



Management of
Large Scale IP Video
Networks

September 2008

INTRODUCTION

Video communications is expanding beyond the conference room and is becoming a standard communication tool and part of the daily workflow for individuals and work groups. This trend is being driven by the integration of video communications into the IP networks of commercial, educational, and government organizations and is having a profound impact on the management and scalability requirements of these video-enabled networks. The increasing demand for visual communications requires networks to support tens of thousands of users and video devices instead of a relatively small number of conference rooms.

Polycom's VC2 vision highlights the importance of management in future visual communication. The key function of large scale video network management is to efficiently deploy and troubleshoot devices remotely throughout their lifespan, thus eliminating the need for on-site technical support and maintenance.

This paper discusses mechanisms for centralized management of large-scale, video-enabled networks that consist of a variety of video endpoints, soft clients, gateways, conferencing servers, video border proxies, and other required devices. It also covers different aspects of management: provisioning, software upgrades, monitoring, licensing, device control and scheduling, and refers to mechanisms described in the Polycom white paper "Scalable Infrastructure for Distributed Video."¹

Note: The soft client is a special video network element that is not literally a "device." To avoid repetition, the paper talks about devices but covers both devices and soft clients in this category. The text will explicitly state it, however, if soft clients behave differently in a particular scenario.

WHAT IS MANAGEMENT?

The term "management" is widely used in the communications industry and means different things to different people, so it is critical here to define the term. In the video communication industry, management includes provisioning, software upgrades, monitoring, licensing, device control, and scheduling. As depicted in Figure 1, management is required for all elements of the video network such as endpoints, gateways, media servers, recording servers, content management servers, and video border proxies.

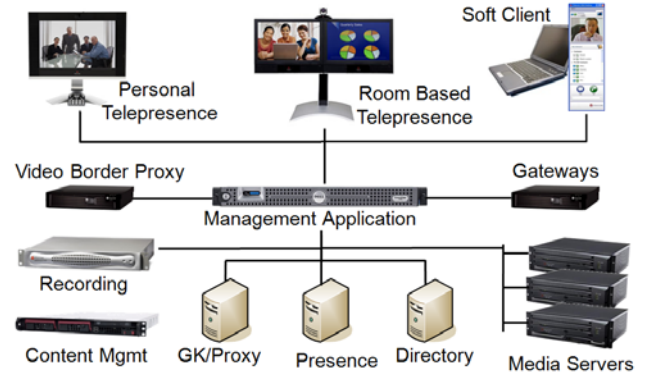


Figure 1: Managing video network elements

The most challenging elements to manage are the video endpoints and soft clients. In a scalable video network there may be tens of thousands of video endpoints and soft clients, and a fully automated way of deploying them is required. In the same network, there may also be a number of media servers, gateways, and so on, that can be configured individually through the administration portal. Since this paper focuses on scalable management, it is primarily focused on the scalable management of video endpoints and soft clients.

Provisioning

"Provisioning" is a technical term for the process of providing users and devices with access to data and technology resources. "Configuration" is a less specific term with multiple meanings, including provisioning, that is, "configuring an end user or a device" is the same as "provisioning an end user or a device". In this paper, we use the more specific term "provisioning" for consistency and to comply with technical terminology. The term "provisioning data" is here used to describe dynamically generated provisioning information. "Configuration data" is related to the term "configuration file" that indicates static configuration information.

User provisioning is easy to understand. When new people join an organization (for example, when new employees join an enterprise or new students join a university), they need access to Web pages, e-mail, the voice communication system, and other organizational systems. The human resources department decides what access is appropriate for the function of the individual in the organization and then the IT department sets up the flags in the Web, e-mail, and voice communication systems, so that the new user can access and begin using these services.

Organizations with more advanced applications and tools have centralized user access information into a single database (master directory), and a new entry in this database automatically triggers the creation of all necessary accounts.

When the user tries to access a particular resource, the authentication service prompts the user to enter a User ID and password. The service then checks the entries against the records in a database and grants or denies access based on the information provided. While the user is part of the organization, profiles and access rights can be modified based on changes of job function. When the person leaves the organization, the IT department must take care of deleting all accounts, so that the person cannot continue to access resources. In some organizations, deleting the user profile in the master database is all it takes to remove the user's access to all tools.

