The low utilization and high cost of data networks

Andrew Odlyzko

AT&T Labs - Research
amo@research.att.com

1. Introduction

The rapid growth of packet data networks is usually ascribed to their lower costs compared to those of circuit-switched voice networks. These lower costs are supposedly coming from greater efficiency in utilization of transmission lines. However, this is a misconception. Packet networks are growing because of their flexibility, not because of low cost. Large file transfers on most corporate packet networks cost more than sending them over a modem on the public circuit-switched voice network (PSTN).

The high costs of corporate packet networks are caused primarily by low utilization rates, far lower than those of PSTN. Furthermore, this situation is likely to persist. Low utilization of data networks is not a symptom of waste. It comes from different patterns of use, the lumpy capacity of transmission facilities, and the high growth rate of the industry.

The circuit-switched PSTN is engineered to provide a low-cost solution to all normal demands. Many calls may get blocked after an earthquake, but peak hour demand is accommodated even on the busiest days. Despite the design that can accommodate peak-hour traffic, the average utilization of long distance links in the PSTN is about 33% over a full week. This efficiency comes from careful traffic engineering and sophisticated routing, from the smoother and more predictable nature of voice traffic in general, from the predictable growth in demand for voice services, and from sharing of the network among users with different calling patterns.

The Internet is slow, as anyone who surfs the Web can attest. The general impression is that Internet backbones are seriously congested. There are indeed many links and nodes on the public Internet that are heavily loaded, especially the public peering points, the expensive trans-Atlantic and trans-Pacific ones, as well as many lines from smaller Internet Service Providers (ISPs) which aggregate traffic from residential modem users. However, a study that Kerry Coffman and I conducted showed that Internet backbones in the U.S. in 1997 had a considerably lower utilization rate than PSTN.

Private lines form the bulk of the long distance data networking “cloud.” These lines are also commonly regarded as heavily loaded. However, there is strong evidence that this is another misconception, and that corporate data networks in the U.S. are lightly utilized, with much lower average utilization than even the Internet backbones. Most corporate networks show patterns such as those of Figures 1 and 2, with most of the traffic concentrated during the business day. For IP networks, average utilization rates (over a full week) tend to be in the 3-5% range. This is not to say that there aren't heavily used private lines. Trans-oceanic lines in particular are typically very heavily loaded, with average utilizations of 20-30% in some cases. However, in the continental U.S., uncongested networks are the rule.

Until my study (see the References at the end) the low average utilization of U.S. enterprise data networks had not been pointed out in any systematic report. Carriers such as AT&T that lease private lines to corporations do not look at the traffic on those lines for privacy reasons. There is no organization that regularly collects utilization statistics from a large sample of enterprises. Many corporations do not measure their own usage. Those that do are more concerned with the busiest hour over a busy week, or the busiest five minutes over a period of months. Many monitor only the bottleneck links. Further, most organizations are reluctant to release any data about their networks.

The general perception is that private lines are heavily loaded, and one often hears estimates that utilization levels are routinely around 70% during the peak hours. In reality, though, the 70% level usually reflects the utilization that is observed over short periods over a few links in a system. Even network managers who are asked to estimate the average utilization without collecting the data tend to overestimate, since they remember best the ”hot spots” that are present in almost any network.

2. Reasons for low utilization rates

Data networks will continue to be lightly utilized for several reasons. Some of the inherent inefficiency of data networks comes from their voice heritage. Having a full channel for each person in a voice call was a reasonable choice in early telephone networks. As a result, data lines are also symmetric despite the fact that traffic patterns are asymmetric. (This asymmetry is most pronounced for Web servers, which transmit much data and receive very little. Such servers are increasingly prominent in corporate networks. Other applications, such as SAP and data backups, are also notorious for their asymmetry.)

Data traffic in corporate networks is concentrated during regular business hours (Figure 2). The busy hour usually carries about one sixth of the day's volume. Since there is little weekend activity, the traffic carried in a 168-hour week is equivalent to that carried over about 30 hours at peak hour utilization. The 3-5% average utilization range that I have observed over a full week corresponds to average peak hour utilization of 15-25%.

Data traffic is growing much faster and less predictably than voice traffic. Internal IP traffic is often growing about 100% a year, and growth within a corporation is uneven as new services are deployed. Installing new links is a slow process, with waits of up to a year reported for private line T3s. It is prudent to over-provision so that internal customers can satisfy mission-critical requirements.

The natural tendency to build in adequate safety margins is aggravated by the lumpy nature of network capacity and the decreasing prices per unit of bandwidth as one purchases higher capacity links. A T3 line has 28 times the capacity of a T1 line, but typically costs only 6 to 10 times as much. Hence it is often more economical to use a T3 instead of seven T1s, even though average utilization of the T3 will be only a quarter of that on the seven T1s.

Perhaps the most important reason for the low utilization of data networks is that these networks have to serve internal and external customers. Although data traffic is bursty (as is shown in Figure 1), it can often be made to fill a large fraction of the capacity of a data link. Web browsing and file transfers that rely on the Internet's most widely used protocol, TCP, can accommodate to any bandwidth that is available. The price of using such mechanisms to reach high utilization is delays. On trans-oceanic links, where costs of transmission capacity are high, network managers generally make the explicit or implicit decision to tolerate the low transmission quality and achieve high utilization. Domestically in the U.S., though, where costs are much lower, they react to demands for better quality. In the words of one branch office manager of a large company,

> I see peak bandwidth as the basic commodity I buy. ... When we had a 256Kb data line it was too slow (it interfered with productivity). With a T1 line, no one has complained. I guess our T1 line is less than 1% utilized. ... I would not go for a T3 line (it would not improve our productivity) but I would not cut back on the T1 line.

