Capacity Management and Optimization of Voice Traffic

Executive Summary

Preface

Today's commerce requires the interconnection of modern business communication systems with legacy voice networks. Demanding legacy requirements of a voice network can rapidly consume scarce capital and operational expense funds if not properly managed. An optimized network can ensure that resources are judiciously engineered and thus mind unnecessary expenditures from your organization's budget. This paper provides a high-level briefing of the leading practices that have application to:

- Network scaling in concert with business agility
- Change stemming from consolidation and mergers
- Managing the effect of discontinuous change

The paper introduces the application of capacity management techniques for an evolving voice network application. These practices have been applied to leading carriers resulting in huge quantifiable savings. Network managers can apply these techniques to ensure their network meets or exceeds their customer's requirements at minimum operating cost.

Introduction

It is imperative for today's networks to have security and safeguards to evolve in step with today's changing business climate. Many factors can rapidly alter the engineering norm. Networks integrate a variety of multi-vendor application specific servers, connection facilities and software applications. External stimuli in the form of failure, rapid growth and traffic variations can alter your network's ability to complete successful transactions. The paper applies leading capacity management principles to the scaling of a voice network. For introductory purposes, the paper focuses on the connection within the standard North American public telephone network. The techniques are applicable to end to end VoIP networks as well and other applications where the customer offering relies on a set of network servers spread throughout the enterprise.

Historical Overview

Telephony service today allows virtually anytime, anywhere communication via wireless, tresterial, microwave and satellite technologies. With any established application, there are legacy systems and networks that although outdated are necessary for ubiqitous service. Inter-connection with the embedded public telephone network can be complex for operators as they integrate to standard telephony elements, emergency service centers, law enforcement agencies, directory assistance, operator services and other telephone standard services. Network capacity must proceed lock step with customer acquisition to avoid call blocking, service denials and intermittent service quality. Voice traffic is random in nature and competes for a limited number of shared resources or voice channels. Various methods have evolved over the years and have lead to industry standards for management techniques. This paper describes the application of those techiques applied to a changing volumes of traffic offered to a voice network.

Leading practices suggest that effective network management include the following:

- **Forecasting and Modeling –** Typically the point of sale function within the enterprise provides an estimate of the business volume over time. Telephone service providers usually predict customer additions, disconnect and net volume over time. Standard customer behavior parameters (e.g. telephone usage frequency and average call duration) can then provide the basis to predict network capacities needed to meet the anticipated customer demand.
- **Trending and Analyzing –** Any model is strengthened with empirical field data. Actual customer usage should be monitored to refine the model thereby improving its ability to predict future needs. Any variations of significance need to be analyzed to determine if all samples are viable. In telephone networks, significant variations can be observed for high traffic days (e.g. Mothers Day, national disaster or radio/television calling campaign). These variations could artificially impact the desired long term engineering results.
- **Augmenting and Optimizing –** Reports and modeling provide insight into resource utilization. Any change, addition or reductions to the operating limit of the resource make require several days to accomplish in voice networks. Thus, it is important to model the augmentation process such that appropriate warning and alert levels can be established to provide the operator sufficient lead time to react to changing conditions. Leading operators define both low and high water marks for critical network resources to allow for additions when needed and right sizing of under utilized resources. The later can have a large impact on recurring network expenditures.

Business Needs

Integrating the above practices into the management of today’s multi-service networks allows operators to ensure that available capacity in tightly coupled to the business forecast. Additionally, these leading practices improve the enterprise’s agility to adapt to environmental changes and external stimuli that typically result in discontinuous change.

Target Market

The principles described in the paper are applicable to all situations where network resources can potentially impact or severe the desired user performance. Although, the application of these practices in a traditional telephone network is described in this paper, the practices can be applied to VoIP networks, any service offering involving network, servers and applications, or a specific network element of finite capacity. Capacity issues are not always resolvable by throwing additional capacity at it. Network throughput, reliabilty, perfromance and efficiency are all manadatory to sustain or grow market share.

