, ranging from very little available to virtually all. The degree of available bandwidth diminished markedly over the course of 1995, but, as for packet loss rates, we also find sharp geographic differences, so the overall trend cannot be summarized in completely simple terms. Finally, we investigate the degree to which the available bandwidth observed by a connection accurately predicts that observed by future connections, finding that the predictive power is fairly good for time scales of minutes to hours, but diminishes significantly for longer time periods.

We argue that our work supports several general themes:

- The N^2 scaling property of our measurement framework serves to measure a sufficiently diverse set of Internet paths that we might plausibly interpret the resulting analysis as accurately reflecting general Internet behavior.

- To cope with such large-scaled measurements requires attention to calibration using self-consistency checks; robust statistics to avoid skewing by outliers; and automated "micro-analysis," such as that performed by `tcpanaly`, that we might see the forest as well as the trees.

- With due diligence to remove packet filter errors and TCP effects, TCP-based measurement provides a viable means for assessing end-to-end packet dynamics.

- We find wide ranges of behavior, so we must exercise great caution in regarding any aspect of packet dynamics as "typical."

- Some common assumptions such as in-order packet delivery, FIFO bottleneck queueing, independent loss events, single congestion time scales, and path symmetries are all sometimes violated.

- The combination of path asymmetries and reverse-path noise render sender-only measurement techniques markedly inferior to those that include receiver-cooperation.

Finally, we believe an important aspect of this work is how it might contribute towards developing a "measurement infrastructure" for the Internet: one that proves ubiquitous, informative, and sound.

To Lindsay —

For making it both possible
and worthwhile

— *with all my love*

Contents

List of Figures

List of Tables

Acknowledgements

This work has its roots in the teaching, help, patience, and inspiration of a great number of people, to whom I wish to express my heartfelt gratitude.

Simply put, Van Jacobson is the reason I have studied networking; the reason I embarked on this study; and the reason I had faith that the work would, with sufficient diligence, yield a host of new insights. I am delighted that, having known him for nearly twenty years, I still find he has much to teach me.

Likewise, this work drew inspiration and invaluable support from Domenico Ferrari. The energy and respect that he affords to both his students' efforts, and to his students themselves, has made it a privilege to be advised by him.

I have also been delighted to have Sally Floyd as my mentor, colleague, and friend. She has listened to countless half-baked ideas of how to analyze and interpret various measurements, and has always patiently separated the promising from the harebrained. This calibration of ideas, and her suggestions on how to then pursue the more promising ones, has proved invaluable for fostering my sense of how to conduct sound research.

A number of others played major roles in shaping this work. I would particularly like to thank John Rice and Mike Luby for their industrious efforts in serving on my dissertation committee, which led to the work being much more solid than it would otherwise have been.[1]

My heartfelt thanks to Greg Minshall, for his detailed, insightful comments on nearly every page of the work (and for his willingness to burn an entire Friday evening discussing some of them); and to Amit Gupta, John Hawkinson, Kurt Lidl, Craig Partridge, and anonymous SIGCOMM and *IEEE/ACM Transactions on Networking* referees, all of whom contributed very helpful comments on earlier versions of the work.

I would like to also thank my colleagues at the Network Research Group: Kevin Fall, Craig Leres, and Steve McCanne, for their much appreciated ideas, support, and feedback.

Special thanks to Kathryn Crabtree, for her untiring help in surmounting innumerable administrative hurdles along the dissertation trail. She is an invaluable asset to UCB computer science.

This work would not have been possible without the efforts of the many volunteers who installed the Network Probe Daemon at their sites. In the process they endured debugging headaches, `inetd` crashes, software updates, and a seemingly endless stream of queries from me regarding their site's behavior. I am indebted to:

Guy Almes and Bob Camm (`adv`);
Jos Alsters (`unij`);
Jean-Chrysostome Bolot (`inria`);
Hans-Werner Braun, Kim Claffy, and Bilal Chinoy (`sdsc`);
Randy Bush (`rain`);
Jon Crowcroft and Atanu Ghosh (`ucl`);
Peter Danzig and Katia Obraczka (`usc`);
Mark Eliot (`sri`);
Robert Elz (`austr`);

[1] Particular thanks to Mike for throwing down the glove, and for knowing which glove to use.

Teus Hagen (`oce`);
Steinar Haug and Håvard Eidnes (`sintef1`, `sintef2`);
John Hawkinson (`near` and `panix`);
TR Hein (`xor`);
Tobias Helbig and Werner Sinze (`ustutt`);
Paul Hyder (`ncar`);
Alden Jackson (`sandia`);
Kate Lance (`austr2`);
Craig Leres (`lbl`);
Kurt Lidl (`pubnix`);
Peter Linington, Alan Ibbetson, Peter Collinson, and Ian Penny (`ukc`);
Steve McCanne (`lbli`);
John Milburn (`korea`);
Walter Mueller (`umann`);
Evi Nemeth, Mike Schwartz, Dirk Grunwald, Lynda McGinley (`ucol`, `batman`);
Francóis Pinard (`umont`);
Jeff Polk and Keith Bostic (`bsdi`);
Todd Satogata (`bnl`);
Doug Schmidt and Miranda Flory (`wustl`);
Sorell Slaymaker and Alan Hannan (`mid`);
Don Wells and Dave Brown (`nrao`);
Gary Wright (`connix`);
John Wroclawski (`mit`);
Cliff Young and Brad Karp (`harv`); and
Lixia Zhang, Mario Gerla, and Simon Walton (`ucla`).

I am likewise indebted to Keith Bostic, Evi Nemeth, Rich Stevens, George Varghese, Andres Albanese, Wieland Holfelder, and Bernd Lamparter for their invaluable help in recruiting NPD sites. Thanks, too, to Peter Danzig, Jeff Mogul, and Mike Schwartz for feedback on the design of NPD.

This work also benefited from discussions with Guy Almes, Tom Anderson, Robert Elz, Teus Hagen, John Krawczyk, Kate Lance, Dun Liu, Paul Love, Jamshid Mahdavi, Matt Mathis, Dave Mills, Pravin Varaiya, Curtis Villamizar, and Walter Willinger.

A preliminary analysis of the $\mathcal{R}_1$ routing dataset was done by Mark Stemm and Ketan Patel.

Often to understand the behavior of particular routers or to determine their location, I asked personnel from the organization responsible for the routers. I was delighted at how willing they were to help, and in this regard would like to acknowledge:

Vadim Antonov, Tony Bates, Michael Behringer, Per Gregers Bilse, Bjorn Carlsson, Peggy Cheng, Guy Davies, Sean Doran, Bjorn Eriksen, Amit Gupta, Tony Hain, John Hawkinson again!, Susan Harris, Ittai Hershman, Kevin Hoadley, Scott Huddle, James Jokl, Kristi Keith, Harald Koch, Craig Labovitz, Tony Li, Martijn Lindgreen, Ted Lindgreen, Dan Long, Bill Manning, Milo Medin, Keith Mitchell, Roderik Muit, Chris Myers, Torben Nielsen, Richard Nuttall, Mark Oros, Michael Ramsey, Juergen Rauschenbach, Douglas Ray, Brian Renaud, Jyrki Soini, Nigel Titley, Paul Vixie, and Rusty Zickefoose.

Finally, this work would never have been realized without the ongoing support provided by the Lawrence Berkeley National Laboratory. I am deeply grateful. In particular, I would like to thank Stu Loken and Ed Theil for their efforts and encouragement.

This work was supported by the Director, Office of Energy Research, Scientific Computing Staff, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.