Low utilization may be technologically inefficient, but it is often the most economically efficient solution when the total system cost is considered. Users do not care about networks per se. What they do care about is that transactions complete fast. If a newspaper doubles the capacity of the private line between its editorial offices and the printing plant, the utilization rate drops by half but the staff gains extra time to work on the edition before going to press. Whether that is worthwhile or not has to be decided by the managers of the business, and not by arbitrary rules about network utilization.

3. Conclusions

The question is whether low average utilization of corporate data networks matters. It is irrelevant for designers of private line networks whose task is to find the most efficient way of providing levels of service to customers at minimal cost. Utilization will stay low for those who must accommodate bursty data transmissions, concentrate their traffic during regular business hours, and be free to suddenly generate increased traffic loads as new services are added.

On the other hand, low average utilization is relevant for other purposes. Aggregation of corporate traffic through use of virtual private networks (VPNs) over the Internet or through public Frame Relay or ATM networks can deliver substantial savings, since utilization rates can be higher, and the links of much higher capacity (and thus much less expensive per unit of bandwidth).

Low utilization rates also cast serious doubt on the advisability of many quality of service (QoS) measures, which are largely motivated by the assumption that networks are congested. When a link is heavily loaded, it makes sense to create a special “lane” for high priority traffic, to avoid the delays and packet losses of the bulk of the traffic. However, when a network is engineered to handle high-speed bursts that arise from mission-critical applications, it will exhibit low utilization, and all traffic will be going through expeditiously almost all the time.

References: This note is based primarily on my paper, “Data networks are lightly utilized, and will stay that way,” which is available at

⟨http://www.research.att.com/∼amo⟩.

That URL also contains links to the joint study with Kerry Coffman of the size and growth rate of the Internet, and to several other papers that consider the economics of data networks and the implications of observed behavior, such as the low utilization rates, in greater detail.

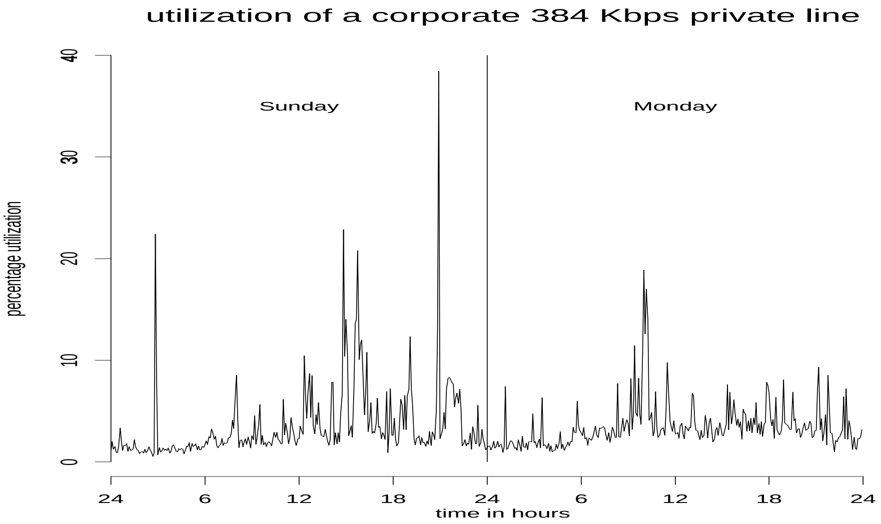

Figure 1: Traffic on a corporate link to the Internet. 5-minute averages.

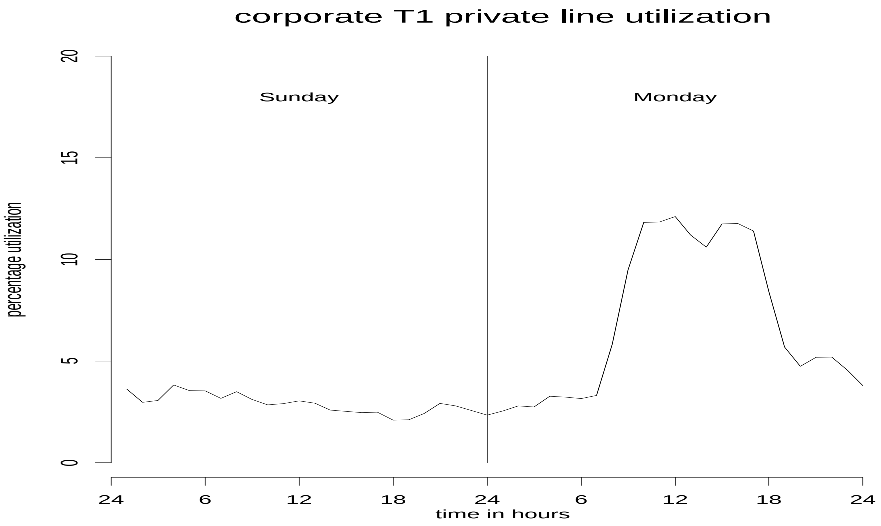

Figure 2: Average utilization of T1 links in a large corporate private line network. Hourly averages.