Growth Trends in Wide-Area TCP Connections

Vern Paxson
Lawrence Berkeley Laboratory and
EECS Division, University of California, Berkeley
1 Cyclotron Road
Berkeley, CA 94720
vern@ee.lbl.gov

November 16, 1993

Abstract

We analyze the growth of a large research laboratory's wide-area TCP connections over a period of three years. Our data consisted of seven month-long traces of all TCP connections made between the site and the rest of the world. We find that many TCP protocols exhibited exponential growth in the number of connections made and bytes transferred, even though the number of hosts at the site only grew linearly. The exponential growth of most of the major TCP protocols began tapering off with the final datasets, while relatively new information-retrieval protocols such as Gopher and World-Wide Web exhibited explosive growth during the same time. Our study also found that individual users greatly affected the site's traffic profile by the inadvertent or casual initiation of multiple, periodic wide-area connections; that exponential growth is fed in part by more users "discovering" the Internet and in part by existing users increasingly incorporating use of the Internet into their work patterns; and that wide-area traffic geography is diverse and dynamic.

1 Introduction

To properly design future networks, we need a thorough understanding of how network use changes and grows with time. Previous studies [Kleinrock76, Quarterman90, Lottor92, Merit93, Adams93] have found that many aspects of network use grow exponentially with time, at least until reaching the carrying capacity of the network. These studies all summarize network "backbone" use of some sort. Because the number of host computers connected to these backbones also grows exponentially, we cannot readily extrapolate the growth of network use by individual sites from the backbone growth: no site-growth, linear site-growth, or exponential site-growth are all consistent with exponential backbone growth.

To our knowledge, no studies have appeared tracking the evolution of a site's wide-area network use over time.[1] In this paper we analyze seven one-month traces of a single site's wide-area TCP connections, spanning altogether three years and nearly three million connections. The key question for such a study is: Does traffic at individual sites also grow exponentially, and if so, what factors contribute to the growth?

Such questions cannot be answered by studying a single site, as there is no foundation for assuming that such a site is representative of Internet sites as a whole. But a single site study provides a beginning for exploring the questions of site growth more fully. With this limitation in mind, Table 1 summarizes our major findings, which we develop in the body of the paper. The first finding states that the site traffic did indeed grow exponentially, exceeding the rate at which the site added computing resources, but the second finding tempers this result by suggesting that in many areas the growth will begin tapering off. The third finding indicates that new protocols may step in to take up the slack as the growth of older protocols diminishes. The remaining findings address the roots of the growth: part is due to increasing inadvertent or casual use of the network; part is due to an increasing number of existing computers using the Internet; part is due to increasing use of the Internet by individual users; and part is due to taking advantage of the widening Internet connectivity.

In the next section we review existing statistics and studies of network growth, which show that network traffic generally grows exponentially with time, at least until the network carrying capacity is reached. We then describe how we captured and reduced the data used in our study. The following sections address the points made in Table 1: the overall growth in the site's wide-area traffic; the appearance of periodic traffic; the growth in network use by individual computers or users; and the changing geographic profile of the traffic.

The final section summarizes the implications and limitations of our results.

[1] The closest available statistics are those published by Merit, Inc. [Merit93], reporting aggregate bytes transferred into and out of each NSFnet stub network.

At a site where the number of Internet hosts grew linearly, wide-area TCP traffic for a number of protocols grew exponentially both in the number of connections made and (at even higher rates) the amount of data transferred. For example, *smtp* connections grew 73%/year, and *ftp* data bytes grew 135%/year.
The most recent datasets indicated decreasing growth in connection and transfer rates for a number of TCP protocols, consistent with exponential growth becoming *logistic* due to reaching some sort of carrying capacity.
Use of information-retrieval protocols such as Gopher and World-Wide Web grew extremely rapidly and will soon outweigh *telnet* traffic in their overall impact. In general, new protocols such as these and *X11* can exhibit explosive growth when first introduced, significantly affecting a site's traffic profile.
The site's traffic profile was also affected by periodic Internet connections created either unknowingly or quite casually through the use of background scripts.
The site's exponential growth in TCP connections was fed in part by new users "discovering" the Internet, and in part by existing users increasingly incorporating use of the Internet into their work patterns.
The geographic profile of wide-area traffic is diverse and dynamic. While at any given time particular states and countries dominate the traffic's geography, this profile changes greatly over time.

