I'm Over-Trunked by How Much?!

A Look at the State of PBX Networks

Talk to most Network Managers and they will acknowledge that their company's voice network has a lot more capacity than is required. This document examines the extent of the excess capacity, its causes and the solutions.

Basis of Analysis

The results discussed herein were drawn from a large number of recent traffic studies. The traffic studies were conducted using the standard engineering methodologies prescribed by telecommunication equipment and service providers. The traffic studies gathered the traffic load (time the trunks were in use) for each trunk group on an hourly basis over a 5-day period. Each company's Network Manager determined the dates for the studies. Typically these dates would be selected when the Network Manager felt the traffic volume would be at its highest.

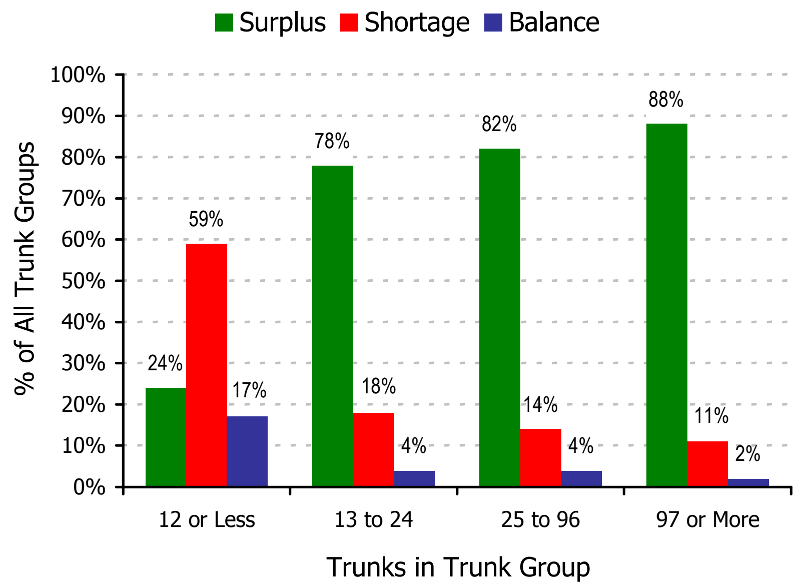
To determine whether trunk groups had enough, too little or too much capacity, a comparison was done of how many trunks (the term "trunk" is used to describe a circuit capable of handling one voice conversation) were in place versus how many were required. The quantity of trunks required was determined by taking the highest hourly value (a.k.a. Busy Hour) of traffic load during the 5-day period for each trunk group and using the Erlang B formula to convert the load into required trunks. The Network Managers specified the targeted service levels, typically p.01 (1% of calls blocked).

The traffic studies on which this analysis is based involved 323 PBXs with 2,765 trunk groups and 61,292 trunks. These PBXs are used by companies of all sizes and from a wide variety of industries.

The Scope

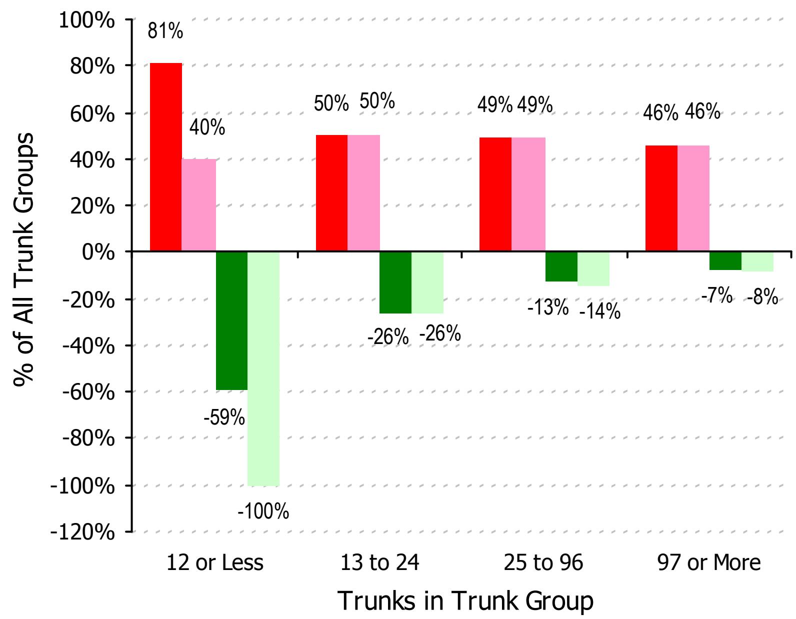
Out of Balance

The table to the right shows the percent of trunk groups that had surplus capacity, didn't have enough capacity and those that had the amount of capacity specified by the 5-Day Traffic Study. The percentages are shown for various sizes of trunk groups.



The smallest trunk groups, those with 12 or less trunks, did not have enough trunks 59% of the time, had too many trunks 24% of the time and had the correct amount only 17% of the time. The results for trunk groups with 13 or more trunks were dramatically different. Of the largest trunk groups, those with 97 or more trunks, 88% of the time they had too many trunks, too few 11% and the right amount just 2%.

Surprised? Well hold on because it gets worse. The percentages noted above wouldn't be so bad if the amount of trunks a trunk group was short or surplus were relatively small. But it's not. The next graph shows the percentage the actual trunk group size deviated from that called for in the traffic study. The darker bars are the mean (a.k.a. average). The lighter bars are the median. Note that not only were 80% of the larger trunk groups over-trunked (as shown in the graph above) but also averaged nearly 50% in excess capacity versus what the 5-Day Traffic Studies said was required.