Solution/Technology Description

A hypothetical example is used here to scope the challenge and provide background information. An existing cable television company has made a sizable investment and now has an addressable market of 30,000,000 households passed (HHP). 40% of the addressable market has subscribed to their CATV services. To increase average revenue per user (ARPU), the cable company has expanded their portfolio to include pay per view, video on demand and high definition broadcasting. ARPU can be increased again by offering telephone over the embedded cable infrastructure. Company executives establish a target of deploying telephone services to 40% of their customer base within four years. To achieve the goal, an average of an additional 100,000 subscribers must subscribe to the telephony services each month.

Landline telephone users in the United States typically use between 500 and 750 minutes of call time monthly. Standard telephone engineering rules suggest that on average, 5 to 7.5 minutes of use per customer should be expected in the busy hour each day. The standard connection in the PSTN is a T1 which contains 24 separate voice channels. Traditional engineering guidelines for telephony suggest a T1 engineering capacity of 850 minutes of usage to achieve the PSTN standard grade of service of 0.5% call blocking. For demonstration purposes, each subscriber addition will demand 6.25 minutes of usage on that T1. Thus every 135 subscribers will necessitate an additional T1. In practice, the call minutes are typically spread across multiple connections based on specific communities of interest. The table below shows a typical growth scenario for our illustrative service provider.

Table 1. Minutes of Use Scaling in Concert with Business Growth

Subscribers	Local	Long Distance	Wireless	Voice Mail	8xx	Other	Total MOUs
	40%	30%	15%	5%	3%	7%	
1	3	2	1	0	0	0	6
125	313	234	117	39	23	55	781
250	625	469	234	78	47	109	1,563
500	**1,250**	**938**	469	456	94	219	3,125
1,000	**2,500**	**1,875**	**938**	313	188	438	6,250
1,500	**3,750**	**2,813**	**1,406**	469	281	656	9,375
2,500	**6,250**	**4,688**	**2,344**	781	469	**1,094**	15,625

The highlighted areas in the table indicate logical call groupings where augments are required to handle the growing telephony needs. If the above company is operating in 80 markets, then 300 subscribers will be added weekly in each market. It is imperative for the operator to have accurate forecasts for business growth so that trunk augmenting can scale in concert with the business. The techniques described in this paper are proven leading practices that simplify the role of the operator.

Network Design Practices

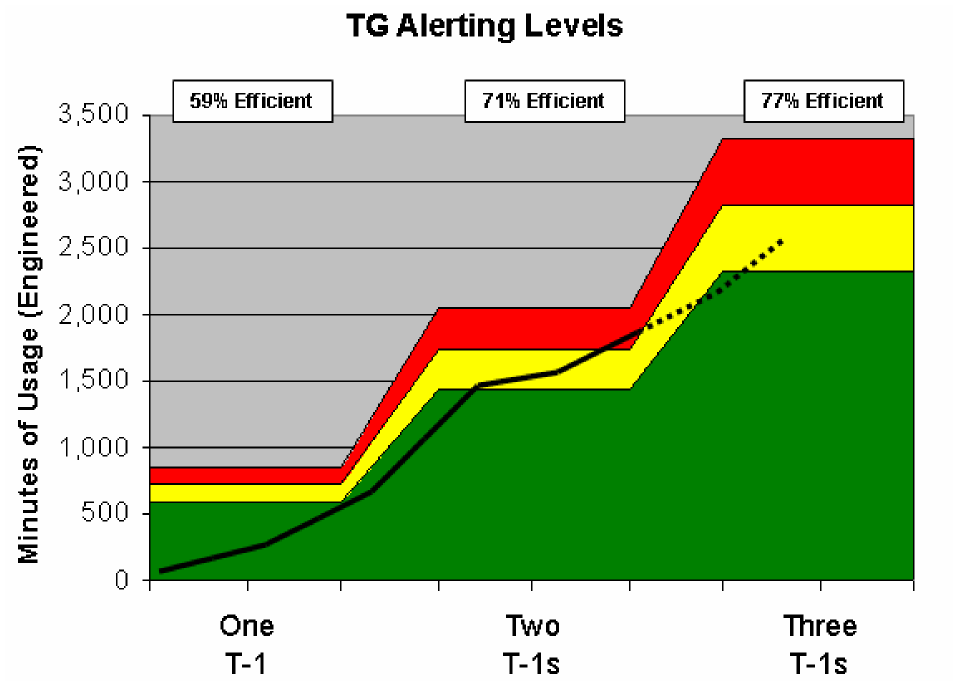
Networks are built based on an engineered size required for the business over some specified planning horizon. During normal operation, key network instrumentation provides insight into the utilization of these engineered resources. As various thresholds are reached, alarms alert operations personnel to initiate appropriate review and augmentation activities. Augmentation strategies typically depend on an incremental size and intervals. Augmentation intervals should encompass approval and purchasing steps through installation, test and turn-up. In practice, augmentation intervals can vary depending on several parameters (e.g. the availability of the associate transport facilities). Network optimization entails carrying the maximum payload across the network at minimum costs while providing the desired grade of service to the end consumer. Engineering decisions clearly influence both the service levels experienced by the customer base and the underlying costs. This paper covers key aspects of service provider voice capacity management and highlights critical optimization opportunities.

