Considerations in VoIP Network Scalability for High Growth

Author and Presenter: Parviz Rashidi, PhD.
VoIP and IMS Network Architect
Network Business Solutions, Nortel
3500, Caring Avenue
Ottawa, Ontario, K2H 8E9
613-765-3659
rashidi@nortel.com

SCTE Cable Tech Expo
June, 2007
Orlando

Abstract

Subscription to VoIP services is growing at a very fast rate fueled by today's competitive pricing, enhanced features brought by converged networks and the efficiencies realized by use of packet technologies. However, the high rate of growth has created its own challenges. Network architectures that can scale effectively to serve the future needs without the necessity to implement expensive network reconfiguration and re-homing of subscribers need to be implemented before issues of network growth slows down opportunities for revenue growth.

This paper looks at the characteristics of the PacketCable 1.x VoIP networks and focuses on the control layers to analyze how to grow a network in different geographic regions. It looks at various models for adding new call processing capacity to the network. The discussions address if it is better to divide the network into different regions (i.e. re-home subscribers) when adding a new control node (CMS/MGC), or if it is better to allow the new node to overlap its control span with an existing switch. It is shown that a combination of these basic options could be a more efficient solution. Uses of spreadsheet models that help to compare the economic impact of the decisions are recommended.

It is expected that as the network grows, routing between switches will become more complex and can become a problem in managing network growth. Traditional tandem switches as back-to-back user agents can provide a solution. However technologies based on SIP redirect agent or SIP routing agent as a "control agent" combined with data lookup from a database such as ENUM as the "control instruction" for routing can offer a highly scalable solution to complex routing problems. It is noted that PacketCable 2.0 IMS architecture also uses data from database sources such as HSS and ENUM to provide "control instructions" on per user basis to offer a solution with limitless network scalability and provide the linkage to integrate segments of existing PacketCable 1.x networks.

Speaker Bio

Parviz Rashidi is a network solution architect at Nortel's Network Business Solutions organization. In the past five years he has been working on VoIP and IMS network solutions for the MSO industry. He has over 26 years of experience in the telecommunications industry, working on solutions and services for different operating companies. He has held technical specialist and management positions in the areas of VoIP and IMS solution architecture design and analysis, network product integration and release management, SS7 network solution validation, network synchronization and network standards.

Dr. Rashidi holds a PhD degree in Telecommunications and a BSC degree in Electrical Engineering from Imperial College, University of London. He has also held university teaching positions in computer science and electrical engineering departments at the University of Edinburgh, Scotland, Shiraz University, Iran and University of Adelaide, South Australia.

Acknowledgements

The author wishes to express many thanks to his colleagues at Nortel, especially Peter Krautle, Brain Lindsay, Russ Coffin, Elaine Smiles and Rick Walczak for their support and contributions in preparing this paper.

Introduction

Scalability is a desirable property of a system, network, or process that indicates its ability to either handle growing amount of work in a graceful manner, or to be able to readily expand in size. Generally, when considering scalability, it is the scalability of a system that one thinks about. Vendors design their systems to optimize granularity of growth against factors such as costs and complexity. In systems, scalability parameters such as incremental hardware to increase processing capacity, incremental ability to expand interfaces to other systems and incremental ports for numbers of users on the system are typical parameters that specify scalability of the system. However, there are other dimensions to consider when it comes to scalability of networks.

Scalability in VoIP networks may be looked at from several dimensions. In one dimension, network scalability may be considered as how well it can scale to support growing number of end users and how they use the network. It is desirable that as number of end users increase, the equipment that defines the nodes of the network grow in proportional steps. Objective should be to manage costs and avoid installing equipment that will sit idle for a long time. Similar arguments can also be applied to scalability of a network to support traffic variations. How well a network can scale to handle traffic increases or changes in traffic distribution over time are some of the scalability issues that network designers need to consider.

The other dimension of network scalability is its ability to interface to other networks. Of course other networks will come in all shapes and sizes and will involve use of different technologies. When designing a network it is desirable to consider how it will scale to support interfaces to other networks that may use the technologies of the past, the present and the future.

By definition, networks have geographic distribution. Geographic scalability is an important dimension to also consider. As the network grows, it is necessary to define how that growth will be managed across different geographic regions. Some regions may have dense populations while others may be thin in population or may be 100’s of miles away from other parts of the network. A geographic scalability consideration would take into account how the network can grow in the geographic environment that it serves. Of course the objective in designing the geographic layout is to keep the costs low and yet allow the size to increase in increments as the demands on the network in various regions increases.

Figure 1 shows a high level view of PacketCable (PC) 1.x VoIP network architecture and how some of the scalability variables discussed above relate to the network architecture and its growth. The objectives of any network design are to balance the costs of managing these variables versus the costs of operating the network in relation to the revenue that it will generate.