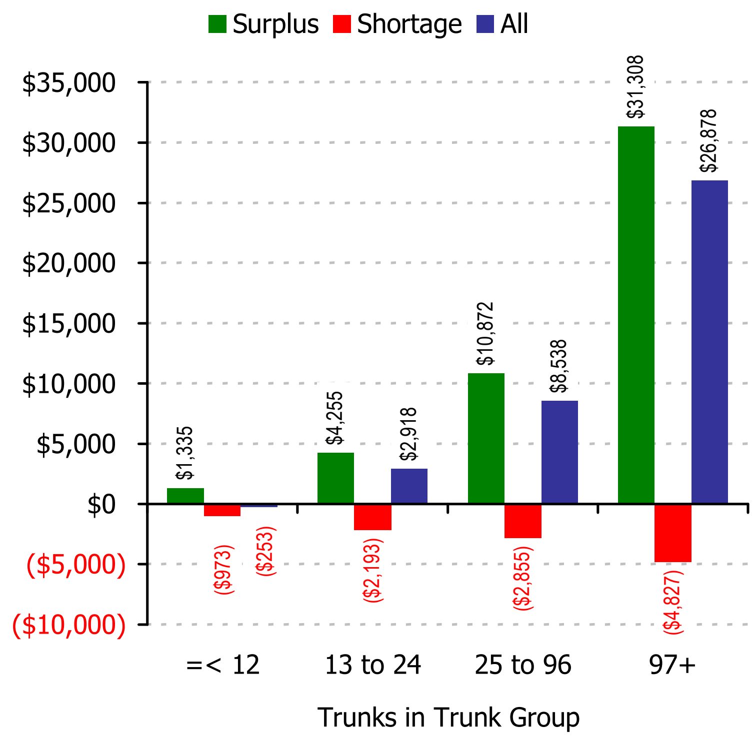
The PBXs in this study averaged 190 active trunks in service. These trunks were spread across multiple trunk groups of varying sizes. Blending the percent and scale of trunk groups with shortage, surplus and in balance the average PBX required only 121 trunks. This equates to an average excess capacity of 36%. Companies with larger PBXs, typically having a greater percentage of large trunk groups, can expect their excess capacity is even greater.

The Dollar Value

Over-trunking is widespread and significant in scope. Before spending money to fix a problem it helps to understand the scale of the potential savings. To the right are some typical annual costs for T1 trunks. The annual average of $408 per trunk will be used to translate the deviations discussed earlier into dollar terms. A company should adjust the following analysis by the percentage their actual annual costs differ from $408.

Annual Cost Per Trunk		
T1 CO Service Fee		$8,660
T1 Card 5 Year Amortization		$698
T1 Card Maintenance		$432
	Total	$9,790
	Per Trunk	$408

The following chart translates the charts on the previous page into monetary terms. Shown are the annual amounts that were either spent unnecessarily on excess capacity or how much more would have to be spent to provide the capacity indicated by the traffic studies. The chart presents the results by trunk group size. The left bar in each set shows the amount spent unnecessarily on surplus capacity. The middle bar shows how much additional money needs to be spent to correct trunk groups with deficit capacity. The right bar sums the surplus and deficit amounts and divides the sum by the total number of trunk groups, including those in balance.



For the largest trunk groups with surplus capacity, $31,308 is spent per year above what was called for in the traffic studies. For the largest trunk groups short on capacity an additional $4,827 would be needed per year to provide the capacity indicated by the traffic studies. The average across all of the largest trunk groups shows that $26,878 is being spent per year over the amount called for by the traffic studies.

The average PBX, having 190 trunks, has $28,108 per year in excess capacity. Companies relying on using 5-Day Traffic Studies to determine voice network capacity requirements can expect to be spending a similar excessive amount per year (adjusted to their size of PBX and cost per trunk).

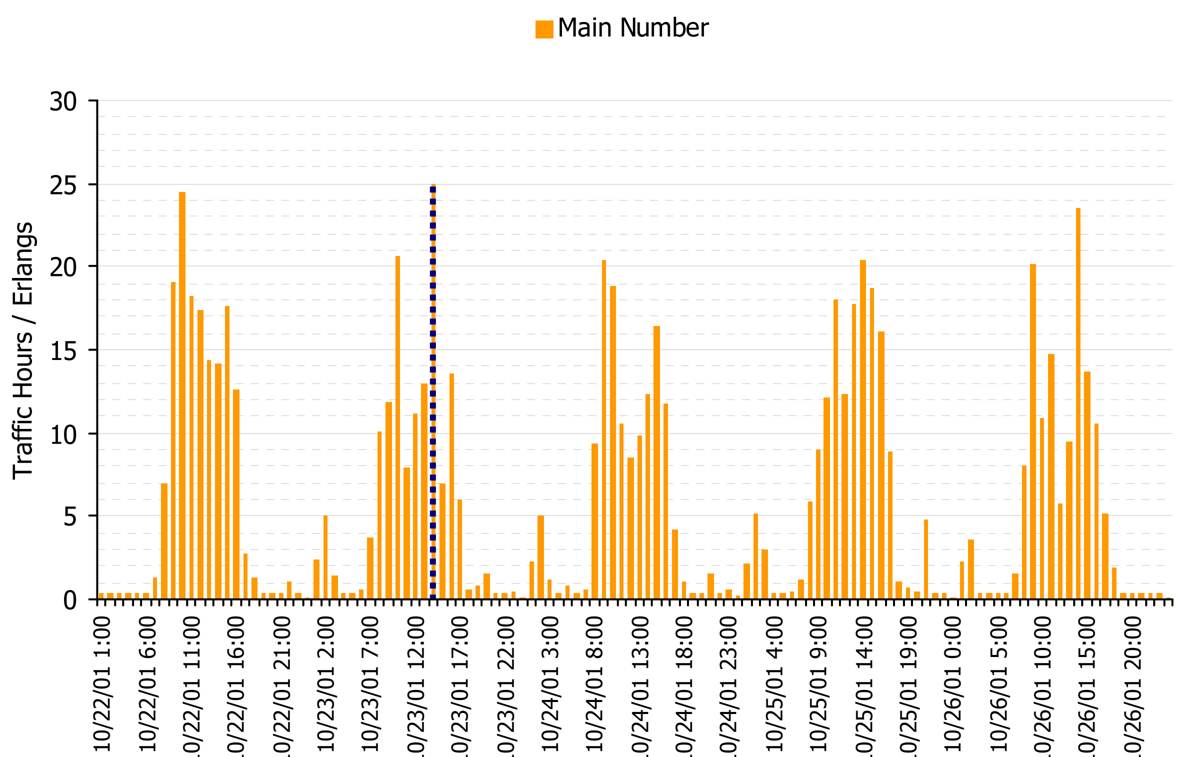
The Cause

So the typical PBX has 36% excess capacity, adding $28,108 of unnecessary annual recurring expense. Why? The answer is really quite simple. Network Managers don't trust the industry's 5-Day Busy Hour Traffic Study methodology to yield predictable results. With a 5-day study period Network Managers aren't sure they are capturing a trustworthy data sample. They aren't confident the data is meaningful for any one of their trunk groups, let alone all of them. Most find the concepts of Busy Hour, Average Busy Hour, CCITT Busy Hour and the like to be less than straightforward. They are uncomfortable explaining to senior management how engineering to a p.01 Grade of Service (a.k.a. GOS) using the Busy Hour from a 5-day traffic study optimizes the corporate network. Knowing no other more trustworthy methodologies, the Network Managers react by erring on the high side, the very high side.

