8-15-1997

# Dial-Up Server Capacity Planning Model

Richard A. Elnicki
*University of Florida*

# Dial-Up Server Capacity Planning Model

**Richard A. Elnicki**
University of Florida

## Introduction

Dial-up servers are widely used to access the Web, Internet, and other computing services. In some cases, such as the recent AOL stories attest, limited access is a major cause of user dissatisfaction. A "busiest-minute" server capacity planning model is proposed here. A limited example of the use of the model is presented.

This study focuses on dial-up capacity for access to the general purpose computer utility at the University of Florida (UF), the Northeast Regional Data Center (NERDC). The components of the access path follow.

- Voice grade phone lines at homes or offices.
- Bell South switching office lines (NARS).
- T1 (fiber) lines from Bell South to NERDC.
- Modems connected to Cisco servers.
- Cisco servers giving TCP/IP addresses.

It is used primarily for access from home, though some use the path for access from their offices. The servers permit anyone affiliated with the University of Florida to access graphic- or character-based services on campus or any Internet site. UserID's and passwords are required to connect to one of the lines. Connect time currently costs $.008/minute for access to Internet services via local lines (long distance access is currently $.20/minuute). The current 424 lines give connections at a maximum of 28,800 bps.

Demand can be measured a number of ways. This study uses charge records that include connect and disconnect times to seconds for each dial-up session. Analyses of these records show the servers' demands peak in the evenings, a logical expectation since most all users work on the UF campus during the day. he busiest hour is usually from 10 p.m. to 11 p.m. The busiest minute is typically late in that hour. In March '97, the busiest minute was 10:54-10:55 p.m., Tuesday, the 18th. March '97 had the highest aggregate demand to date, with 256,942 sessions and 6,562,253 connect minutes. The average session length was 24.2 minutes. Aggregate demand has been doubling every 12 months, see http://nervm.nerdc.ufl.edu/~dicke/ts/tsuse.html and http://nervm.nerdc.ufl.edu/~dicke/ts/ptuse.html.

## Background

Measuring demand in terms of capacity units used in short time periods is common. Electricity and gas peak-load demand is typically measured by the hour. Peak-load demand on limited-access highways and toll booths is measured for rush-hour traffic periods (that are usually longer than an hour). A study of water use in Stettler, Alberta, Canada, required peak-load demand analysis. An electricity spot market in England and Wales sets prices every half-hour. A "hundreds of call seconds per hour" metric is commonly used to measure telephone line use. One hour, 60 minutes x 60 seconds, has 3600/100 or 36 CCS's. [References omitted to save space.]

Bell Atlantic used the CCS measure to show the impact of internet service providers (ISPs) on line use rates in the Virginia, Maryland, Washington area February 25 to March 23, 1996 http://www.ba.com/ea/fcc/index.html and Telephony, July 29, 1996]. Use by 7 IPS on about 5,000 circuits and by 800 business customers was measured for 24-hour days over the 28 days at 9 Bell Atlantic central offices. The peak hour usage averages over the period were 72 percent for ISP's on traditional switched public network lines, 78 percent for ISM's on ISDN lines, 33 percent for businesses on multi-line hunt groups, and 8 percent for all customers for the entire central office. ISP calls averaged 17.7 minutes as compared to 4 to 5 minutes for all other calls. The ISP maximums occurred at 10 and 11 p.m. respectively.

The business and all-customer maximums were at 5 and 4 p.m. respectively.

An industry rule is that a group of lines is considered saturated if 1 percent of the call attempts get a busy signal. Busy signals on lines used by the ISP's exceeded this level. Bell Atlantic noted that major investments were required to "…provide acceptable levels of service to all end users…" For example, in the 9 central offices, 5 times more trunk lines were installed during the year ended June 1996 than were expected to be added without the ISP's, costing $17.8 million. The objective of the study was to show the current FCC "Enhanced Provider Exception" promotes saturation of capacity with no market-related pricing mechanism to control use, giving Bell Atlantic's ISP's a 73 percent cross subsidy during 1996.