Capacity Management Simplified

Networks consist of resources working in concert to provide an end-to-end service offering. These network resources span customer premises equipment, access and transport nodes, inter-connection facilities and a backbone network. These elements are typically grouped into access, aggregation and core layers. Much of today's voice traffic has migrated to wireless, cable, internet and other media while the traditional Public Switched Telephone Network (PSTN) provides legacy voice connectivity. Voice networks virtually transport packets of information from sender to receiver, and vice versa. The PSTN employs discrete packets of information at fixed time intervals. This technology, termed Time Division Multiplexed (TDM) traverses digital pipes throughout the PSTN over digital transport facilities. A T1 is a 1.544-Mbps transport typically configured as 24 separate voice channels and is the minimal connectivity between two nodes in the PSTN. A single T1 is capable of carrying 24 simultaneous TDM voice calls. The same T1 facility can be configured to carry significantly more calls using VoIP technologies that use various compression techniques and eliminate dedicated bandwidth consumption for periods of silence in normal telephone conversation.

Standard industry guidelines specify that a single T1 in the PSTN should be expected to carry about 850 minutes of call traffic or about 59% of the available minutes for a fully loaded T1. These minutes of usage (MOUs) include call establishment times, alerting, talking, disconnect and call tear down. A detailed explanation of the engineered T1 capacities is provided in Appendix I. For our purposes, this resource has an engineered limit as depicted in the figure below. As business scales and more call traffic is offered, it may become necessary to augment the transport with additional T1s. A second T1 elevates the engineered capacity of the transport to 2,050 available minutes of call traffic. There is a 20% efficiency improvement with the increased size of the pipe.

Figure 1. Trunk Group Alerting Levels



As part of the resource augmentation process, alarm levels should be adjusted. These alarms are designed to alert operations personnel when resource utilization reaches various severity levels of concern. Each augment effectively alters the engineered limit of the particular network resource (augment increment) and requires some amount of implementation time interval (augment interval) based on a variety of factors. Capacity management is simply the continuous monitoring of the consumption of these network resources and taking appropriate augmentation actions. In practice, various techniques are used to forecast or predict how network resources will be consumed as the business evolves. Trending information is then used to compare against the forecast. Capacity engineers may re-program the alarm levels as they gain experience with there augmentation practices. Carrier grade voice providers have a 45-day target for most augments. Alarm levels should be set to allow sufficient lead time to perform an augment before reaching 100% resource consumption as the business continues to scale. The line in Figure 1 depicts resource consumption experienced over a 5-month period and forecasts the following two months.

Marketing divisions of large service providers submit business plans that provide roadmaps for subscriber acquisition, retention and growth in their focused areas. While individual telephony use is mostly random and can vary substantially among subscribers, voice networks are typically engineered with an average subscriber usage. A subscriber forecast in conjunction with this usage can predict or forecast the rate of consumption of essential resources. As shown above, minutes of usage are tracked over time to determine when an augment is required. For illustrative purposes, we used a 30-day augmentation interval. Sales campaigns, new service offerings and new market launches can affect the rate of resource consumption and warrant close attention. It is essential for the network capacity engineer to keep abreast of these market factors to ensure adequate network capacity for the offered traffic load.