When is the Busy Hour Anyway?

The chart below graphs the hourly traffic load for the trunk group servicing a particular company's Main Number. Traffic load is shown in units of hours (a.k.a. Erlangs).

Before getting to the graph, a definition of Erlang is in order as many have heard the term but few are sure what it really means. An Erlang is classically defined as 36 CCS. A CCS, standing for Centum Call Seconds, equates to 100 seconds of time a trunk is in use. So if a CCS is 100 seconds of trunk usage (a.k.a. volume) and an Erlang is 36 CCS, then an Erlang can also be described as 3,600 seconds of trunk usage. As there are 3,600 seconds in an hour (60 seconds/minute and 60 minutes/hour) an Erlang can also be simply described as an hour of network usage.



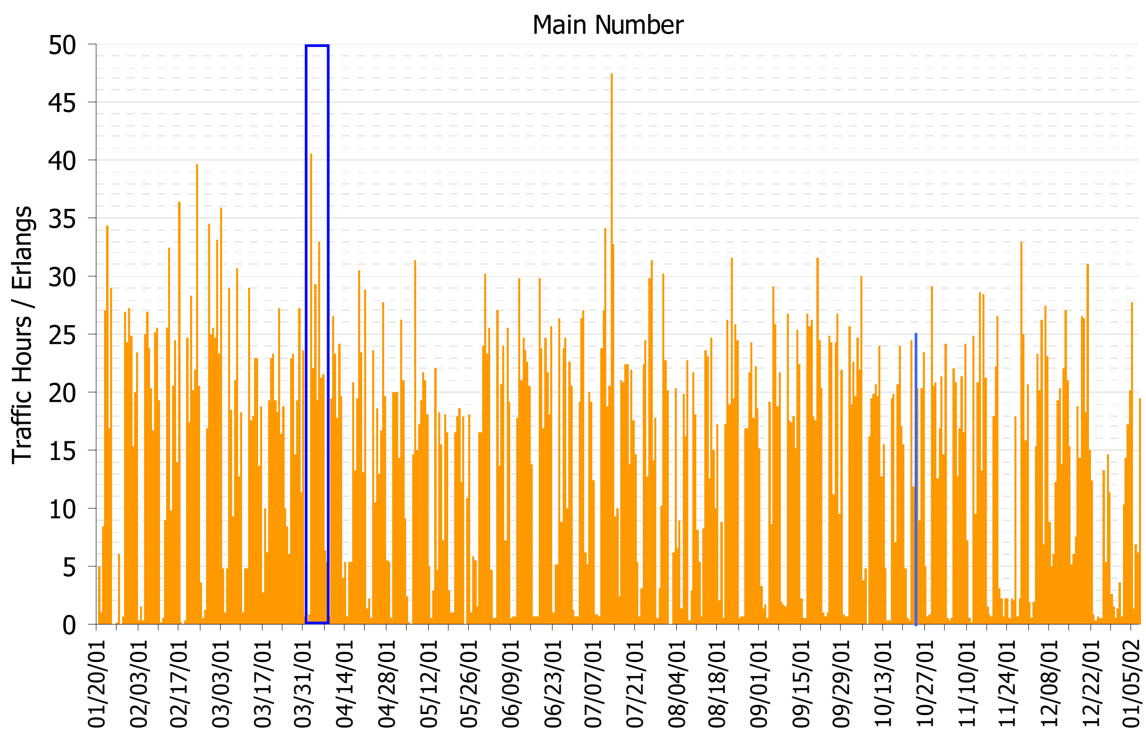
The orange lines in the graph above show the traffic volume on an hourly basis for the 5-day period of October 22nd to October 26th. One can quickly see that the Main Number is predominantly used on an

8am to 5pm basis, which is not atypical. The dashed blue line highlights the highest hourly value (a.k.a. the Busy Hour) of the week, which occurred on 10/23/01 from 2pm to 3pm with 25 Erlangs.

The standard industry methodology would then use 25 Erlangs as the basis of determining trunking requirements. One would reasonably conclude that if provisioned to 25 Erlangs using a p.01 GOS that only 1% of the calls would be blocked during this trunk group's busiest period. But this is only valid if the busiest hour value that occurred during the selected 5-day period is the same value of the busiest hour for whatever future range of time trunking resources are being planned, purchased or contracted.

The graph below shows the hourly data for the same trunk group for a much broader range of time, a full year. This expanded scope of data reveals several interesting things.

First, note the blue line in the October time frame. This line is highlighting the 25 Erlangs from 2pm to 3pm on 10/23/01, as was the preceding graph. While this was the Busy Hour during the 5-Day Traffic Study, it doesn't come close to reflecting the busiest hour of a year. One wouldn't know this unless they had access to this expanded data set.



Note that the highest value occurred in July, more specifically July 12 from midnight to 1pm. This is an interesting time of day for the highest value to occur given this trunk group is used to support an 8am to 5pm operation. Further investigation revealed that the PBX seized and hung all the trunks in this trunk group for the hour following some diagnostic routines run at midnight, which apparently weren't so routine the night of July 12th. So an expanded data set also improves the ability to identify situations where systems are operating outside expected parameters.