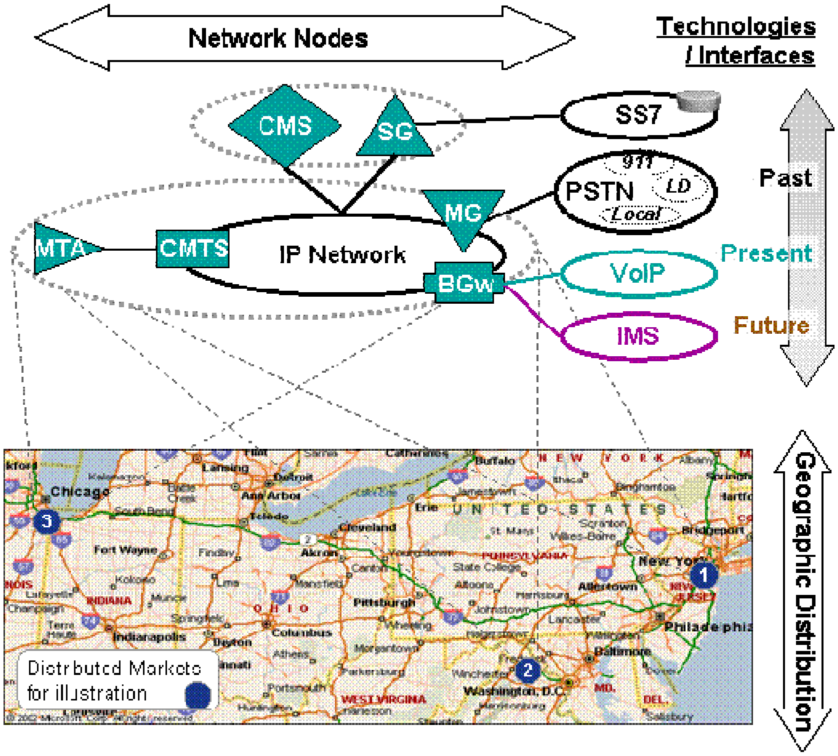

Figure 1: Network growth variables

This paper looks at the key PC1.x VoIP network functional elements and identifies the problems of scaling the architecture as number of subscribers increase. It discusses solutions to the problem of how to manage a growth of a VoIP network in relation to geographic areas that it serves. References to the network in these discussions relate to the PacketCable VoIP network definitions although similar concept also can be applied to other types of networks. Objectives are to study VoIP network scalability from the geographic distribution point of view. The focus of the discussions is to separate the choices that network designers have in order to simplify the decision models. The methods proposed can then be used by network planners to develop simple spreadsheet geographic models of the network with variables to help understand the impact of their decisions and thus find the most cost effective choices for the layout design of the network. Finally technologies such as SIP redirect user agent or SIP routing agents with data from ENUM database or use of PacketCable 2.0 technologies based on IP Multimedia Subsystems (IMS) are shown to provide solutions to network scalability problems.

Hierarchical Network Architecture

Architecture of the network plays a key role in how its scalability is managed. In one extreme, a network with flat distribution of functions in which every node has the same functional capability as every other node forms an architecture with maximum scalability. In such a flat architecture the nodes can be added as required when a new subscriber or a group of subscribers are added. Problem with this architecture when used for VoIP applications is that the equipment at each node or at the subscriber location will need to be complex to support all of the VoIP networking features. The costs per subscriber for such architecture can grow quickly as the network grows.

PC1.x has defined the functional architecture of the VoIP network in a hierarchy. A way to view this is illustrated by the functional layers in Figure 2. In this example each layer illustrates a functional plane in the same geographic area and can represent how the nodes in that plane are scaled as well as how they are distributed in the network serving areas.

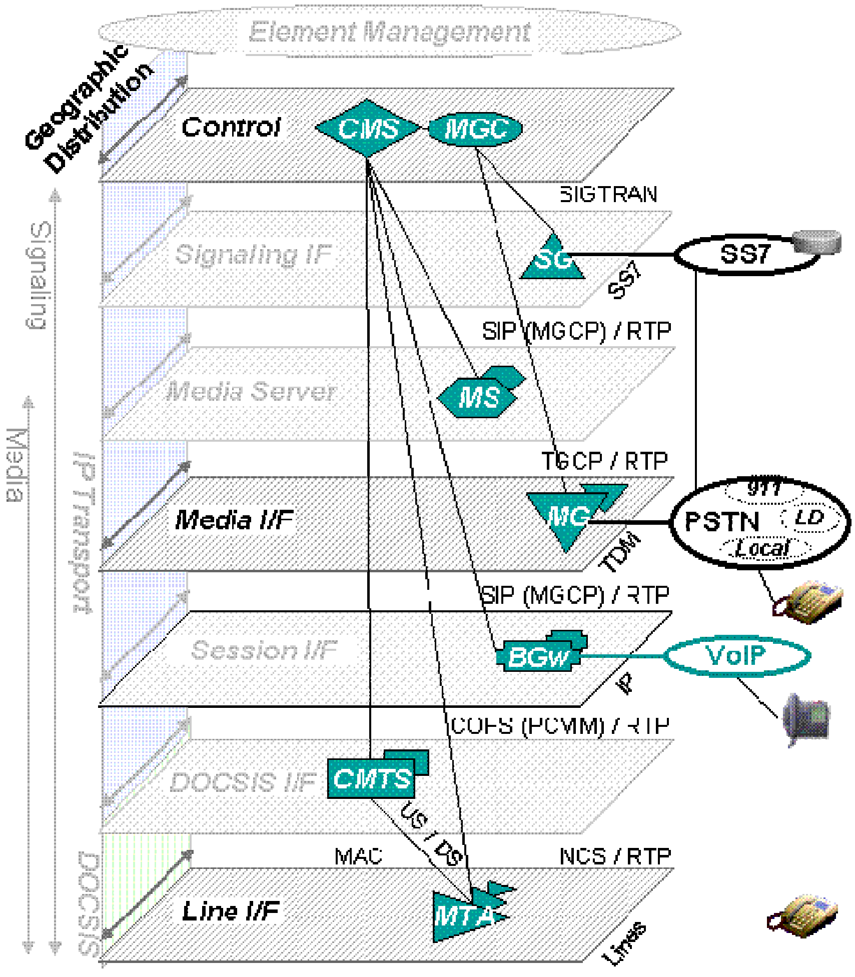

Figure 2: PacketCable 1.x functional layers

In Figure 2, the highest level in the call control hierarchy is the Call Management Server (CMS) and the Media Gateway Controller (MGC) functions that provide the call processing functions for lines and for Time Divisions Multiplexed (TDM) inter-machine trunks respectively. At the lowest level in the hierarchy, the Media Terminal Adaptors (MTA) provides the gateway functions to interface to the individual subscriber lines. For discussions in this paper, the focus is on elements that get directly involved in call processing and thus the management layer is not considered.