Brief Overview of the PSTN

In North America, there are over 20,000 central offices or hubs that terminate traditional voice access facilities to customers. Cellular, cable, Internet telephony and other providers employ other techniques to reach subscribers; however interconnection with the traditional PSTN for the foreseeable future will be required to reach legacy telephony users. Carriers interconnect with the PSTN over groups of T1s in what is termed a Trunk Group (TG). These TGs connect two PSTN nodes or two communities of interest. TGs also provide access to special services such as voice mail, operator services, directory assistance, 8xx traffic and emergency services.

Today's service providers strategically locate their network equipment based on a variety of factors including facility robustness, environmental factors, cost and ease of PSTN interconnection. The equipment usually consists of a service authentication server, a call manager, a media gateway and other maintenance and security systems. The media gateway provides the interconnection to TDM-based PSTN facilities. These media gateways typically convert IP-based voice packets to TDM packets and transport them on T1 facilities. These T1s may be aggregated into large metallic or optical transport pipes. A single media gateway may be configured for several TGs:

- Trunk Groups connecting local communities of interest are termed Direct End Office Trunking (DEOT)s
- Network aggregation points for several local communities are configured in Tandem Switching Nodes. These provide default market ingress and egress for call traffic in cases where DEOTs either do not exist or do not have sufficient capacity to carry the offered call traffic.
- Long distance (in region and national), and international traffic is typically carried over dedicated TGs also.
- A variety of TGs exist for 8xx traffic, operator services, directory assistance, emergency 911 service and more.

Dialing rules and routing guides in the Call Manager dictate how voice packets are carried through the network. Voice networks are typically engineered for busy hour traffic volumes. For commercial services, this typically occurs during normal business hours; residential service typically peaks during the evening hours. Voice networks should be configured to accommodate a burst of call traffic due to mass calling campaigns, severe weather and significant national events.

Operations personnel monitor the availability of network resources and take corrective action for operational difficulties and outages. Capacity engineers monitor busy hour resource usage. Samples over time can provide the necessary trending information to compare against the supplied, or predicted market forecast. As the business matures, the forecast usually closely matches the observed trend information. In essence, the role of a capacity engineer can be termed "just in time capacity availability" to allow for prudent use of both capital and recurring expense funds.

Management of Voice Networks

Network nodes provide real time operational information in the form of SNMP events, syslogs and application specific reports and measurements. The reports guide operational personnel to identify and resolve faults and outages, maintain network security and performance, manage network configurations and administer network data bases. OSS tools and methodologies can correlate information from these various sources and guide operational personnel as they ensure network reliability and availability.

Management of voice services additionally requires the close management of the network application (i.e. voice). In other words, packets are transported throughout the network within quality of service (QoS) tolerances for latency, jitter and loss; however, the application may not be providing the desired grade of service (GoS). This is the case if trunk groups are not properly sized resulting in blocked calls. Routing errors and inadequate safeguards for mass calling campaigns can also impact the GoS. Incumbent voice service providers closely monitor the rate of blocked calls as there is a direct correlation with customer reported service troubles. Where telephony is regulated, these performance metrics must be reported to oversight commissions monthly.

Reports and measurements from the call manager and associated media gateway are essential for traffic analysis and capacity management. Some commercial tools are available in this space and must be customized to meet the needs of the service provider. Below is a sample custom report developed for a major VoIP service provider. The report correlates various application level service indicators and guides the traffic analyst and capacity engineer where manual intervention is required.

