2007 SCTE Cable-Tec Expo

VoIP Scalability – Tiger by the Tail?

How to Scale VoIP to Best Accommodate Subscriber Growth

Cisco

Bil Dry, Technical Leader

Version 1.0

Corporate Headquarters
Cisco, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

Document History

Version No.	Issue Date	Status	Reason for Change
1.0	30-March-2007	Released	First release

Change Forecast: <Low>

This document will be kept under revision control.

A printed copy of this document is considered uncontrolled.

How to Scale VoIP to Best Accommodate Subscriber Growth

Delayed dial tone, annoying pops and clicks, one-way voice and incorrect billing

Annoyances like these afflict VoIP customers when their provider's VoIP network is swamped with traffic without the capacity, processes and systems to deliver it. This paper discusses capacity and scale leading practices to help MSO's and other service providers' design, deploy and operate large scale VoIP networks. It begins with a discussion of general aspects of VoIP capacity planning then a deeper look at VoIP call control scaling techniques followed by sizing and growth considerations for the HFC, IP and TDM portions of the network.

Overall VoIP Capacity Planning & Scaling Considerations

While VoIP customers are using the service, receiving bills and occasionally contacting customer care, the operation and interaction of various systems and processes like IP address allocation for new VoIP network components, intelligent, load-based call routing to minimize blockage and dial plan additions to facilitate new calling patterns must function in harmony to ensure user experiences meet expectations. VoIP network and process growth spurred by the constant influx of new customers should be a routine task that ensures the network and its adjunct systems are ready for any subscriber surge. A thorough capacity planning process provides insight and analysis of growth assumptions then makes measured steps to augment the network based on the analysis.

Capacity planning distills into key steps. The steps are planning, monitoring, alerting, augmenting, trending and forecasting. Forecasting, the final step, feeds data into planning, the first step, to complete the loop and deliver critical feedback. These steps are sequential and rely on the actions and results of prior steps. For instance, trending cannot occur without the data collected during monitoring. While monitoring occurs continuously, over the life of the VoIP network, the capacity planning process pulls the data from monitoring in intervals for the analysis, trending and forecasting steps. Figure 1 shows the capacity planning steps in context of their feedback loop.

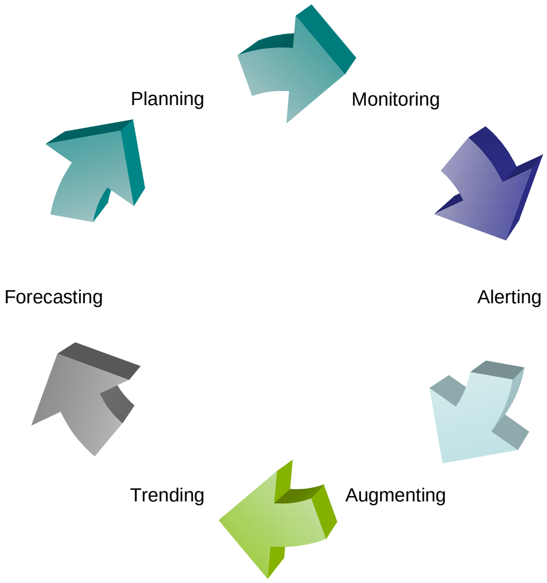


Figure 1 – The cycle of VoIP capacity planning

Capacity Planning Steps Defined

- Planning – Looks at anticipated load and defines the necessary increases to components, links and systems to handle the predicted load.
- Monitoring – Surveys current consumption and usage rates then compares the actual load to the engineered limits
- Alerting – Notifies staff when usage rates or volumes surpass predefined levels. These levels are fractions of the engineered limit and crossing levels signals capacity planning staff to act. Alerts are set at the device, link and system levels. Augmenting – Grows the engineered limit incrementally.
- Trending – Documents growth and provides insight to the direction of growth
- Forecasting - Predicts future growth and translates growth predictions into estimated utilization rates

Carefully conducted capacity planning lets the operator know the engineered limits of the overall network and its individual components. The operator can set the proper augmentation intervals for all components using this insight plus various component consumption rates and growth trends identified during the VoIP capacity planning process.

A Closer Look at VoIP Capacity Planning

Begin with a business plan that predicts subscriber take rates, growth rates and call/usage patterns. Use those rates and patterns to size and build the access, aggregation and interconnect portions of the VoIP architecture. Likewise, construct back office systems to handle the volume of provisioning, billing and trouble ticketing initially and in the longer term.