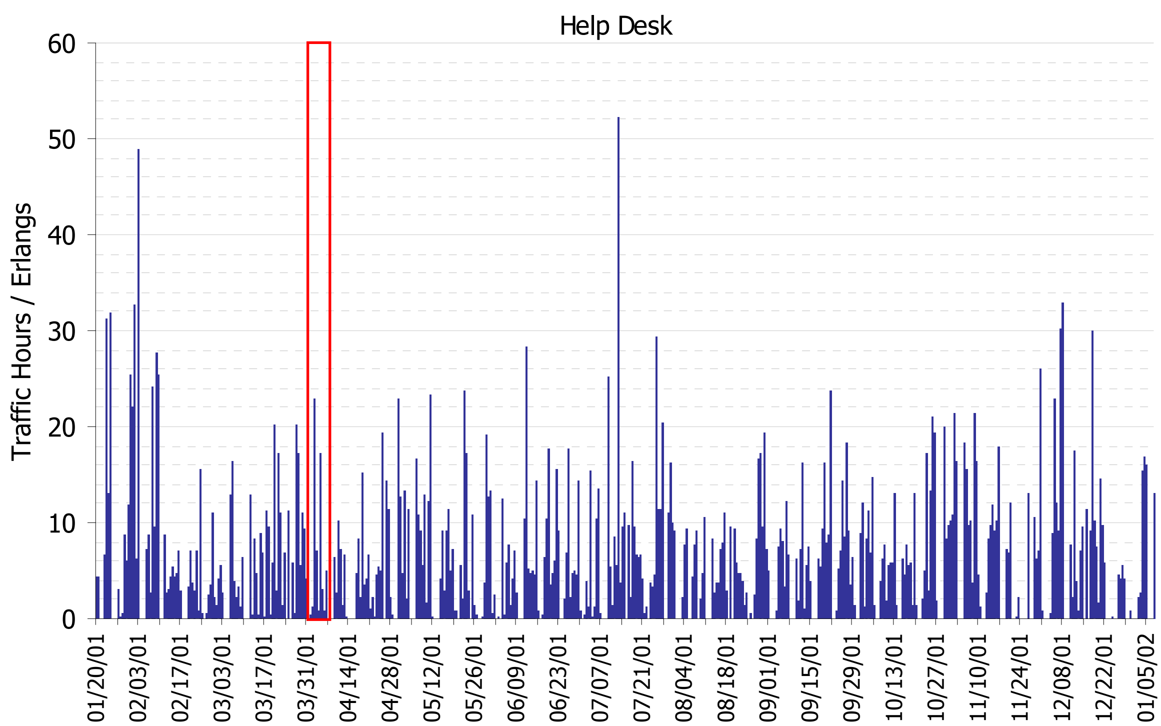
Given the expanded time frame one can now see that the true Busy Hour for this trunk group occurred during the first week of April, around which a blue box is drawn. If one truly wants to engineer to the Busy Hour, a concept that will be re-visited later, instead of the 25 Erlangs indicated by the 5-day period in October one should use the 41 Erlangs from the first week of April.

If a company knows that the Busy Hour for the previous 12 month period for this trunk group occurred during the first week of April should they run a 5-Day Traffic Study during the first week of April next year? If they do so, how much confidence would they have that the Busy Hour is again contained inside that same week? Using a 5-Day Traffic Study to determine the trunking requirements over a much broader range of time yields widely fluctuating, unreliable results. Hardly something one would consider "engineered".

The Busy Hour for Different Trunk Groups Occurs at Different Times

Let's presume for a moment that a company clairvoyantly predicts that the Busy Hour for the Main Number Trunk Group will occur the first week of April. If they conducted a 5-Day Traffic Study during this period should they expect that the Busy Hour for all trunk groups associated with the same PBX will occur during this same time period?

The graph below shows the same data elements as the previous graph but for another trunk group, which services a Help Desk. There are a couple things to note. First see the same "hung trunk" phenomena on July 12th from midnight to 1pm as the Main Number Trunk Group experienced. Having the broad spectrum of data for multiple trunk groups makes this operational aberration easy to spot. Second, the first week of April, outlined in the red box, clearly does not contain the Busy Hour for this trunk group.



If a company's goal is to engineer to the Busy Hour, clearly this can't be done using the same 5-day period for all trunk groups. The Busy Hour for the Help Desk Trunk Group occurred during the last week of January. Using the first week of April instead would have meant engineering to 23 instead of 49 traffic hours, producing obviously very different results. Correspondingly, if the last week of January had been used to determine the Busy Hour for the Main Number Trunk Group, a value of 27 traffic hours would have been used instead of the 41 from the first week of April. Using data from a single 5-day period to engineer multiple trunk groups does not yield reliable, effective or equivalent results across those trunk groups.

The Solution

Expand the Data Sampled

Back in the 1970s Bell Labs used a simple data collection and reduction methodology to obtain the data it used for engineering the trunking capacity of its toll switches. Attached to each switch was a Teletype that would print out the traffic volume for each trunk group of each switch every hour. At the end of each day, the Teletype print out was examined to determine the highest value for each trunk group. This value was recorded on a form designed to capture this data for every day of the month. At the end of the month, the highest eight (8) values for each trunk group would be taken from this form and entered on to another form. This would be done for twelve (12) consecutive months. Thereby, over the course of a year, 96 of the highest datapoints for each trunk group would be gathered. These datapoints would then be sorted from high to low. The top end of this list could be fairly well relied upon to represent the busiest times each trunk group experienced over an entire year. Even though this methodology produced only 96 datapoints, the process by which the data was gathered creates a data sample with much higher integrity than the 120 datapoints gathered during today's industry standard 5-day study.

The Bell Labs methodology was effective in gathering data from a broad range of time. This diminishes the risks noted earlier associated with limited data samples, such as from 5 days. This methodology also reflected the state of information technology in place at that time. Compared to today, computing power was expensive. Most of the data collection and analysis was done with people not computers. So this methodology accepted some risk that the integrity of the data sample would be compromised in order to contain the labor costs of gathering and analyzing the data.

Limited data samples, even expanded ones using Bell Lab's methodology, also compromise the analysis of the combined traffic load from multiple trunk groups or PBXs. The peak load for each of the trunk groups to be combined most likely didn't occur at the same time. With limited data one can just add the peak values of each trunk group together. This most likely overstates the true peak traffic volume of the combined trunk groups.