Figure 2. Trunk Group Alerting Levels

MMPXTNPS3	07/02/07 Usage: 55,436 Lines, 10,728 Trunks									12/29/07 Forecast: 73,653 Lines, 10,728 Trunks		
Trunk Group Identity	DS0s	T-1s	CCS Capacity	%IN	%OUT	IN CCS	OUT CCS	Total CCS	%Util	Planned CCS Usage	%Util (Planned)	TG Utilization Grey is Baseline; Black is Forecasted
1001_CRVLTNMADS0	48	2	1,233	52%	48%	504	404	908	74%	1,206	98%	
1002_CVTNTNMTDS1	120	5	3,578	45%	55%	67	82	149	4%	198	6%	
1003_MMPHTNBADS0	240	10	7,660	50%	50%	2,510	2,483	4,993	65%	6,634	87%	
1004_MMPHTNCTDS0	72	3	1,996	49%	51%	695	721	1,416	71%	1,881	94%	
1005_MMPHTNELDS0	72	3	1,996	43%	57%	345	802	1,147	57%	1,524	76%	
1006_MMPHTNGTDS0	144	6	4,383	45%	55%	1,378	1,707	3,085	70%	4,099	94%	
1007_MMPHTNMADS0	168	7	5,196	22%	78%	148	531	679	13%	902	17%	
1008_MMPHTNMADS1	240	10	7,660	46%	54%	357	422	779	10%	1,035	14%	
1009_MMPHTNMTDS0	120	5	3,578	68%	32%	1,364	630	1,994	56%	2,649	74%	
1010_MMPHTNOADS1	192	8	6,014	48%	52%	1,633	1,946	3,579	60%	4,755	79%	
1011_MMPHTNSLDS0	144	6	4,383	47%	53%	1,257	1,294	2,551	58%	3,389	77%	
1012_MMPHTNSTDS0	96	4	2,781	37%	63%	568	944	1,512	54%	2,009	72%	
1014_SOVLTNMTDS0	48	2	1,233	53%	47%	256	223	479	39%	636	52%	
1015_HRNNMSDSDS0	48	2	1,233		100%		257	257	21%	341	28%	
1016_BYHLMSXADS0	72	3	1,996	73%	27%	93	35	128	6%	170	9%	
1017_OLBRMSXADS1	456	19	15,178	87%	13%	264	38	302	2%	401	3%	
1018_JCSNTNMADS0	24	1	511		100%		279	279	55%	371	72%	
2001_MMPHTNMA84T	2,040	85	71,512	58%	42%	24,004	17,617	41,621	58%	55,298	77%	
2101_JCSNMSCP06T	768	32	26,179	88%	12%	9,970	1,318	11,288	43%	14,997	57%	
2201_MONRLAMA06T	600	25	20,242	17%	83%	1,442	6,393	7,835	39%	10,410	51%	
2202_SHPTLAMA0GT	720	30	24,480	58%	42%	7,739	5,594	13,333	54%	17,714	72%	
4001_NWORLAMA0GT	240	10	7,660	73%	27%	428	160	588	8%	781	10%	
8501_ATLNGAHPG0C	1,344	56	46,662		100%		32,431	32,431	70%	43,088	92%	

This report provides a busy hour baseline summary for the call traffic in a VoIP market served by the Cisco® BTS 10200 SoftSwitch. Standard reports and measurements in conjunction with the service provider subscriber roadmap are used to portray current and projected usage. In this example, the switch scales from 55,436 lines at the beginning of the period to over 73,000 lines in six months. This particular report focused on trunking. Other critical system resources (e.g. CPU, memory, disk, software and data base limits, licenses, media terminations and transport facilities) must be monitored in a similar manner.

Trunk Groups beginning with digit one (1) are DEOT groups connecting to focused communities of interest. TGs beginning with digit two (2) are Tandem groups for local traffic aggregation. The TG beginning with digit four (4) is used to terminate all inbound long distance calling and also 8xx calling. The TG beginning with digit eight (8) carries all outbound long distance calls. Other characteristics of the report include:

1. A section providing the baseline network usage, directional flow and percent of engineered capacity currently utilized in a busy hour.
2. The forecast section contains the planned usage, extrapolated from the baseline usage and scaled according to the market supplier subscriber roadmap.

3. A visual scale of TG utilization shows the current and forecasted resource consumption level for each TG. Note that alarm levels vary according to the type of TG to meet the needs of this service provider. In practice, DEOTs are designed to run at very high utilization levels since denied or overflow calls attempting to traverse the DEOT will be routed to the corresponding Tandem aggregation point.