**Busiest Minute Planning Model**

The governing board of the computer center running the server wants to provide a service level that satisfies its users, e.g., no more than 1 user in a 100 does not get connected in a time period. The context includes a State of Florida law requiring that the center's services pay from themselves over the long run. The capacity providing a given service level will have an average cost at a demand level. That demand must result in revenues that will cover the costs. This capacity planning model addresses the following question.

What price must be charged at a given service level to meet the financial requirement?

"Service level" means the percentage of users that will get a phone line and TCP/IP address for a dial-up session within one minute of the time their systems send dialing commands from their micros to Bell South. It takes about one minute for a typical micro and modem to make 3 calls to the server with a script that redials when "BUSY" is reached. So, one minute was chosen as the initial period for this study. The period will be increased at least up to one hour to compare results with studies using CCS's.

Major tradeoffs the governing board must consider include the following: At a demand level, a given capacity level has a service level and average cost. A service level of 1 percent (1 percent of the users do not connect in 1 minute) will require more capacity than a 5 percent service level. In general, there is an inverse relationship between service level and capacity, and a positive relationship between capacity and average cost.

At some sufficiently small non-service probability, the price will be too high and drive away demand. Similarly, if the non-service probability is too great, dissatisfaction by users will also drive away demand. The proposed planning model will identify what service level a capacity level will provide given existing demand distributions of busiest minutes. Related average costs infer prices the governing board must consider and compare to other market offerings. The optimal point is between "excess" capacity that provides satisfactory access times but drives away demand via higher prices and "insufficient" capacity that is low priced but drives away demand via user dissatisfaction with a high non-service probability.

If that optimal point implies a price lower than or equal to market offerings providing similar service levels, the service should be continued. Otherwise the service should not be offered by the computer utility.

**Model Application**

This application considers only the service level dimension of the planning model. Predictions of future demands, timed capacity additions, average cost and price implications for various service levels are not included here. A brief review of these aspects is at http://nervm.nerdc.ufl.edu/~dicke/ts/capmod.html.

The 1,416,980 charge records of all local sessions from October '96 through March '97 were put into a 182-day by 1440-minute matrix. The busiest minute in each day was identified. Daily data for the servers were obtained from the engineers responsible for daily maintenance and periodic upgrades. SAS's GLM procedure was used to find the best estimator of the busiest minute (MAXI) per day. The result of this iterative procedure was a simple linear estimator. The right-hand-side has a time variable specified as the 1

to 182 days, one supply-side variable, and two demand-side variables.

The capacity by day is not included. It was not significant in the model when the variables noted above were in the model. This was expected since in no day were all lines used except some of the days when bad lines/modems/servers (BAD) reduced the actual capacity. The supply-side variable BAD captures this actual capacity effect -- BAD was significant at conventional levels in all models tested.

The major demand-side variable is a binary (1,0) identifying days when the university was/was not on vacation (VAC). This vacation effect for the 1996 Thanksgiving and Fall 1996 semester end could be seen by casual inspection of busiest minute lines (and average lines) starting a few days prior to the actual start of vacations. In these vacation periods the lines in the busiest minute are 105 lower (about 41 percent!) than the average over the 182 days.

The day of week is another demand-side variable. Visual inspection showed that Saturday always had the lowest busiest minute (and average lines) in a week. It is the intercept default in the model. Each of the other days is structured as a binary (1,0) variable. Monday (MON), Tuesday (TUE) and Wednesday (WED) have the highest busiest minute in that order, always in the 10:00 to 11:00 hour and about one-third higher than Saturday. Sunday (SUN) and Thursday (THU) are about 18 percent higher than Saturday. The estimation equation, "T" statistics, and significance levels from SAS's GML are in Table 1.