The state of today's information technology does not require compromising data integrity. Data can now be cost effectively gathered on a 24 hour a day, 365 day a year basis. Such a comprehensive data sample:

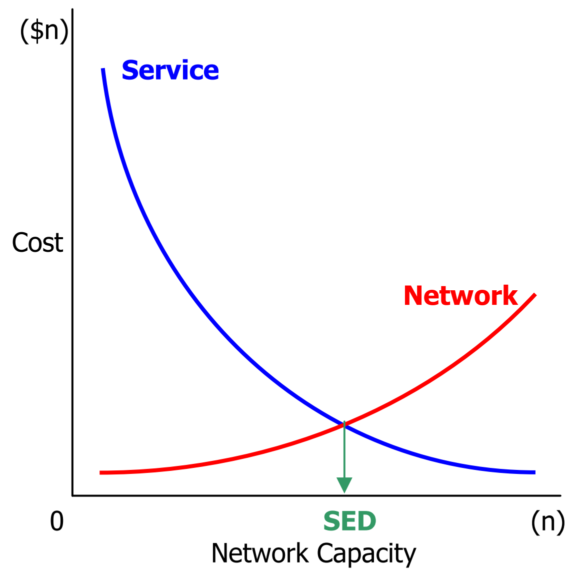
- assures the height and duration of rising and falling traffic volumes are captured.
- allows the ready identification of trunk groups performing out of specification.
- provides opportunities to fine tune a network by seasonally adjusting capacity.
- gives an accurate picture of traffic under consideration to be consolidated from multiple trunk groups.

Pick the Right Datapoint

With any data sample one must be careful of which data is used for projections. The Bell Labs used the term SED (a.k.a. Significant Engineering Datapoint) to describe which of the 96 datapoints should be used to determine the proper level of network capacity. The theory behind the SED was rooted in basic microeconomics. SED is the datapoint that balances the incremental costs of increasing network capacity with the incremental costs of diminished customer service.

Limited network capacity results in calls being blocked and callers receiving busy signals. In such situations companies suffer the immediate costs of lost revenue and the longer-term costs of negative goodwill. Lost revenue and negative goodwill are shown as Service Costs in the graph to the right. The blue lines show the general relationship of how Service Costs fall as network capacity is increased.

Network Microeconomics



While increasing network capacity decreases the Service Costs it increases Network Costs. More Network Capacity means more trunk service fees, more trunk cards capital expenditures and more trunk card maintenance fees.

Pick a datapoint too low and more money will be lost due to poor customer service than on the incremental trunking to avoid it. Pick a datapoint too high and more money will be spent on trunking than the incremental improvement in customer service was worth. So a properly selected SED would lie were the sum of the Service and Network Costs is the lowest.

How Network Costs rise as network capacity increases is generally well understood. For the most part incremental Network Costs are fairly consistent across PBXs and trunk groups, though not always so. In the Bell System the cost of trunking between switches located hundreds of miles apart in rural areas can be quite different than between switches located inside the same city.

How Service Costs rise as network capacity decreases is typically challenging to determine. The Bell System was no exception. Bell System revenues were tied to rates established by Public Service/Utility Commissions. Poor service led to complaints to the commissioners. In response the commissioners would be less likely to grant the size of rate increases requested by the Bell System. The Bell System viewed any negative impact on rates as a Service Cost. Bell Labs studied and thereby knew the number of commission complaints that were generated per (n) busy signals. They also knew how many complaints it took to get the negative attention of the commissioners. So they modeled how different levels of network capacity resulted in different levels of rate increases by Public Service/Utility Commissions.

The Bell System engineered network capacity to maximize its rate increases with a minimum of network costs. As callers typically tolerate an infrequent busy signal they found engineering to the highest datapoint didn't optimize network capacity. Instead they determined engineering to the ninth (9^{th}) highest of the 96 datapoints gathered in a year tuned their network to its optimum.

The ninth (9^{th}) ordinal was used as the basis of the SED for all of the Bell Systems toll switches. This was regardless of their locality, the nature of the people in the locality or density of the population. As noted earlier, trunking costs will differ from urban to rural areas. And clearly the tolerance to busy signals is not consistent among the people served across all regions of the Bell System. Bell Labs knew the Service and Network Costs differed among its switches. They knew that using the ninth (9th) ordinal across the Bell System meant some localities were providing better service than was economically justified and others were doing the opposite. Relying solely on the 9th ordinal meant many switches were not optimized on an individual basis.

Why? It goes back to labor costs. Bell Labs determined that it would cost more in additional labor to gather and analyze the data necessary to determine switch specific SEDs than would be saved in Service and Network Costs at less than optimized sites. So Bell Labs engineered the "system" to its optimum, sacrificing the Service and Network Cost performance on a site-by-site basis to save labor expense. But now with today's technology it is now possible to cost effectively collect and analyze the data necessary to optimize each trunk group of each PBX without compromise.