A close look at the example by traffic analysts and capacity engineers may identify several work items. The following actionable issues are extracted from this particular sample:

- Possible switch routing error since virtually no traffic is traversing TG 1016 and TG 1017 in the report period.
- Excess capacity in several TGs (1002, 1007 and 1008) since they reach less then 50% of their engineered capacity at the end of the planning horizon. In this case, 22 T1s have been built to accommodate traffic that could easily be handled with 4 T1s (see Optimization Section below).
- Possible switch routing errors and corresponding service problems in the connecting network node for TG 1015 since there is no evidence of inbound traffic traversing this group. Note that local DEOTs are engineered as high usage two-way groups handling both ingress and egress traffic.
- Need for an augment of TG 8501 to allow its engineered capacity to scale in concert with the subscriber roadmap.

Network Optimization

The previous section describes how real time network reports and measurements can be used to manage the performance of the voice application and thus the corresponding grade of service experienced by the customer base. A few critical areas that have a direct impact on the cost of providing the service and resultant gross margins are now described.

At this point, we need to assume that the underlying data network is providing carrier grade reliability and availability. As shown above, several additional factors can influence the application-level service performance. Voice and voice applications directly benefit from a few optimization areas.

Increased Utilization

Unused Trunk Groups capacity is cost that should be avoided. Right-sizing TGs in concert with market scaling can minimize these costs. In addition to daily fluctuations, voice traffic may have monthly, seasonal and holiday variances. The service provider needs to establish a busy hour and day as the basis. It is not necessary to attempt to engineer a voice network to handle all the calls offered on a holiday or in response to a calling campaign.

There are significant costs associated with unused trunk capacity. As noted in the example above, three TGs have a total of 18 T1s that are not planned to carry traffic in the sample study period. Each of these T1s has a monthly recurring cost of approximately $200 to $300 for a total monthly recurring cost of $3,600 to $5,400. These expenditures quickly scale when under utilization exists throughout the network.

Increased TG Efficiency

Ideally, a single server is available for 60 minutes each hour. Figure 1 demonstrates that the amount of usable minutes (i.e. TG efficiency) increases as the size the Trunk Group grows. Voice networks should be designed with larger Trunk Groups to take advantage of these efficiency gains.

Traffic Analysis and Optimized Routing

Voice networks have changed significantly in the last 25 years and will continue to change going forward. Growth, landscape, opportunities and next generation IP based communication services will alter routing rules and call treatment. It is essential that the traffic engineer keep abreast of the offered call load and incorporate the necessary routing changes to maintain high TG utilization levels and larger TG pipes. It may not be possible to predict the next American Idol like impact on call volumes.

As stated in the document, there are several limiting factors that may restrict the growth of an individual network node. These must all be considered to determine the optimal and least costly configuration. A standard metric for voice networks is the Line to Trunk Concentration Ratio (LCR), also referred to as an over-subscription ratio. When a market launches, the LCR may be as low as 4.5 (commonly depicted as 4.5::1) indicating 4.5 lines compete for each trunk in the switch. However, as the market matures and customer calling behavior is validated, you should expect to see LCRs in the range of 9::1 for a predominantly residential service and lower for call managers servicing commercial enterprises.

Alternate Architectures and Interconnection Alliances

Although the fundamental subscriber behavior has not changed significantly in the last 25 years, wireless and cellular providers carry large amounts of call minutes daily. In lieu of traditional access arrangements through aggregation points and localized points of presence, it may be prudent for the voice service provider to explore alternate interconnection arrangements with these carriers to minimize cost and leverage the synergies of both providers. For example, an IP peering arrangement can be established for all traffic to a particular cellular carrier.

Conclusions

Today’s VoIP technologies are predominantly on the access side of the network, be it cable, DSL or wireless. Much of the traditional PSTN backbone has migrated to carrier grade IP-based networking. Voice providers are beginning to use their inter-connection LANs and WANs to carry on-net traffic and minimize the cost associated with handoffs to the traditional PSTN. The evolution will place more demand on the management of the voice application. It is imperative that networks be instrumented to provide application level insight for managing the voice GoS and the basic QoS as packets traverse multi-provider networks.