Table 1: Major Findings

2 Prior Work

Many existing collections of network growth statistics show that network traffic grows exponentially with time:[2]

- Kleinrock [Kleinrock76] includes a discussion of ARPAnet growth from October, 1971, through March, 1975. For the first half of this period, traffic grew exponentially, rising from 10^5 packets/day in October, 1971, to about $3 \cdot 10^6$ packets/day in August, 1973. This increase corresponds to 630% growth per year. After August, 1973, though, traffic growth leveled off, with about $4.5 \cdot 10^6$ packets/day during March, 1975; 27% growth per year. The declining growth was probably due to approaching the carrying capacity of the network, which was 10^7 packets/day.

- Lottor [Lottor92] reports on how the number of Internet hosts grew from August, 1981, up till January, 1992. While the Internet grew exponentially over this period, the rate lessened with time. Fitting an exponential to the entire body of data yields a growth rate of 139%/year. During the last three years of the study, growth diminished to 99%/year (from 80,000 at the beginning of 1989 to 727,000 at the beginning of 1992).

  More recent data [Lottor93] shows this growth rate leveling off further still, with 2,056,000 hosts at October, 1993; the most recent three years of growth (October, 1990, through October, 1993) correspond to an annual growth rate of 85%/year.

- Statistics available from Merit, Inc. [Merit93], show that NSFnet backbone traffic has been growing exponentially, from $1.27 \cdot 10^{12}$ bytes/month in March, 1991 (the first month for which data is available), to $7.02 \cdot 10^{12}$ bytes/month in September, 1993. This growth corresponds to an increase of 106%/year. During this same time, the number of networks connected to the NSFnet has risen exponentially from 2,501 to 17,979, a growth rate of 112%/year. Unfortunately, the number of different *hosts* making connections over the backbone is not available.

  The NSFnet T1 backbone traffic is further studied by Claffy et. al. [CPB93], who find that the number of bytes traversing the T1 backbone grew quadratically between June, 1988, and June, 1992, though this growth trend is at least partially influenced by the traffic switching over to the T3 backbone later in the study period.

- Statistics for USENET network news traffic [Adams93, Quarterman90] show steady exponential growth since October, 1984. Traffic volume grew from $4 \cdot 10^5$ bytes/day to $5.8 \cdot 10^7$ bytes/day; these totals exclude news article headers, which add a fairly constant 25% overhead to the volume. Over this nine-year period, traffic grew at a steady rate of about 77% a year, as shown in Figure 1. This constant growth rate sustained over nearly a decade is striking.

[2] Plots of many of the statistics discussed in this section can be found in [Quarterman93].

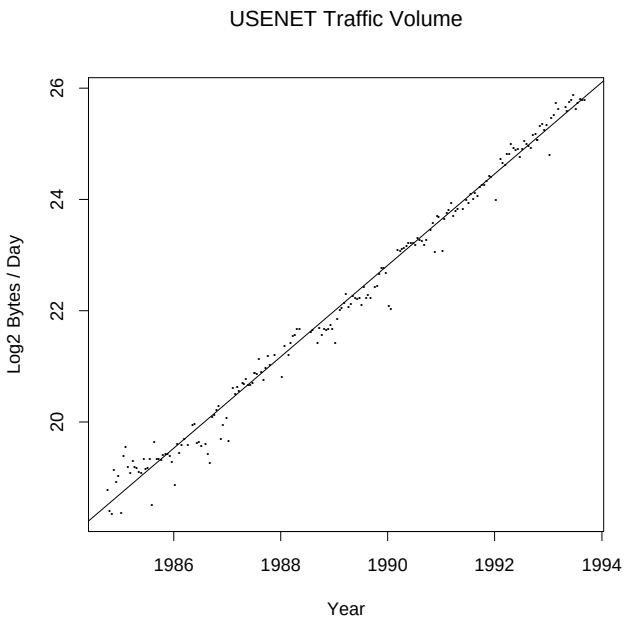

Figure 1: Exponential Growth in USENET Traffic Volume

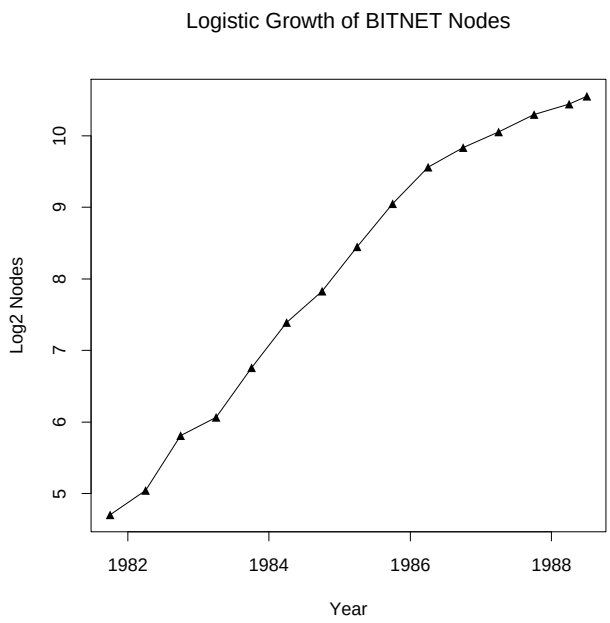

Figure 2: Logistic Growth in BITNET Nodes

During the same interval, the number of sites carrying USENET news also grew exponentially, from 520 to 31,747, but at a slower rate of 61%/year, consistent with an increase of about 10%/year in the news traffic generated by each site.