Seek out the capacity limitations of every device, link and system in the network delivering VoIP service. Approach the limit analysis from multiple directions. Broaden the analysis past the bandwidth of the link or the CPU utilization of the device to include the needs of every function within VoIP service delivery. For example, consider operational angles such as simultaneous provisioning sessions supported. If the CMS only supports eight CLI sessions but permits much higher provisioning speeds using CORBA or SOAP interfaces, consider a provisioning system upgrade to take advantage of these bulk interfaces. Time to market considerations may have dictated the use of CLI provisioning to add new subscribers to the CMS' database but the demands of a growing customer count translate to larger call centers, more order entry and the requirement to take advantage of the CMS' higher capacity bulk interfaces for provisioning information exchange.

Identify the forcing factor(s) in the augmentation process. Is it circuit delivery? Is it installation time? Is it new facility certification? Continue this brainstorming exercise with additional questions to decide the forcing factor for every device, link or system that must grow to support subscriber growth. Use these factors to determine your augment lead times and augment sizes.

The augment size increment should last the augment lead-time plus the growth rate period. For success, the actual load must not cross the engineered limit. Know the engineered limit, understand how to grow to it and act to grow it in accord with known augmentation times to avoid unwanted consequences such as blocked calls, suspended installs or customer-initiated service cancellations.

Use payback and ROI calculations to fine tune the size of capacity augments relative to the expected increase in revenues over the concerned time frame.

Step back and consider every aspect of VoIP network growth then determine its impact on every device and process involved with VoIP service delivery. Consideration and reflection on VoIP service's growth impacts should happen regularly to ensure any required change to process, procedure or components happen before limits are reached.

Scaling VoIP with an Eye on Facilities

Planning and forethought allow operators to build VoIP POP's that service current demand and easily accommodate future growth while minimizing VoIP network operation expense. For instance, know the BTU ratings of each VoIP component and determine the temperature gradient for the POP. Locate VoIP gear producing the most heat in the section of the POP with the greatest cooling and exhaust ventilation capacity if possible. This avoids the extra costs of cooling and exhausting naturally warmer sections of the POP to temperature and airflow levels required to meet the environmental requirements of VoIP gear with the greatest heat dissipation rates.

Larger floor space requirements often accompany VoIP subscriber growth. For example, assume an MSO is growing at 175K subs per month in one section of their service footprint. At 0.1 Erlang per sub and 14 Erlang per T1 using 0.5% blocking, this subscriber growth rate translates to 1250 T1's, nearly 4 OC-3's, of new trunk gateway capacity needed every month.

Determine the additional VoIP components needed to scale such as trunk gateways or session border controllers. Know the power, cooling, cabling and space requirements of these VoIP components then budget for their arrival, installation and operation in advance so any necessary facility augments happen before the component is needed in operation.

Scaling the VoIP Call Control Architecture

Figure 2 shows a typical PacketCable 1.X VoIP deployment.

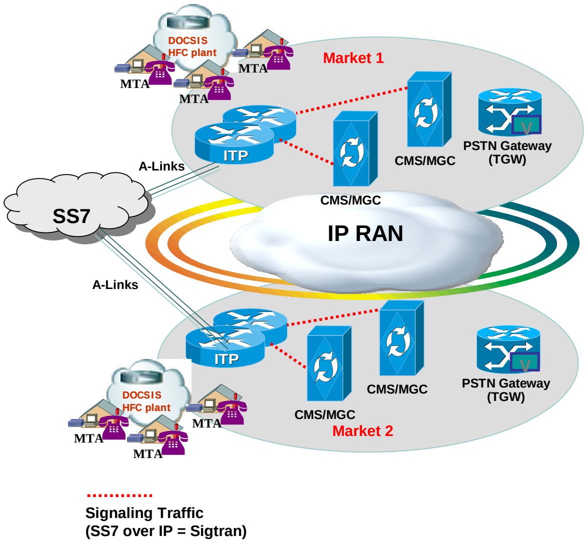
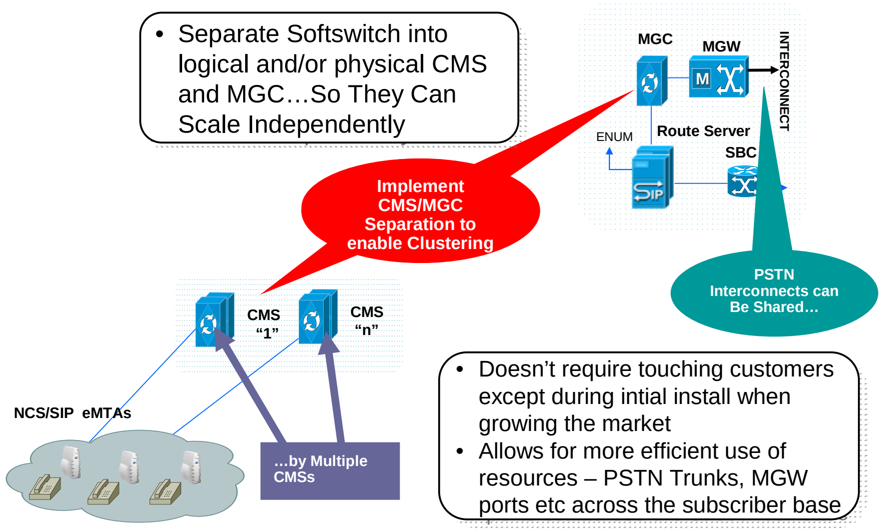