In between there are functions such as Media Gateways (MG) and Signaling Gateway (SG) that provide the interfaces to the TDM networks. The Cable Modem Termination Systems (CMTS) provide the interface to the DOCSIS layer of Hybrid Fiber Coax (HFC) access network. As direct packet interface with other VoIP networks become popular, then more and more nodes, represented as Border Gateways (BGw) in Figure 2, will need to be added to control the direct packet interfaces to other VoIP networks.

Figure 2 also shows the IP transport layer plus the DOCSIS access layer, illustrating how traffic is passed between the nodes on the same plane as well as between the nodes in different planes. The illustration also shows how signaling and media traffic spans between the different planes. Understanding of how much traffic needs to flow within and in between the planes is important to develop a geographic layout that optimizes the use of transport facilities.

The media traffic only spans across the lower planes as shown. However, the signaling spans across all planes following the path shown by the lines connecting the nodes. To minimize the unnecessary backhauling of traffic across the transport infrastructure the nodes in the planes that span media traffic need to be close to where the population densities are. However the nodes that only handle signaling can be centralized, as signaling uses much less bandwidth.

The exception to above rule is the Media Server (MS) node which usually would carry an order of magnitude less traffic per subscriber line than the MG or BGw and thus it may make sense in most implementations to also locate the MS functions such as voicemail and announcement servers at a central location to have more efficient scaling of the equipment for these servers and save on equipment costs.

Scaling Dependency

In any hierarchical architecture such as PC1.x as depicted in Figure 2, the nodes for different functional layers will scale differently. The scaling characteristic of the network is then determined by how the combined layers scale together in relation to their geographic serving area and relative to each other. The scaling dependency between the nodes is defined by the control dependency, which in turn is set by the function of the node, and the protocols used. The geographic serving area dependency of each node is then defined by how the network layout for dependent nodes is designed.

The control layer being the highest in the hierarchy sets the boundary for geographic coverage. If nodes in other layers such as the session interface layer are not controlled as slaves by the control layer, then the geographic coverage of those nodes will be independent and can cross the service boundaries for the control layer nodes. On the other hand the geographic coverage of a node such as MG and its interfaces that map to the TDM network will be dependent on the geographic coverage of the MGC node in the control layer.

The discussions presented in this paper assume that the CMS and MGC control functions are implemented as combined in one system (the CMS/MGC or softswitch). This simplifies the analysis but does not restrict the methods and the conclusions to this assumption. It is important to note that with both control functions together in one system, the engineering of the system can be defined so it can function as 100% CMS (for lines only) or 100% MGC (for inter-machine trunks only) or any combination in between thus allowing flexibility to engineer the control layer.

Studying the protocols between the layers and using the dependency relationship discussed above, the key functions that depend on each other in a geographic layout situation are the following:

- CMS/MGC with MGs and MTAs
- CMTS with MTAs

The geographic scalability of CMS/MGC in relation to MGs and MTAs is discussed in this paper in greater details. However, the scaling and the geographic coverage of CMTS in relation to MTAs relates to how BW for the Upstream (US) and downstream (DS) channels of the DOCSIS HFC network are engineered in the CMTS. This requires a separate discussion in its own merits and is out of scope for the discussions in this paper.

The session interface layer involving the Border Gateway (BGw) node of Figure 2 that allows direct packet handoff with other VoIP networks also needs a separate discussion in its own merits. For direct VoIP packet handoff with other networks, there are several methods relating to how security can be managed with each having its own scaling implications. To understand these methods the reader is referred to references (1) and (2) which give background to the network interconnection solutions and the network border security solutions respectively. A brief outline of the different methods is discussed below to allow a comparison of scaling implications.

One solution for direct packet handoff is to rely on security at the IP layer using routers and firewalls to manage the traffic across the border interfaces. In such implementations, the scaling of equipment at the border will be independent of the call processing equipment and needs to only follow the normal IP traffic engineering considerations.

When considering Session Border Control (SBC) function in the session interface layer of Figure 2, there are two architectures for the SBC to consider. These architectures are different in how the VoIP media and signaling are separated across the network borders (as discussed in Reference 2). If the SBC is a standalone SBC in which both signaling and media are controlled in the same node, then the scaling will be only related to the traffic across the border at that interface. This requires similar considerations as for the IP layer solutions, and will not depend on the scaling of other layers.

An alternative implementation of the SBC is to have the control of signaling traffic at the central location with the CMS/MGC and distribute the control of media traffic to different regions in the network near where media traffic flow takes place. This provides advantages in having a hierarchical architecture for control of border traffic. In such implementations the considerations will be similar to the relationship between the MGC as the controller and the MG as the dependent node. Since discussions on the SBC solution has other implications related to border security and requires other topics to be introduced, its detailed discussion is outside the scope of this paper. Here it is only sufficient to note that depending on SBC's functional architecture, similar arguments also can apply to design the geographic layout of SBC node. It is also important to note that regardless to which solution is used for the SBC, it is always necessary that the SBC function for media control is close to where the population densities are to minimize backhauling of the media traffic on the transport layer.