- Monthly statistics for the number of hosts in the European RIPE network [Terpstra93] show exponential growth of 104%/year for the period from January, 1992, through October, 1993. Prior to that time (the data go back to October, 1990) the growth is considerably more rapid, about 261%/year, but also much more fitful.

Exponential traffic growth must at some point slow down. The Kleinrock study shows such a turndown clearly, as does the declining rate of the Internet host count (but not, at least yet, the USENET or NSFnet statistics). Some researchers propose modeling network traffic growth with a *logistic* distribution[3] rather than an exponential, to take into account the carrying capacity of the network [Solensky92] and saturation of demand [Gurbaxani90]. In this regard, we should be alert to the presence of inflection points in growth curves, as these may correspond to approaching the current carrying capacity or demand limits.

Figure 2 shows the number of BITNET nodes over a seven year period, taken from [Gurbaxani90]. The X-axis gives the year and the Y-axis $\log_2$ of the number of BITNET nodes. The inflection point at 1986 shows the characteristic slow-down of logistic growth as network demand or capacity begins to saturate.

3 The Site Data

For our study we collected seven month-long traces of all wide area TCP connections between the Lawrence Berkeley Laboratory (LBL) and the rest of the world, taken between November 1990 and September 1993. The acquisition of the traces, which we summarize here, is discussed in greater detail in [Paxson93].

The University of California operates LBL, which is dedicated to basic research, under contract with the U.S. Department of Energy. During the time period covered by this study, LBL as an institute experienced little growth: staffing levels grew from 2,558 full-time equivalents to 2,681, an increase of about 2%/year, and funding rose from $256 million/year up to $290 million and back down to $262 million.

The traces were captured using the *tcpdump* packet capture tool [JLM89] running the Berkeley Packet Filter [MJ93]. We used a *tcpdump* filter to capture only those TCP packets with SYN, FIN, or RST flags in their headers, greatly reducing the volume and rate of data (but at the cost of losing any per-packet information). From SYN and FIN packets one can derive the connection's TCP protocol, connection duration, number of bytes transferred in each direction (excluding TCP/IP overhead), participating hosts, and starting time.[4]

[3] The logistic distribution is defined by

$$F(x;\alpha,\beta) = 1/(1+\exp(-(x-\alpha)/\beta)).$$

[4] In principle we could derive the same information using RST packets instead of FIN packets, but we found that often the sequence numbers associated with RST packets were erroneous. Since we could not derive reliable

Dataset	Packets (days)	Starting Time	Ending Time	Drops
LBL-1	124M (36)	Thu 01Nov90	Sat 01Dec90	$0 + 0$
LBL-2	?	Thu 28Feb91	Sat 30Mar91	$0 + ?$
LBL-3	207M (47)	Thu 07Nov91	Sat 07Dec91	$9 + 24$
LBL-4	210M (36)	Thu 19Mar92	Sat 18Apr92	$6 + 233$
LBL-5	337M (35)	Thu 24Sep92	Sat 23Oct92	$8 + 1808$
LBL-6	447M (31)	Wed 24Feb93	Fri 26Mar93	$3 + 0$
LBL-7	560M (32)	Thu 16Sep93	Sat 15Oct93	$0 + 7959$

Table 2: Summary of Datasets

Table 2 summarizes the datasets. The second column gives the total number of network packets received by the kernel for each dataset, along with the number of days spanned by the entire trace.[5] Each dataset was then trimmed to span exactly 30 days, beginning at midnight on a Thursday and ending at midnight on a Saturday (i.e., just after 11:59PM Friday night), except for LBL-6, which begins on a Wednesday and ends on a Friday midnight. The "Drops" column gives the drop count reported by the Ethernet driver followed by the drop count reported by *tcpdump*; this last value represents dropped SYN/FIN/RST packets. As noted in [Paxson93], the increasing number of dropped packets in the LBL-4, LBL-5, and LBL-7 appear correlated with "RST storms"—periods during which two hosts furiously exchange RST packets.