Figure 2 - Common PacketCable 1.X VoIP network

Commonly, PacketCable 1.X implementations have unified software components that contain subscriber side “line” functionality and network side “trunk” functionality. With this software design, signaling components scale in tandem. If demands on any signaling component grows at a faster rate, all the signaling components within the software application grow at that faster rate too. If call signaling patterns shift, say more calls route via CMSS/SIP to other service providers instead of via traditional SS7, the TGW will not need to augment but the CMS lines continue to grow as new subscribers are added.

Splitting the soft switch into its logical CMS and MGC parts separates line side expansion due to subscriber growth separate from trunk expansion because of SS7 offnet call rates. The CMS manages call signaling and features for the NCS and SIP endpoints in this separation architecture while the MGC manages offnet call signaling and control of the TDM trunks as well as IP interconnects to other voice service providers. Migration to the CMS MGC separation architecture commonly is a logical separation of CMS and MGC on the same hardware platform. However, CMS and MGC must grow to meet demand so either one can offload onto physically separate hardware and provide the necessary processing power to accommodate the offered subscriber call load. Figure 3 shows the CMS MGC separation architecture.



- Separate Softswitch into logical and/or physical CMS and MGC…So They Can Scale Independently

Implement CMS/MGC Separation to enable Clustering

PSTN Interconnects can Be Shared…

- Doesn't require touching customers except during intial install when growing the market
- Allows for more efficient use of resources – PSTN Trunks, MGW ports etc across the subscriber base

Figure 3 - CMS MGC separation architecture

Expanding the HFC, IP and TDM Infrastructures to Handle Subscriber Growth

HFC/DOCSIS Infrastructure

Now, consider the DOCSIS portion of the delivery network. DOCSIS UGS minislots slots are akin to PSTN trunks. Like trunks, the number of available UGS minislots determines the number of simultaneous VoIP calls. It is the operator's charter to decide how many minislots are available for VoIP traffic by specifying the number of UGS flows admitted on the carrier. This decision influences the experience of customers using high speed data services on the same upstream link. Available high speed data bandwidth shrinks and its delay increases as more slots carry UGS VoIP flows. As such, the operator can use "what if" analysis and usage trials to tune an acceptable level of data response time while maximizing the simultaneous voice calls per upstream/downstream.

Make sure to calculate the downstream utilization when linking additional upstreams to a downstream. It is helpful to calculate a constant average throughput rate per subscriber, 48Kbps for example, based on observed busy hour utilizations and actual subscribers per upstream/downstream. Compare this actual average to the forecast average throughput per subscriber and compare actual subscribers to forecast subscribers per upstream/downstream. If the product of the actual average throughput and forecast subscribers per upstream/downstream exceeds the bandwidth of the upstream/downstream, consider a capacity augment.

To minimize wasted space inside minislots, specify the minislot size at 8 or 16 bytes to promote efficient insertion of VoIP packets into the DOCSIS layer. Note how changing modulation technique from QPSK to 16-QAM nearly doubles the number of simultaneous VoIP calls per upstream as well as the number of supported subscribers per upstream. Payload Header Suppression (PHS) provides additional call capacity per upstream but not nearly the magnitude of a modulation change from QPSK to QAM. Consider enhancing plant quality to support DOCSIS 2.0 64-QAM upstreams and triple the size of the upstream "pipe." For capacity planning, tune DOCSIS parameters and enhance plant RF characteristics to ensure call quality while minimizing the stress on network elements.

IP Transport Infrastructure

Every IP router and switch contains key monitoring points -- CPU utilization, interface utilization, packet throughput, etc. Extract data on these monitoring points with automation like periodically run scripts and SNMP polls. Generate statistics and trend lines using the data to ascertain current IP transport baseline performance and its vector (change rate and direction.) Use the performance vectors to determine when established engineered limits will be reached. Feed "what if" analysis with proposed changes to determine feasibility. Refine the proposed changes based on the results of the analysis. Implement the capacity augment for the IP transport.