Table 1: Busiest Minute (MAXI) Estimate

| Variable | Coefficient. | T-Stat | Pr>\|T\| |
|---|---|---|---|
| Intercept | 177.16 | 25.22 | 0.0001 |
| Time (Days) | 0.44 | 10.46 | 0.0001 |
| Bad Lines | -0.92 | -3.04 | 0.0027 |
| Vacation Day | -104.99 | -17.12 | 0.0001 |
| Sunday | 47.84 | 5.86 | 0.0001 |
| Monday | 80.41 | 9.84 | 0.0001 |
| Tuesday | 77.85 | 9.53 | 0.0001 |
| Wednesday | 73.93 | 8.96 | 0.0001 |
| Thursday | 46.49 | 5.70 | 0.0001 |
| Friday | 7.66 | 0.94 | 0.3490 |

The R-Square is .786. The standard deviation, 29.40, and the MAXI equation gives an estimate that 2.75 percent of the users would not connect in a minute or less on the busiest day in the week when capacity was last added. The addition of those 48 lines reduced the probability of non-service to .02 percent for that Monday.

A different rule is currently used to add capacity. A "high-water mark," the highest line/modem hit for at least an instant of time, is recorded for each day. Judgmental estimates are made about the future day that mark will hit the current capacity. The six-week capacity addition cycle is started in anticipation of that day. The linear model used for MAXI was fit to the same 182 days' high-water marks. The results, including significance levels, were remarkably similar, with a R-square of .776 and a standard deviation of 29.47. The average high-water mark for the last week in March '97 was 322.5. The average estimation equation value of MAXI for that week was 304.2, a difference of 18.3.

If a busiest minute rule with a .01 service level had been in effect, capacity would have been added in February '97. But a smaller increment could have been added, e.g., 2 instead of 3 Cisco servers. This would have resulted in a lower average cost and permitted a price decrease. And, the non-service probability of

the busiest day that week would have been .13 percent. The access difference would be the expectation that 13 in 10,000 dialing attempts would not get a line in a minute or less as compared to 2 in 10,000.

**Additional Issues**

On-going study of the data revealed additional issues that must be resolved to implement this model. The importance of vacation days on the MAXI and average demand levels was not expected. Capacity additions in these time frames may be preferable even if a few weeks earlier or later than the forecasted date. Plots of aggregate data by month show summer semesters will also be low demand periods as compared to fall and spring semesters.

The estimation model presented above assumed the variance is homoskedastic through time. Casual review of the 6 months' data suggest it may be heteroskedastic. This potential problem will be analyzed when the daily data by minute is created back to August 1, 1995.

Another question is how overall demand will grow. Three times in the period since aggregate demand data collection started, it appeared that the series had moved into the third, slowing-growth segment of the "lazy S" growth curve i.e., the point where the second derivative becomes negative. Subsequent data indicated growth was still in the middle, hyper-growth segment. So it now appears the simple linear model giving a best fit for the 6 month period reported here will not give a best fit for the 20+ month data set that is in progress.

If existing demand has slowed, another major hyper-growth segment could be added since the UF is planning to move to a "universal access" ID for all students, faculty and staff. It would guarantee some minimum e-mail and the Web access and pay for some level of dial-up access. However, some students may stay with their existing private ISP's. In the Spring Semester 1997, 20 percent of the author's undergraduate telecom students prefered to use their private ISP's for e-mail rather than use "free" class accounts. When questioned about this, all indicated they considered the time costs of changing to the class accounts greater than their monthly ISP bills.

Bandwidth capacity itself has not been an issue to date. The current bandwidth constraint point in the path is on a hub through which the Cisco servers connect to two ethernets. For that 10 megabit hub to be saturated, all lines would have to be simultaneously moving bits, e.g., all doing file transfers, and be doing so at the maximum possible rate of 28,800 bits per seconds. To date, all lines have never been in use simultaneously transferring data and an actual transfer rate of 28,800 bits per second seldom occurs even when users' modems do connect to the server's modems at that rate.

Calls to the NERDC server use the same pool of network access registers (NARs) at Bell South's central office as do all office and voice phones and a number of "private" modem pools on campus. During an unspecified number of times in recent months the use rate of the NARs led to a busy rate greater than the 1-in-a-100 standard. Additional NARs were added to the campus pool. However, the information currently available states that the times when the standard was exceeded were on Thursday and Friday evenings from 11 to 12 p.m., times when the NERDC server was not at peak use. Efforts are currently under way to obtain daily information on NARs use. It will be studied to determine if it could have constrained access to the NERDC server. The time period of this planning model will be changed from one minute to one hour make the NARs use rates comparable to the NERDC server use rates.