Consumer confidence and loyalty has required that we look beyond availability and reliability toward application level performance and capacity management. Various optimization strategies are described here for the voice application. Similar techniques apply to other service applications and are required as the market demands more SLA-based service assurances. An early focus on application performance in conjunction with service performance will provide a key differentiator at the outset, and will be required for market entry in a few short years. Service providers need to closely align with their major network equipment providers to ensure organizations design the necessary instrumentation, management processes and service/performance metrics.

These principles have been validated in two major VoIP accounts. Using the techniques described here, we have been able to:

- Reduce CapEx expenditures by over $15 million by right sizing network nodes
- Reduce annual NetEx expenditures by over $35 million by increasing trunk utilization and efficiency, and thereby reducing monthly recurring fees
- Related but not measured OpEx savings associated with reduced management requirements.

Acronyms

Acronym	Definition
OSS	Operations Support System
PSTN	Public Switched Telephone Network
TDM	Time-division multiplexing (TDM) allows multiple data streams to be transported as a single signal. The steams are fixed length and arrive at prescribed intervals.
VoIP	Packetized voice packets transported over a data network utilizing the IP protocol.
LCR	Line Concentration Ratio. The ratio of lines to servers (trunks) in a telephone switch.

References

1. Trunk Traffic Engineering Concepts and Applications, Telecordia Document Number SR-TAP-000191, Issue No. 02, Dated Dec 1989
2. General Information and Technical Papers Available at website: www.erlang.com
3. Voice Design and Implementation Guide, Cisco Systems Document ID: 5756
4. Traffic Engineering and QOS Optimization of Integrated Voice & Data Networks, Gerald Ash, AT&T Labs

Appendix I – Erlangs Design Guidelines

Background – Long gone are the days of rotary dial phones, coin first public phones and exchange identity dialing (e.g. Spencer 9, 3776). Yet telephony usage and consumer behavior have not changed significantly in over 50 years. There was a noticeable change about 15 years ago when the internet was only accessible via dial-up modems. This stimulated many additional telephone lines and changed some fundamental network engineering parameters (e.g. average call duration). Today, most internet access has migrated to high speed media returning telephony to plain old telephone calling.

The duration of a telephone call is measured in minutes and spans the actual talk time, rounded up in some fashion for billing purposes. Telephony engineers determine an average call hold time to encompass initial setup, alerting, talking, disconnect and tear down. Average call duration can be expected to range between 3 and 4 minutes based on the market characteristics (i.e. residential or commercial service). Another important engineering criterion is call attempts. These include the complete call with talk time and also misdials, invalid dials, partial dials and abandon attempts during call setup. Call attempts place a demand on the call manager and can consume scarce CPU processor availability.

Telephony engineers measure call duration is Centum Call Seconds (CCS) or hundreds of call seconds. Our average call hold time of 3.5 minutes is 2.1 CCS. Busy hour subscriber usage is typically measured in CCS for bandwidth sizing and call attempts for determining required CPU capacities.

On-net, local calls are usually served within the call manager and do not require interconnection with the PSTN. Call traffic to and from legacy PSTN subscribers is typically carried over digital facilities. These usually start with 24 separate channels or digital paths multiplexed on to a single T1 facility. The T1 is in effect 24 separate servers. A T1 fully loaded can handle 864 CCS (24 servers x 60 minutes per server hour x 60 seconds per minute / 100). In practice, 864 CCS for a T1 is rarely experienced. This is due to a variety of factors such as:

- Calls do not typically arrive immediately after the prior call completes
- Call establishment and tear-down times can consume 2-3% of available capacity
- All 24 servers may not be fully operational and available for the entire period

In general, call patterns are random in arrival times and duration. Service providers are able to draw upon many years of telephony studies and consumer behavior. These resulted into a set of Erlang engineering guidelines.