We reduced the traces by extracting only full TCP connections, i.e., two exchanged SYN packets followed by two exchanged FIN packets. We discarded connections that failed to include both pairs of SYN and FIN packets.

Because of the close administrative ties and the short, high-speed link between LBL and the University of California at Berkeley (UCB), traffic between the two institutes is likely to be atypical wide-area traffic, so we also removed these connections (comprising 20-40% of all connections). We made one exception, keeping LBL-UCB *nntp* traffic; by including all of LBL's *nntp* peers, we can study the net inflow and outflow of network news, the total rates of which should be unaffected by the close ties between LBL and UCB.

We furthermore removed connections that transferred no data, and those that purported to transfer impossibly large amounts of data[6], attributing the latter to protocol errors. Details of the removed connections are given in [Paxson93]; the number removed was almost always less than 1% of a protocol's connections.

Finally, to simplify the analysis and presentation of the data, we aggregated into an "other" protocol those connections that never accounted for more than 500 connections during a month.

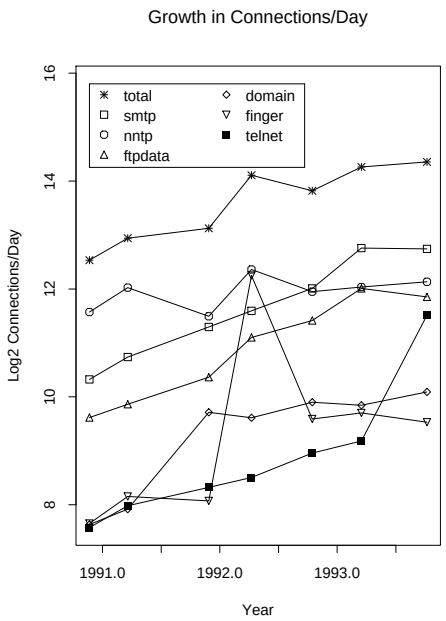

Figure 3: Daily Connection Rate for Most Popular Protocols

4 Growth in Traffic Volume

In this section we discuss the overall growth in traffic volume for the different TCP protocols. Figure 3 shows the daily connection rate for the six most popular protocols, along with the total connection rate for all of the protocols. The X-axis gives the year, and the Y-axis $\log_2$ of the average number of connections made each day.

The varying spacing between the lines of the different protocols immediately shows that the traffic "mix" varies considerably with time. This observation complements that made in [DJCME92] and [Paxson93] that traffic mix also varies significantly from site to site.

Addressing now each protocol in turn:

- The considerable variation in the *nntp* (network news) connection rate is due to at least four factors other than just growth in USENET traffic. First, LBL has two *nntp* servers, one primary and the other secondary. These servers sometimes independently connect with their outside peers to receive news, and sometimes receive their news from the other LBL server; the former results in two WAN connections per incoming batch of news, the latter only one. Hence the intra-LBL news-propagation dynamics considerably influences LBL's external *nntp*

byte counts from RST-terminated connections, we excluded them from subsequent analysis.

[5] The statistics missing for the LBL-2 dataset are due to abnormal termination of the tracing program; this termination, however, did not imply any extra-ordinary loss of packets during the 30-day study period.

[6] Typically close to 2^{32} bytes; after removing these connections, the largest connection remaining in any of the datasets was 300 MB.

We removed one *X11* connection from LBL-7 that purported to transfer 1.1GB in 9 days; it may have been legitimate, but the next largest *X11* connection in that dataset was only 42MB, making the 1.1GB transfer a huge spike.

connection rate. For example, the proportion of the *nntp* connections involving LBL's primary server varied from 58% (LBL-3) to 86% (LBL-4).

The second factor is that the rate at which new news arrives depends heavily on the configuration of the remote *nntp* peers. For example, the proportion of connections between LBL and its UCB peers varied from 37% (LBL-1) to 57% (LBL-5).

The third factor is that LBL *nntp* servers keep their connection to their peers open for one minute after they last sent data to the peer, in the hopes that new news will arrive in the interim and can be propagated without requiring a new connection set-up. Thus if new news tends to arrive within a minute of any earlier news, the total number of connections will be lower (other things being equal). One measure of this tendency is the proportion of *nntp* connections that were *inbound* (originated by a remote peer), which will be high if LBL's servers tend to coalesce multiple outbound news batches into a single connection, and low if when propagating news LBL's servers tend to use multiple connections. This rate varied from a low of 15% (LBL-3) to a high of 45% (LBL-2).