Finally the SG in the signaling layer and its SS7 links are engineered to match the traffic flow to the SS7 network. For this system the geographic coverage of the SG is therefore dependent on the coverage of the MGC (the controlling node). If the SG can support a number of different MGCs, then the relationship simply relates to the summation of those nodes.

Initial Network Growth

It is desirable that growth of the network initially starts with limited investment in equipment but allows the capacity of the network to grow in small increments as the number of subscribers and geographic coverage increases. The hierarchical architecture allows the problem of scalability to be divided and solved at each functional layer.

The ultimate limit to the initial network growth is the capacity of the CMS/MGC (for simplicity referenced as CMS). The decisions made as how to expand the reach of the control layer in different geographic regions can impact the decision on when and where to add the second CMS node. Following are some considerations in planning for growth of the CMS.

The first is how the network geographic architecture should grow to allow scalability of CMS services in different regions. A related question is at which locations the network nodes that interface with other networks should be planned?

The answer to these questions relates to how important it is to minimize transport bandwidth requirements. A simple answer is to position the interface nodes at locations where there are significant amount of traffic that terminate or originates between the two networks. Since local traffic is usually the largest of all traffic, then it makes sense to have the interfaces at high density subscription regions. The method to help decide where to locate the interface nodes may include calculating traffic backhaul distances and comparing different decision models to minimize backhauling of traffic. Note that although traffic over IP network can be carried efficiently, it still uses bandwidth. So the objectives should be to reduce the volume of backhaul traffic while controlling the costs of having too many gateways distributed around the network.

As network grows and new equipment are added, the other consideration is if it is better to increase the number of gateways at existing interface nodes or negotiate new interface agreements at other geographic locations. The question then becomes how to manage the new areas on the original CMS if it is known that in a future time the load on the CMS will reach its capacity limitations. The timing to add a new CMS node will need the consideration of some of the factors discussed in the next section.

Figure 3 shows an example of the architecture as it is initially configured while having nodes in all of the initially chosen regions. To illustrate the geography of the architecture in Figure 3, the reference to the areas marked as markets 1, 2 and 3 refers to the geographic market examples of Figure 1.

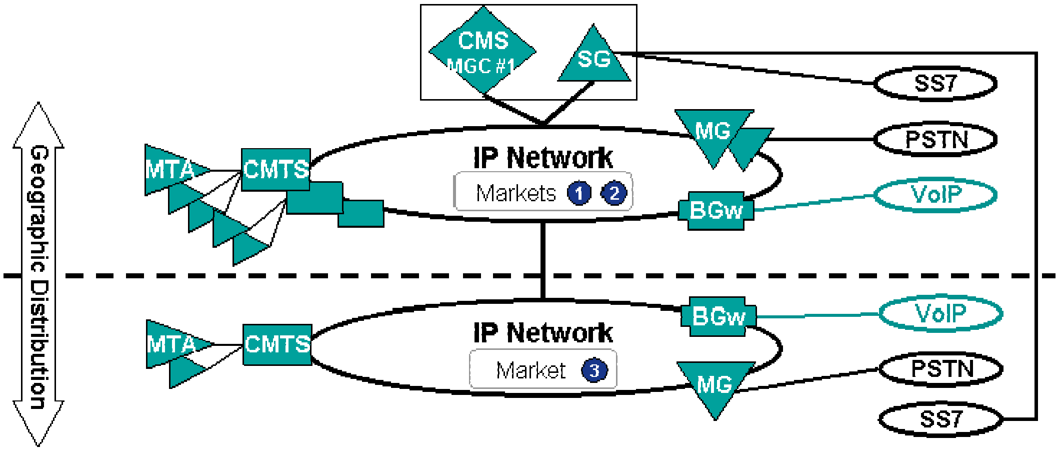

Figure 3: Initial network

Options to Add a New CMS / MGC

As the VoIP network expands, one problem is what to do when the initial CMS reaches its capacity. When the capacity limit of the CMS, as the highest point in the control hierarchy, is reached the question is if the new subscribers in the same geographic area that are served by the first CMS should go on to a second CMS to create an overlap network architecture or if it is necessary to separate the network into two regions that are controlled by different CMS nodes and re-home the subscribers from one CMS to the other. Both models may be valid decisions depending on the situation.