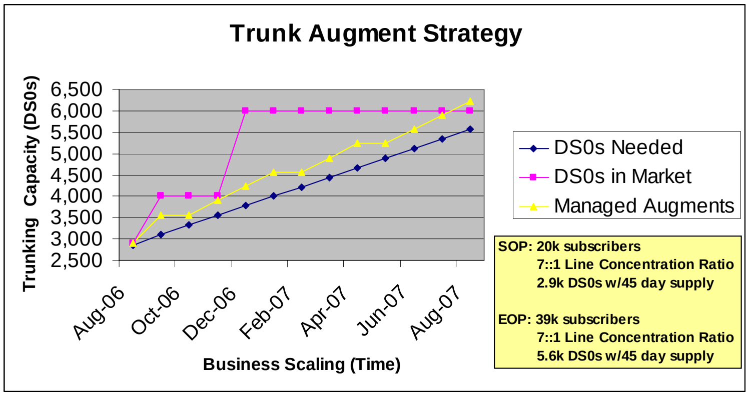
Call statistics provider directions to cost savings opportunities. For instance, as call volume grows between MSO service areas, identify alternate voice traffic transports, like an IP interconnect, between the areas. Next, engineer an IP solution that avoids TDM circuit, access and interconnect charges as well as any other tariffs.

TDM Interconnect Infrastructure

Mine call statistics to understand peak demands and growth trends. Use the findings to augment engineered limits and refine forecasts. For example, routine data mining of trunk group utilization uncovers incremental growing trunk utilization rates. Combine these rates with per unit trunk costs to decide augment size and frequency for each TDM trunk group. The utilization rate sets the growth rate of the engineered limit and associated warning lines that trigger capacity augments.

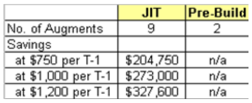
Instead of a trigger managed approach, an operator may choose to manage trunk capacity augments in an ad hoc, "best guess" fashion based on an intuitive "feel" for the growth trend. Frequently, the ad hoc approach works but the operator purchases trunk capacity in an unnecessarily large quantity and thinks "well… we will not have to worry about that problem for a while." Unfortunately, the expense of the bulk trunk purchase often surpasses incremental revenue growth rates from new and existing subscribers so cash flow and profits fall until VoIP service revenue increases to offset the cost spike. A comparison of the two trunk capacity augment methods is displayed in Figure 4, below.

Trigger-Managed (JIT) vs. Intuitive "Pre-Build" Trunk Capacity Augments



SOP: 20k subscribers
7::1 Line Concentration Ratio
2.9k DS0s w/45 day supply

EOP: 39k subscribers
7::1 Line Concentration Ratio
5.6k DS0s w/45 day supply



	JIT	Pre-Build
No. of Augments	9	2
Savings		
at $750 per T-1	$204,750	n/a
at $1,000 per T-1	$273,000	n/a
at $1,200 per T-1	$327,600	n/a

Market scaling 195%

Managed Augmentation assumes:

1. Trigger at 94% BH utilization
2. 45 day supply of trunk capacity increments

Figure 4 - Comparison of trunk augmentation approaches

Notice the ad hoc, "best guess" and the trigger-managed trunk augmentation approaches ensure adequate trunk supplies. However, managed trunk capacity additions based on real capacity triggers offer dollar cost savings compared to bulk trunk capacity augments performed in an ad hoc, "best guess" fashion.

Summary

This paper discussed a stepwise approach to VoIP capacity planning driven by subscriber growth. To begin the approach, understand the time sensitive capacity planning and augmentation process for the consolidated VoIP network as well as is individual components, links and requisite back office systems. Identify sources of capacity planning data within the VoIP network. Next, assess the impact of subscriber count and call volume growth on network elements, bandwidth and, ultimately, call quality / user experience then calculate metrics, interpret results, make augmentation decisions and, finally, take timely action.

About the Author

Bil Dry

Technical Leader, Cisco Systems

Bil is a Technical Leader in Cisco's Advanced Services consulting group with more than 10 years of experience in network and software engineering at Cisco. Recently, he led a team that designed and deployed a nationwide Voice over IP (VoIP) network for a Tier 1 cable MSO.

Prior to designing and deploying VoIP networks, Bil managed one of Cisco's Advanced Services teams that offered design consulting, escalation support and best practice recommendations to service provider packet telephony customers.

As a consulting engineer at Cisco, Bil leveraged his work with ATM, IP MPLS & WAN to co-author *Cisco Multiservice Switching Networks* and published an article on WAN troubleshooting for Cisco's Packet Magazine. He holds two CCIE certifications and graduated Magna Cum Laude with a bachelor's of science degree in Electric Engineering from North Carolina State University.

Reach Bil by sending email to him at bil@cisco.com