The Erlang Contribution

Intense study of consumer behavior, call handling systems and call traffic under a variety of conditions have lead to a well defined set of design expectations of "N" servers. These are based on a few key parameters:

- Treatment for blocked/denied calls (queue for service or deny the call attempt)
- Consumer behavior for blocked/denied calls (a subset will retry the call immediately)
- Service provider's desired probability of blocking to achieve the corresponding grade of service offered
- Number of servers planned

Voice networks are typically design parameters based on an **Extended Erlang B Model** using a P005 blocking criteria. That is, it is permissible for 5 calls per 1,000 to encounter blocking and thus require a user reattempt or redial. Some high-usage facilities are designed to accommodate the bulk of the traffic load between two communities and overflow to higher aggregation points. These high-usage facilities may be designed at higher blocking levels (e.g. P12 or 12% blocking).

The Erlang Model sizes server groups in what is termed an Erlang. An erlang is a measure of call handling capacity as follows: one server fully occupied for one entire hour is equivalent to one erlang. Erlangs can be derived from CCS, and vice versa as follows:

CCS / 36 = Erlangs

Erlangs x 36 = CCS

For a single T1 with 24 separate channels and a P005 blocking probability, the Extended Erlang B Model recommends an engineering limit of 14.2 erlangs or 511 CCS. Recall from the above discussion a fully loaded T1 can handle 864 CCS. Thus, in order for a single T1 to provide the desired grade of service (GOS), engineers should not design for more then 511 CCS. In practice, the T1 may accommodate the additional CCS within the desired GOS, but the randomness of call durations and arrivals indicate that is become highly probable to encounter call blocking as more CSS is offered. Note that a single T1 operates at 60% efficiency (i.e. equivalent of 60 cents of every investment dollar at work). Efficiency is defined as:

Erlangs for N servers / (Number of T1 servers x 24)

Sample T1 efficiencies at P005:

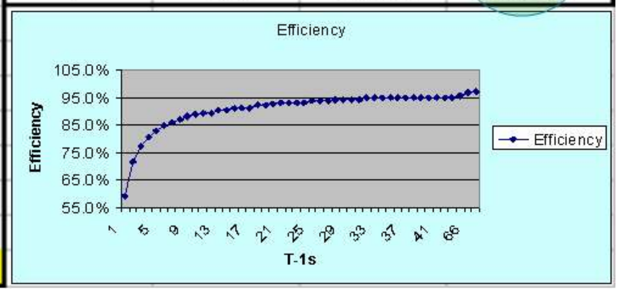
One T1: 14.2/(1 x 24) = 59.2%

Seven T1s: 143.8/(7 x 24) = 85.6%

Larger grouping of T1s are more efficient, more desirable and usually available at lower costs. The table below shows the affect of varying the call blocking criteria from the traditional P005 to P14. Note that at P14 a single T1 is very efficient (minimal bandwidth unused), however 14% of the calls are denied service. The table also shows the increase in efficiency as multiple T1s are combined while holding the P005 blocking criteria constant.

These guidelines should be taken into consideration when sizing interconnection facilities. While this overview covers discrete TDM voice packets for simplification purposes, the methodology and principles are directly applicable to compressed voice packet technology (e.g. VoIP networks)

One T-1			Multiple T-1s			
DS0s	Blocking	Erlangs	Number	DS0s	Erlangs	Efficiency
24	0.140	23.65	1	24	14.20	59.2%
24	0.130	23.20	2	48	34.20	71.3%
24	0.120	22.75	3	72	55.45	77.0%
24	0.110	22.25	4	96	77.20	80.4%
24	0.100	21.75	5	120	99.35	82.8%
24	0.090	21.25	6	144	121.75	84.5%
24	0.080	20.75	7	168	144.30	85.9%
24	0.070	20.20				
24	0.060	19.60				
24	0.050	19.00				
24	0.040	18.35				
24	0.030	17.55				
24	0.020	16.60				
24	0.010	15.25				
24	0.005	14.20				



Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
168 Robinson Road
#28-01 Capital Tower
Singapore 068912
www.cisco.com
Tel: +65 6317 7777
Fax: +65 6317 7799

Europe Headquarters
Cisco Systems International BV
Haarlerbergpark
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands
www-europe.cisco.com
Tel: +31 0 800 020 0791
Fax: +31 0 20 357 1100

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Printed in USA C11-439774-00 11/07