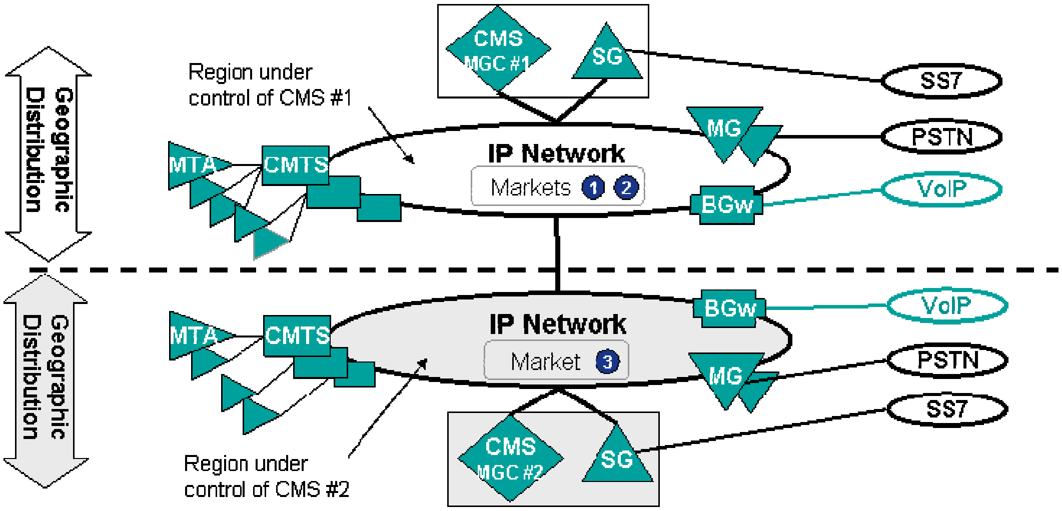

Figure 4: Re-homing model

As illustrated in Figure 4, let's first consider the re-homing model in which the network gets divided into two regions. For this model, at a decision point in time, a second CMS is introduced and the network is split in two sub-networks based on their geographic distribution of subscribers. The MTAs that support subscriber lines in some regions and gateways that interface to other networks in those regions (such as MG / SG) will need to be re-homed to the second CMS. As part of the re-homing plan, the links to SS7 network and the physical trunking facilities to other networks will need to all get re-homed while keeping service outage to a minimum. At the same time the routing in switches in other networks also need to change to match the changes discussed. This requires close coordination with operators of all neighboring networks.

As can be noted, re-homing will require planning and expenses to implement. However for some situations the costs of re-homing may pay off by having separate operations in different regions to manage the network and save in operations costs.

Re-homing makes sense if regions are geographically distinct and the growth of the network in one of the region is controlled until a decision time when re-home takes place. By planning from the first day that some of the regions may need to be re-homed in the future, the costs of re-homing can be controlled. A decision model as shown in figure 5 may be used. A decision threshold needs to be defined which may be based on capacity of the CMS node and factors such as costs of re-homing, or having a safety period before CMS reaches its capacity. This helps the decision process on when to re-home and how much to re-home to add sufficient growth time in the life span of the new architecture.

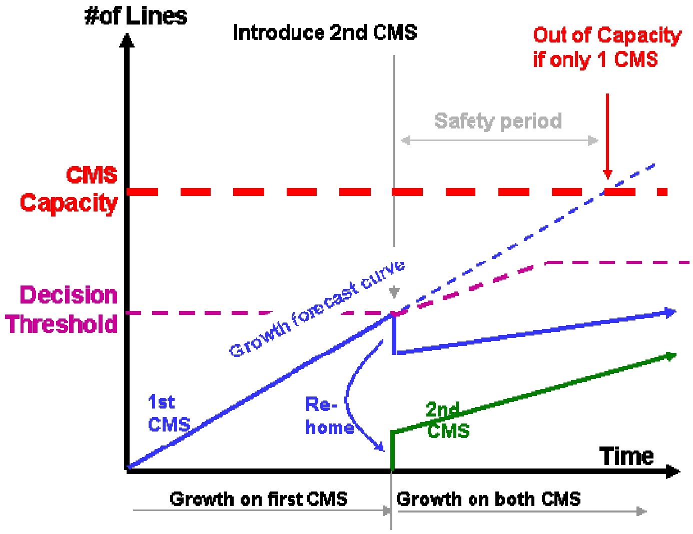

Figure 5: The Decision model

In the example of Figures 4, the regions referred to as 1 and 2 are planned for the first CMS while Region 3 is planned for a second CMS. However, during the initial growth phase, all subscribers will be on the initial CMS. When the decision threshold is reached, then a controlled group of subscribers in region 3 are re-homed to the second CMS freeing up capacity on the first CMS. Figure 5 shows that after the re-homing, the growth curve will have a longer time to reach the capacity limits of the CMS as the rate of growth will be shared between both CMS nodes. With that consideration, the decision threshold can be raised to a higher level. Also the decision model can be customized to take into account factors such as future CMS capacity increases, or other advances in the CMS technology from the vendor.

Alternative to re-homing is to design the network to overlap control between two CMS nodes. This model may especially be economical in dense populated regions with high growth forecasts. Problem with overlapping architectures is to have to manage more complex procedures to support the split of subscribers between the two switches. In a region with high density of subscription and large growth, overlap operation may be a necessary part of managing the network.