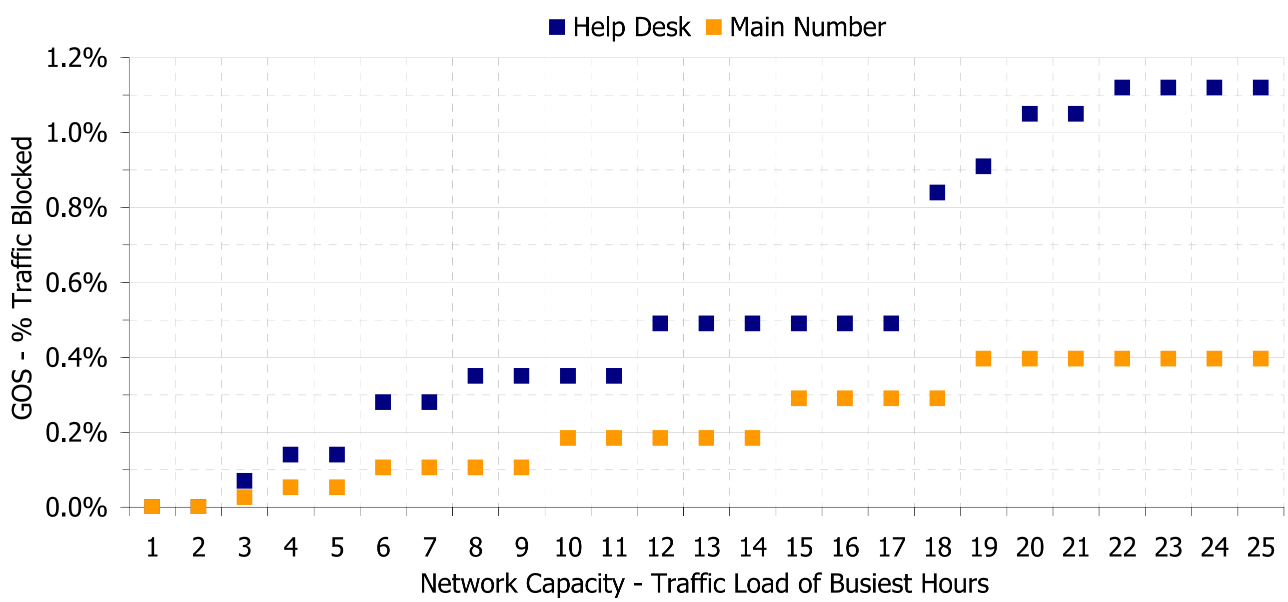
Service Costs vs. Network Capacity

Determining the incremental change in Service Costs per incremental change in network capacity is rarely easy. One way, but by no means the only way, to simplify the task is to breakdown Service Cost into two factors:

- the unit value of the traffic carried by each trunk group, and,
- the incremental change in traffic blocked for a given incremental change in traffic volume

The value of the traffic carried by various trunk groups can fluctuate greatly. Trunk groups serving revenue generating call centers would typically be considered to be of very high value. Credit card companies place a much higher value on calls from their elite card customers than calls from those holding their basic card. The traffic into a Help Desk called by internal employees might be viewed as having an even lower value. There are many examples, which can be extreme given the nature of a company's business.

The incremental change in traffic blocked for a given incremental change in traffic volume differs based on the traffic profile of each group. The graph below shows the GOS, as measured by the %Traffic Blocked over a year, when Network Capacity is equal to any one of the highest twenty-five (25) hours of traffic volume for that year. If Network Capacity is engineered to the busiest hour of the year (a.k.a. the first ordinal) one would expect that zero, or nearly zero, percent of the traffic would be blocked. As the busy hour ordinal increases so does the percent of traffic blocked.



The same ordinal yields different results for the two trunk groups. At the twentieth (20th) ordinal the Main Number would have blocked 0.40% of its traffic for the year. At that same ordinal the Help Desk would have blocked 1.05% of its traffic. So if both trunk groups were engineered to the twentieth (20th) busiest hour, callers to the Help Desk would have been over two and a half (2.5) times more likely to get a busy signal than the callers to the Main Number. This increases to a little over three (3) times if the eighth (8th) ordinal was used.

To engineer the two trunk groups to achieve similar levels of service, different busy hour ordinals would have to be used. To run a Percent Blockage of 0.4% or less, the traffic volume from the 19th through the 25th+ ordinal would be used for the Main Number. For the Help Desk, the 8th though the 11th would be used.

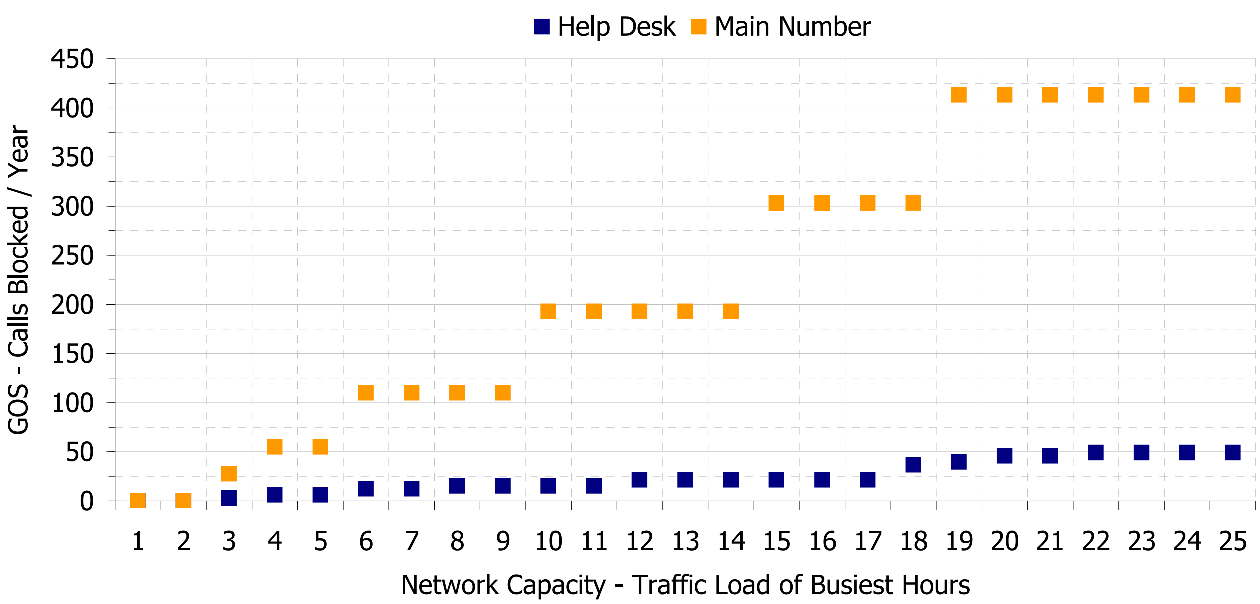
Given the same traffic values, the Service Cost for two trunk groups may differ based on their traffic profile alone. Given similar traffic profiles the Service Cost may differ based on differing traffic values. Given that one or both of these factors will differ among trunk groups, it is highly unlikely that each would be optimized basing its SED on the same ordinal.

Think Calls Blocked

Traditional traffic engineering methodologies use Percent Blockage as a metric for GOS. It's a fairly simple notion dating back to, if not before, Agner Erlang's study of telephony probabilities in the early 1900's. Erlang developed mathematical formulas relating traffic volume, network capacity and GOS. Given any two parameters one can calculate the third.

Lacking any automated computing power, simple methodologies were developed for using Erlang's formulas. Erlang's formulas were often reduced to tables. Looking up traffic volume on one axis and GOS (such as 1% blockage a.k.a. p.01) on the other, one would find the required network capacity at their intersection. So once a SED was determined, translating it into trunking requirements was straightforward.