The final factor is the proportion of "failed" *nntp* connections. As explained in [Paxson93], an *nntp* connection during which the originator transfers exactly 6 bytes corresponds to a "failure" in the sense that the responder stated it was unable to accept news at that time. Such failures are likely to lead to repeated connections as the originator later tries to again propagate the news to the responder. The failure rate fell steadily from 38% (LBL-1) to 3-6% (LBL-4 through LBL-7).

- *smtp* (electronic mail) connections have grown at an approximately exponential rate. A line fitted in a least-squares sense to all of the datasets except LBL-1 and LBL-6 (which are not collinear with the others) gives growth of 73%/year.

- *ftpdata* traffic (corresponding to the data-transfer and directory-listing portion of an *ftp* file-transfer session) is quite noisy, in part due to the presence of weather-map scripts (discussed in Section 5 below). If we remove weather-map traffic, a least-squares fit to the first three datasets gives exponential growth of 61%/year, and for the last three datasets, 59%/year, with a substantial discontinuity at LBL-4.

- *domain* (Internet domain name service) traffic increased greatly between LBL-2 and LBL-3, and otherwise shows fairly flat growth. The large increase is due to the addition of a probably-misconfigured remote peer, as the LBL-3 through LBL-7 traffic is all dominated by connections to just one remote site.

- *finger* (remote user lookup) traffic shows a great deal of variation, including a huge spike in the LBL-4 dataset that returned to an elevated level in LBL-5. The causes for these increases are discussed in Section 5 below.

  We also investigated whether the higher level of *finger* traffic might be due to use of resource discovery tools such as *Netfind* [ST91]. For each dataset we checked the *finger* connections of the ten most popular remote (non-LBL) hosts to see with how many different LBL hosts they connected. We deemed a remote host connecting to more than 25 LBL hosts as engaging in resource discovery. We found no such instances in the first three datasets, one in LBL-4, two in LBL-5 and LBL-6, and four in LBL-7. Of the nine resource discovery instances, six involved hosts from the same university as the authors of *Netfind*, indicating our heuristic did successfully identify use of resource discovery tools. Assuring minimal network load was one of the goals of the authors of [ST91]. We found that during the busiest resource-discovery dataset, LBL-7, resource discovery accounted for 8% of all *finger* connections and 3% of all *finger* bytes, providing evidence that this goal has been met.

- *telnet* (remote terminal login) growth appears roughly linear, increasing about 150 conn./day/year, except for a huge increase with LBL-7, all due to a single pair of hosts (again, see Section 5 below).

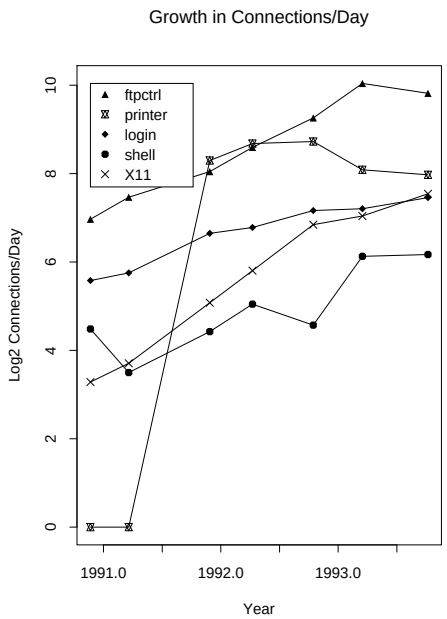

Figure 4: Daily Connection Rate for Less Popular Major Protocols

Figure 4 shows the connection rates for the remaining protocols that had connections during all of the datasets. The Y-axis is again logarithmic:

- *ftpctrl* (the control side of an *ftp* session) shows strong exponential growth of 140%/year for the first six datasets, but then a decrease in connections for LBL-7. It turns out that *ftpctrl* connections are dominated by weathermap connections (as discussed in Section 5 below); after removing them we obtain a good exponential fit across all datasets corresponding to growth of 66%/year.

- *printer* (remote printer access) connections increased enormously between LBL-2 and LBL-3, and then leveled off and began heading down again, a phenomenon again discussed in Section 5 below.

- *login* (Unix remote login) traffic grew fitfully, much as did *telnet*, but can be fit fairly well to linear growth of 46 conn./day/year.

- *shell* (Unix remote command execution) traffic is very sporadic (but see the discussion of bytes transferred, below).

- *X11* (X11 network window system) traffic, on the other hand, shows very consistent exponential growth between LBL-2 and LBL-5, growing at the ferocious rate of 300%/year. Fitting all seven datasets gives growth of 198%/year, but the fit is not all that compelling. Instead, the lower growth rate between the last three datasets suggests that *X11* traffic is exhibiting logistic growth (see the discussion of *X11* traffic volume in bytes below).