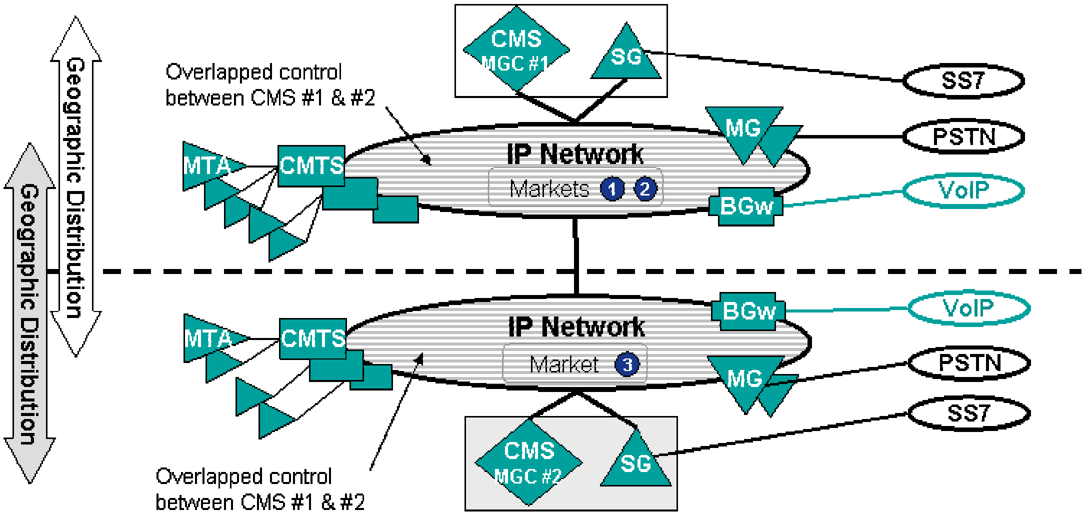

Figure 6: Overlap growth model

Figure 6 illustrates an example of the overlap network model. In this example, the first CMS/MGC is allowed to grow to its capacity limit (or a decision threshold based on a safety buffer period). When that threshold is reached, the first CMS/MGC is capped and the second CMS/MGC is allowed to take over the remaining growth. This process can carry on to the third CMS/MGC and so on. Of course, in the detailed planning the decision threshold need to be set so some spare capacity is reserved on the switch to allow room for traffic between the overlapping switches to grow as well as allow a safety period before the original switch reaches its capacity limits.

Composite Growth Model

One of contribution of this paper is to show that the architecture for network growth can also be managed through a combination of models discussed in the previous section. Figure 7 illustrates an example of this.

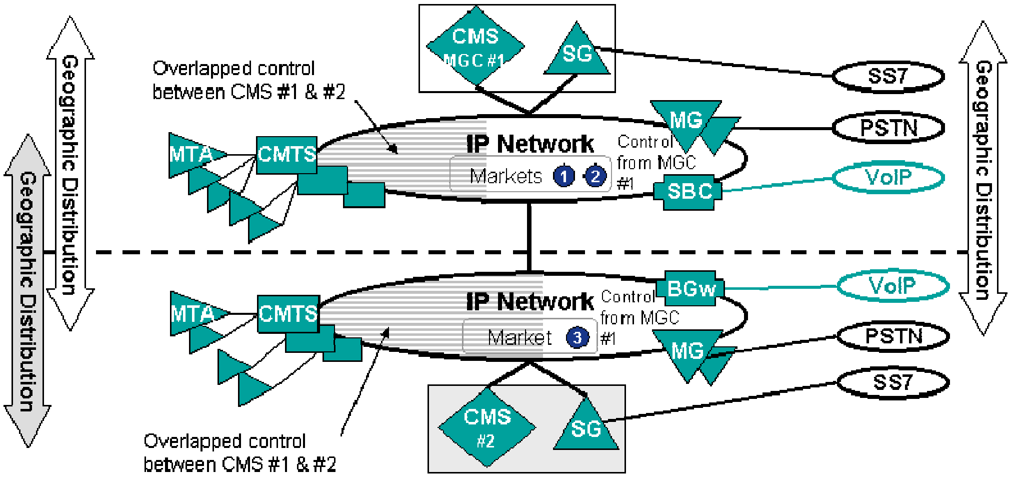

Figure 7: Example of a composite growth model

In the example of Figure 7, the initial growth of the network is allowed to carry on up to a threshold after which the growth of lines on the first CMS/MGC is capped, but growth of trunks is allowed to continue. All traffic to the PSTN from the second CMS is sent to the first CMS/MGC, which is allowed to grow its trunking (or its MGC) capacity to serve both switches. In this model there will be no re-homing necessary because the trunking facilities to PSTN are kept only on the first CMS/MGC node and lines will be operated in overlap mode. Sharing the interface facilities to other networks also increases the efficient use of these facilities. It is expected that as direct packet interconnect becomes more popular, the growth of traffic to the PSTN network will slow down and keeping all of the TDM trunking on the initial switch to manage the interface facilities may be the right choice. However as the result of passing traffic between the switches, additional packet interface resources will need to be engineered on each switch.

Using above examples and extending the arguments, other growth models can be defined. Following is a list of promising possibilities:

- Grow lines and trunks in region and re-home as necessary (Figure 4)
- Overlap lines and trunks by adding a new CMS/MGC when first CMS/MGC reaches capacity (Figure 5)
- Overlap lines but re-home trunks to create regional trunk distribution on the switch serving each region
- Overlap lines, but grow trunks only on one CMS/MGC (Figure 7)

- Grow lines in regions and re-home as necessary, but grow trunks only on one CMS/MGC
- Assign a switch to handle all interfaces to other networks (i.e. use it as 100% MGC) while expanding with overlap lines on other CMS nodes
- Others - Other options are also possible, but there effectiveness is not as significant as above examples.

Planning for growth of SBC nodes in the network also can follow similar arguments but needs to allow for the differences in the functional and control architecture of the SBC in relation to the control plane as discussed earlier in this paper. Solutions for SBC model should be defined on the basis that SBCs, like MGs need to be located at suitable geographical locations with high density of subscribers to minimize backhauling of traffic.