These simple methodologies remain mainstream to this day but are they effective? Is a p.01 GOS, as is the industry norm, optimal for all trunk groups? Is the use of a common GOS, p.01 or something else, optimal for all trunk groups? While Percent Blockage is a meaningful metric perhaps it isn't the most useful.

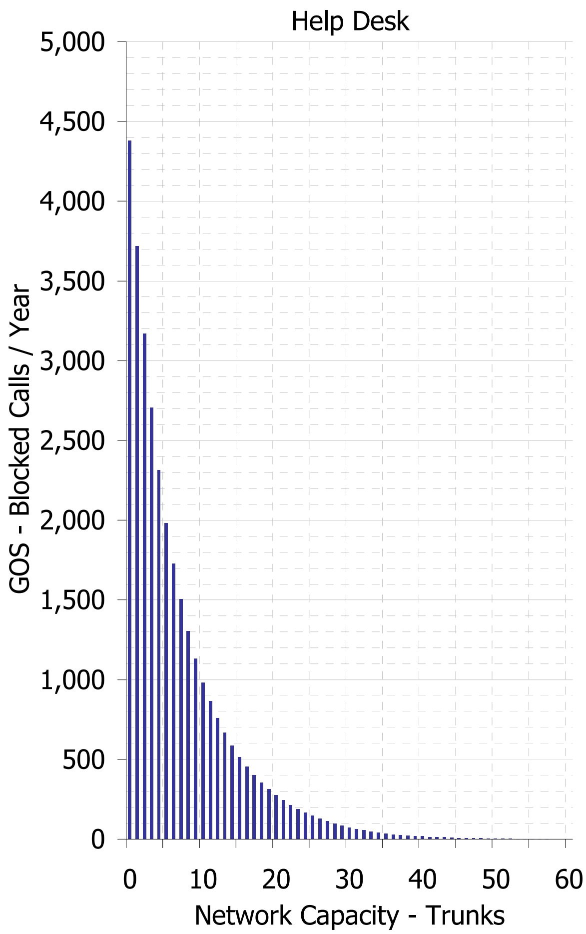


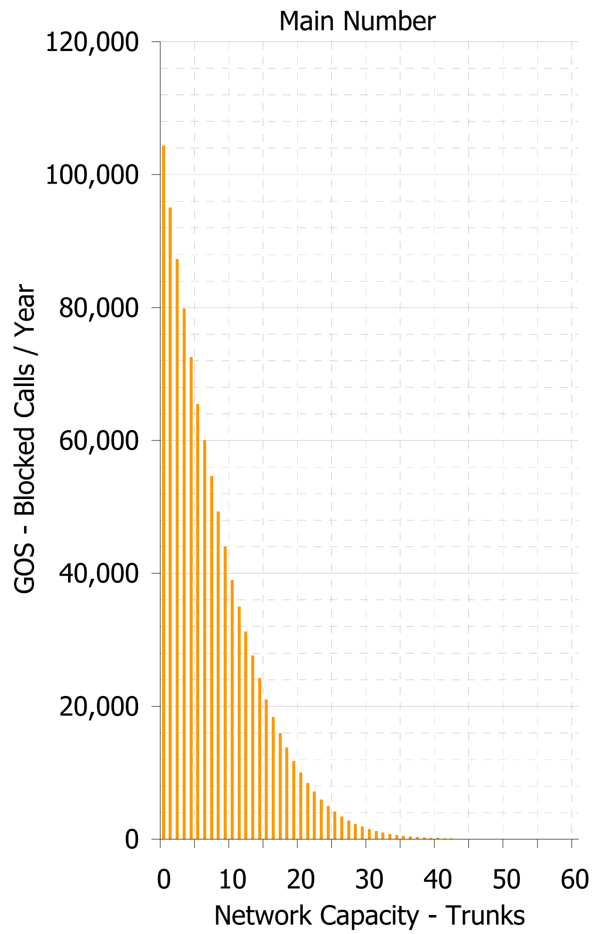
The graph above is the same as the previous with the exception that the number of Calls Blocked is used instead of % Traffic Blocked as the metric for GOS. Now you see a very different picture. The previous graph portrays the Help Desk GOS rising much faster than the Main Number's. This graph shows just the opposite. Here the GOS, as measured by the number of Blocked Calls, rises much faster for the Main Number than the Help Desk. Again note the 20^{th} Busy Hour ordinal. The previous graph gives the notion that the Help Desk GOS would have been ~2.5 times worse than the Main Number GOS if both were engineered to the same busy hour ordinal. This graph shows that the Main Number Trunk Group would block 413 calls. This is nearly 9 times the Main Number Trunk Group, which would block only 46 calls.

Both answers are technically correct. The question is which is the easiest from which to make an optimal network capacity decision and communicate them effectively to all levels of management.

Move Past Busy Hour

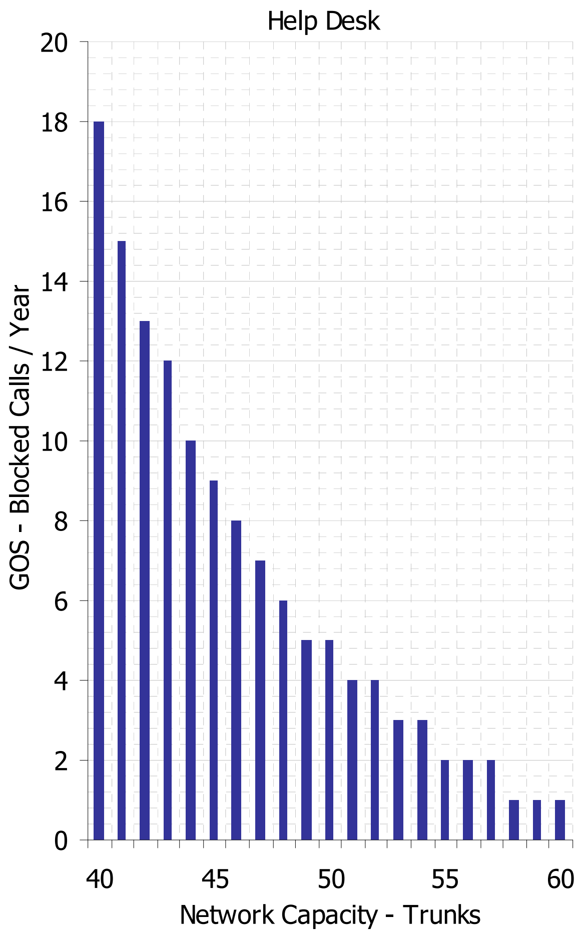
The next step in simplifying network engineering is to replace the concept of busy hour ordinals with trunk quantity as the metric for network capacity. The graphs below show as Trunks are increased how Blocked Calls decrease. The quantity of Blocked Calls is for a yearlong period.