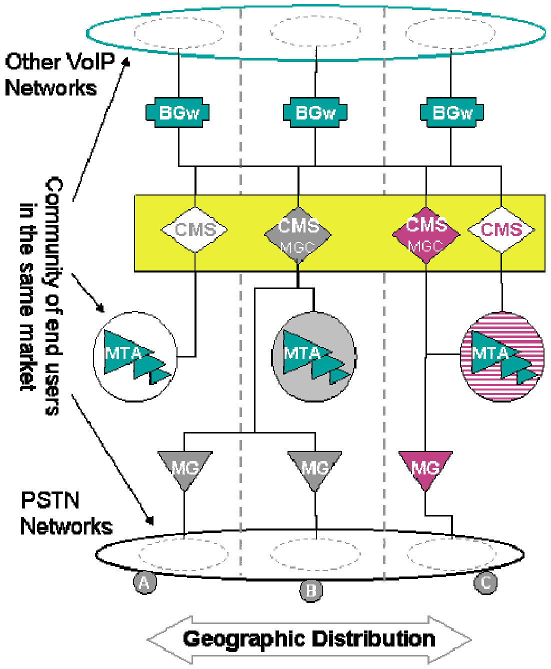

Figure 8: Illustration of growth models and network interconnect layers

Figure 8 shows a 3 region example of how a network architecture may evolve as the results of different growth decisions. The example illustrated shows different layers of the network, different interfaces to other networks and different models for adding new CMS/MGC nodes for each of the 3 regions in the example. Note that in the example, two of the nodes are shown as 100% CMS function, while the other two are shown as combined CMS/MGC to illustrate how a combined CMS/MGC node can provide flexibility in engineering decisions.

Routing Scalability

As the number of CMS nodes increase, a function that becomes increasingly under heavy usage is the routing between the CMSs and routing to other networks. With more CMS nodes, the routing function will have to filter through more possibilities. Today, different vendor switches can scale their routing capabilities to different degrees, but eventually, one expects that as the network becomes larger, the routing capacity of any switch will reach its limits and therefore a solution to the routing problem will be needed.

The traditional solution to connect a number of switches is the tandem switch (or Class4 switch if additional features are needed). A tandem switch with SIP trunking interface acts as back-to-back user agent allowing a mesh connection between several switches to simplify into a star architecture. This enables solving the routing problem at a single node instead of at many nodes. The tandem switch then will be required to build the algorithms and work with the Local Number Portability (LNP) systems to scale to the increasing routing complexity as the network grows and number of nodes and subscribers increase.

Solutions involving processing of SIP messages either as SIP Redirect Proxy or SIP Routing Proxy (SRP) also can offer the same results as the tandem switch but without the need to manage each call for its duration as back-to-back user agent. This offers a simpler system architecture, which can scale to virtually indefinite limits. However there are some implications to each solution

The SIP redirect proxy as redirect agent will create increased load on each CMS as a side affect. With more messages being redirected back to the CMS, it will need to do more work and process more messages. However, the SIP routing proxy will process every message as a router at the SIP layer. In this case the system then needs to have more processing capacity of its own.

In an SRP node (whether as redirect agent or routing agent), each hardware system element references a database to get routing information for every message. Because the SRP does not require the overhead of back-to-back user agent, it has simpler and scalable hardware architecture. Now the hardware system is a “control agent” for routing with limitless scalability. The data that the hardware system uses then becomes the “control instructions” for what the hardware should do. Data in turn will also offer limitless scalability.

Although the routing information for the SRP node can be stored in its internal database, a simpler scalable solution that is emerging is the ENUM database as the source for routing information. ENUM as a DNS database can provide addressing for each individual numbers and can scale to indefinite limits. The MSO operators can either use the ENUM data from a data service provide, or can build their own carrier ENUM database as a solution to the growing routing problem.

Taking this concept further, either the CMS switch can do the DNS ENUM query directly to get the routing information, or the ENUM server can be the SIP proxy (as either redirect or routing proxy). Each of above suggested solutions could provide a simple answer to the routing scalability problem in a growing network.

With routing between switches simplified, one of the major disadvantages of the overlap architecture is now removed. Of course other issues such as managing regional diversity when using overlap architecture still remains and needs to be solved in the planning process. However, with routing process simplified to a simple ENUM lookup, the overlap solution should be a favorable option for the MSOs to consider for their high growth regions.

Network Technology Evolution

The previous section introduced the concept of using SIP redirect or SIP routing technologies with lookup of routing data from a database such as ENUM to solve the routing scalability problem. This allowed the complexities of routing to be defined in the data structure and control of routing to be managed by database lookup. The principle is the use of data as “control instructions” and the hardware systems as “control agent” to provide limitless scalability for the SRP node.

PC2 IMS technology is also based on this principle, allowing the data in a database such as Home Subscriber Server (HSS) to define the “control instruction” environment for the hardware system to operate as the “control agent” for every individual end user. Also data in the ENUM database define the “control instruction” environment for routing to every destination. The hardware system as the “control agent” is then scaled to offer the required services for the user population and the data in HSS and ENUM is scaled to offer the “control instructions” for each individual end user.

In PC2 IMS system architecture, the hardware for each function or group of functions is added independent of other functions as the network subscribers and traffic grows. With aTCA technology for IMS platform, the hardware is a standard blade in the aTCA chassis that offer an efficient system scalability and redundancy architecture. As the network subscriber population grows, more aTCA blades are added to support more subscriber population and more data added to define the customized controls through the system architecture.

Through interface functions such as BGCF, MGCF and IBCF, the IMS network allows linkage between the existing PC1.x segments, the PSTN networks, other VoIP networks and other IMS networks. With the scalable routing and interface architecture the PC2 IMS can also provide a solution to the routing scalability problem in the existing PC1.x segments. A high level view of the evolving network architectures starting with the PC1.x segments and growing to integrate with PC2 technology segments is illustrated in Figure 9.

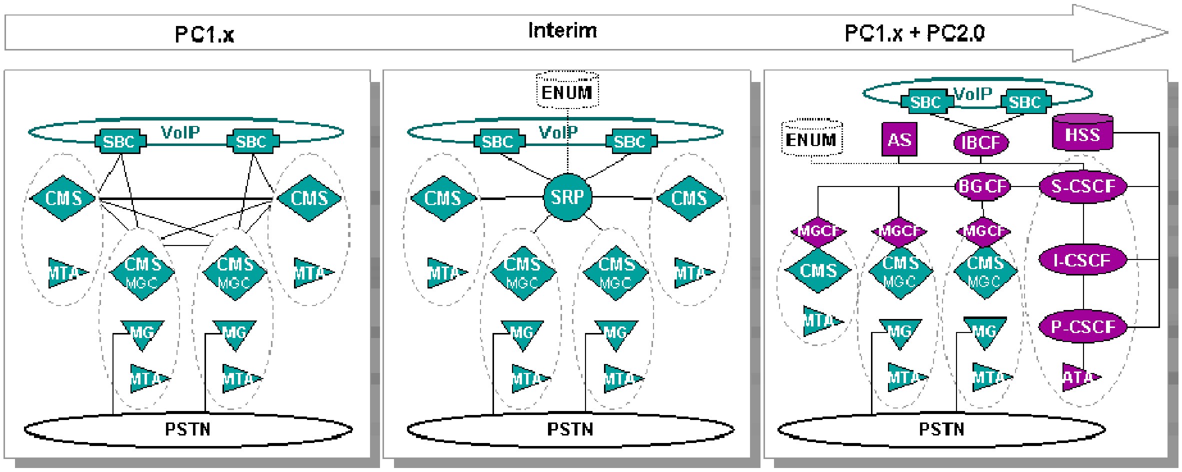

Figure 9: Evolution of network architecture

The examples in Figure 9 show the functional architecture of the network separated in different control segments and illustrate a summary of the solutions discussed above. Different segments involving 100% CMS control or mix of CMS and MGC control are also shown. Each segment can represent different geographic regions or two or more of the segments combined can represent the overlap scenario in one geographic region. So in an area, whether the end subscribers are serviced by MTAs that are homed off PC1.x CMS node or ATAs serviced by PC2 P-CSCF nodes, the control of the region can be in overlap model or separated into regions. Of course, by definition, the network also has to overlap with networks of other service providers in the same serving area.

Figure 9 also illustrates how PC2 IMS technology can provide the overall linkage between each segments of PC1.x. For operators who may not choose to adopt PC2 technology for a while, the Interim solution with ENUM based technology together with the models discussed in this paper can be used to manage growth of the network. The network can be planned in regions to manage administrative differences and allow the network in each region to grow in overlap architecture as subscriber population grows.

Conclusions

This paper has introduced a systematic approach to address the planning for growth of PC1.x VoIP networks in a geographic area. Although the concepts introduced are in practice to various degrees in existing networks, the discussions have shown that there are several other possibilities to manage growth of a network in relation to its layout architecture.

The concept of evolving the network in regional areas and re-homing of lines and trunks to balance the load between CMS/MGC nodes is one of the basic approaches. The other approach is to allow overlap architectures to develop in regions thus eliminating the need to re-home. But overlap models could result in more complex administrative procedures that may not be practical, especially in regions far apart. More importantly there may be inefficiencies when there are overlapping interfaces to other networks from CMS/MGCs in the same area.

A solution proposed suggests solving the growth scalability problem in different geographic regions with a mix of different models. As example, a solution may be to allow given CMS/MGCs to be the only controller(s) for interfacing to other networks for all CMSs in a give region, while allowing all CMSs to overlap control for lines in that region. Decision analysis using the concepts discussed with simple spreadsheet models can provide the tools to help the decision process for optimum design.

As the network grows, a problem that will surface is the increase in routing complexity between the CMS/MGCs. Eventually the ability of the individual CMS/MGC switches to handle the increased routing complexity will be saturated and other solutions will be needed. A tandem switch allows simplifying a mesh of routing possibilities into a star architecture thus focusing the solution for routing at one node instead of many. However tandems switches, as back to back agent will have limitations and if other capabilities of the switch are not required, a better solution would be to use the technologies based on SIP Redirect Proxy or SIP Routing Proxy with the capability to query ENUM database for routing information. This will realize a virtually limitless growth solution for routing.

The architecture for querying an ENUM database can be simplified and yet be fully scalable when the CMS directly interfacing to the ENUM server. This can be done either via SIP interface with the ENUM server functioning as a SIP redirect (or SIP routing) agent for the CMS SIP messages, or having the CMS node make the DNS ENUM query directly to get the routing information.

Finally PC2 IMS based technology is shown to offer limitless scalability for growth in every dimension including automatic and dynamic reassignment of resources. PC2 solves the scalability problem by using the limitless scalability of data in HSS and ENUM databases as the sources for “control instructions”. This allows the hardware systems as the “control agents” to scale independent of each other while the “control instructions” of the data from HSS/ENUM will customize their functions to scale to the individual needs of each end user. With this powerful scaling capability, PC2 also can offer the solutions as the integrating agent to the existing segments of PC1.x network, the past technologies of the PSTN and the evolving networks of the future.

References

1) The packet interconnect opportunity for MSOs
http://www.nortel.com/solutions/cablemso/collateral/nn113860.pdf

2) Extending VoIP across network boundaries
http://www.nortel.com/products/01/succession/cs/softswitch/collateral/nn107880-041404.pdf