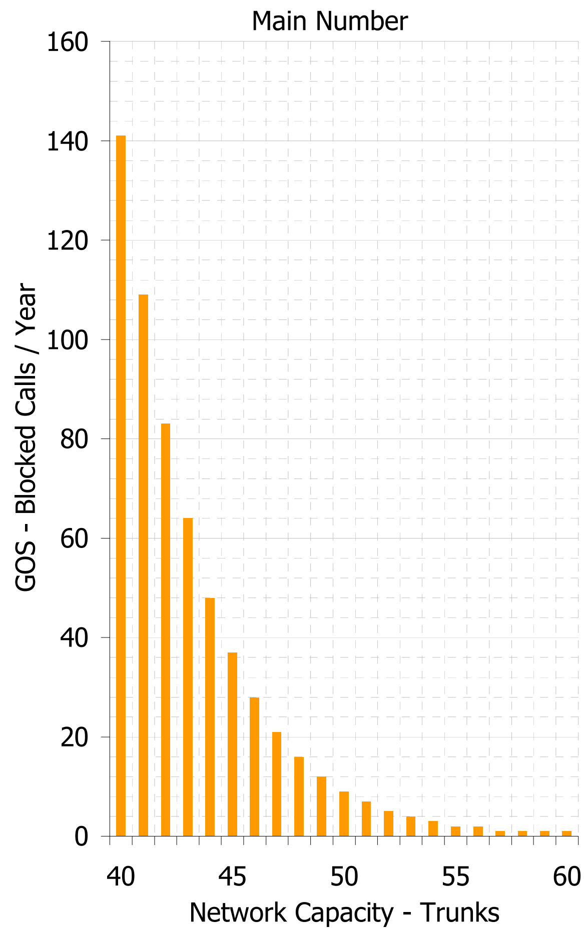




The leftmost values on the graphs explain why the previous two graphs presented opposing assessments. The Main Number had over 20 times the number of calls that the Help Desk did during the same period. Therefore a smaller %Blockage could produce a greater number of Blocked Calls. So while both %Blockage and Calls Blocked are valid GOS metrics they can present very different pictures of the same situation. Companies should consider which metric is easiest for them to make and communicate network capacity decisions.

Isolating the Trunk quantity to the high values reveals the levels of Blocked Call Attempts from which Network Capacity decisions can be made. With 44 Trunks the Help Desk would block only ten (10) calls in a year. To reduce Blocked Call Attempts to just one (1) would require 58 Trunks. At an annual unit cost of $408 the additional 14 Trunks would cost $5,712. This would save just nine (9) Blocked Call Attempts in a year.





For the Main Number a quantity of 44 Trunks would result in 48 Blocked Call Attempts in a year. An addition of 14 Trunks would reduce the Blocked Call Attempts by 47.

One can see some simple tradeoffs inside a trunk group and across trunk groups. If a company has 14 trunks of network capacity to allocate, do they want to reduce nine (9) Help Desk Blocked Call Attempts or 48 for the Main Number, a ratio of about 5 to 1? So are Help Desk calls worth five (5) times those to the Main Number? This question seems much easier to answer than what %Blockage and Busy Hour Ordinal should be used to optimize network capacity.

Leverage Today's Technology

Agner Erlang's work in the probability of telephone network behavior was a significant step forward for its time. It has been so revered that network engineers have been schooled in it for almost 100 years. Telephone companies around the world have developed elaborate methodologies based on Erlang's mathematics to engineer their vast networks.

It is important though to separate Erlang's mathematics from the data collection and analysis methodologies. Erlang would be distraught at the thought of his mathematics being used as part of a 5-Day Traffic Study. Erlang would never have expected his mathematics to be used in a manner that results in a 36% surplus in network capacity. Given the technology of the day, Erlang probably would have been comfortable with Bell Labs' approach for determining their SED. Given today's technology Erlang would expect his mathematics to be used without compromised data collection and analysis methodologies.

With today's technology, new data collection and analysis methodologies can be used. Large samples of data can easily be collected and analyzed to dramatically improve the reliability of the decisions drawn from it. Network capacities can now be optimized on a trunk group by trunk group basis. No longer do results have to be compromised because of generic approaches. Network Managers now can engineer their networks with confidence. They can even engage senior management in an understandable manner to assure the network is tuned to the corporate strategies. With today's technology, the uncertainty that leads to highly over-trunked networks can be a thing of the past.

Solutions based on today's technology can be found in a couple of forms. One is Software Tools. The other is through Managed Service Bureaus. With either solution companies should look for:

- 24 x 365 data collection and analysis, and,
- alarm notifications when systems are operating outside of expected parameters.

Software Tools are typically installed at a company's facility and used directly by a company's personnel. Software Tools provide the most robust basis for performing network capacity analysis. Companies with dynamic networks will want to acquire a Software Tool. With it, they will be able to model the effect of growth, changes in traffic mix, changes in call lengths and varying GOS levels. Software Tools can combine the traffic load from many trunk groups into one, from the same or different PBXs, to study changes in network design.

Managed Service Bureaus collect the data from a company's sites, analyze the network requirements and forward a report on a periodic basis. Managed Service Bureaus minimize the amount of hardware required at a company's facility and, the expertise and time required of company personnel.

Solutions based on today's technology deliver network capacity analysis without compromise. Network Managers can make network capacity decisions with confidence. Network Managers can collaborate with senior managers regarding network capacity decisions. Companies can enjoy the economics of a highly optimized network.

IMPACT technologies

16647 Chesterfield Grove
St. Louis, MO 63005
voice: 314.743.1400
fax: 314.743.1401
web: www.impacttech.com