Now let's look at device provisioning. A device connected to a network is typically a computing element and, in our case, a video-enabled network element. Therefore, a device can be a video endpoint, a gateway, a media server, a recording server, or a content management server. Devices are deployed in the organization's network, perform their duties for several years, and are then removed or replaced by newer devices. When the device is first deployed, it has to be provisioned, that is, it has to learn what other network elements are there, and how to function properly in this environment. Figure 2 describes the concept of device and soft client provisioning.

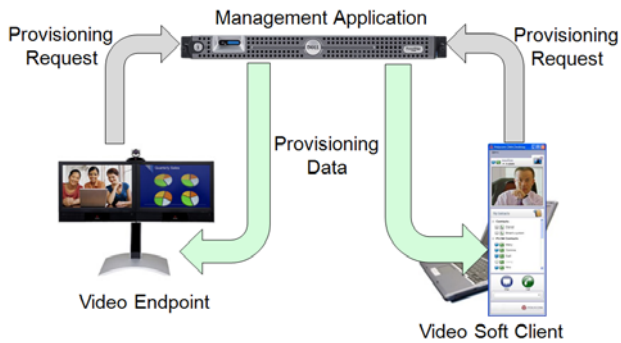


Figure 2: Device and Soft Client Provisioning

If a device malfunctions and has to be taken out of the network, the replacement device may need to have the exact same configuration; therefore, the provisioning service has to store the device's provisioning information, and be able to apply it to the replacement device.

Soft clients are provisioned independently from the computer they run on. While IT departments deploy a variety of tools to provision computers, these tools do not provide the capabilities for configuring every application running on the computer so the soft client application has to be provisioned separately. The trend towards integration of video into the IT network will eventually lead to a converged provisioning mechanism.

Software upgrades

Devices arrive from the factory with the latest available software version at the time of their manufacture but it can conceivably take months before a device is actually connected to a network. In the meantime, software development continues and new software versions with bug fixes and new features are tested and released. As a general rule, all devices should be upgraded to the latest software version upon connecting to the network for the first time. In the past, management tools were limited and could not perform software upgrades of video equipment remotely. Instead, an on-site technician had to plug the device into a server and load the latest software. This process required on-site support and was very time consuming and expensive.

Soft client software distributed on CDs leads to similar issues: the software may not be up-to-date and, immediately after installation, the client will need to contact the management server for a software upgrade.

Figure 3 is a software deployment example. The software for endpoint A is different from the software for endpoint B and different from the soft client software.

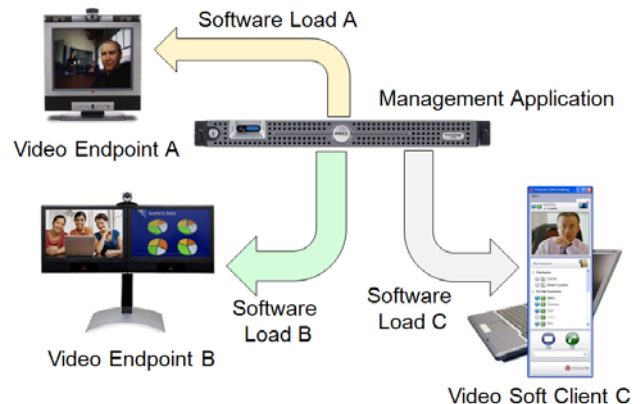


Figure 3: Software Upgrades for Devices and Soft Clients

Modern management tools, such as the Polycom® Converged Management Application™ (CMA™) 5000

(Polycom CMA 5000), allow a device to connect to a management server over the IP network, and retrieve the latest software version for its model.

When the device downloads the new software version, it overwrites the older software version but keeps the provisioning parameters, including addresses, numbers, and names. Often, the new software has new functions and therefore requires additional provisioning parameters that were not available in the older version. If the device does not function properly after a software upgrade, it is a best practice to reload the provisioning information, now formatted for the latest software version, to the device. XML provisioning is technically the superior approach to address the issue and is therefore supported in Polycom HDX endpoints and the Polycom CMA 5000 management application.

Since vendors may release new versions of software several times a year, devices will sooner or later drop so far behind the latest version that this may impact interoperability with other devices that are running newer software. In extreme cases, the older version of the software may no longer be supported by the vendor. This calls for a software upgrade, and the most painless way to update software is over the IP network on a weekend night or other time when few people are using the network. Modern management servers such as the Polycom CMA 5000 server allow scheduling the upgrade of a device, a group of devices, or all devices at a preset date and time.

For soft clients, the best strategy is to avoid software distribution on CDs and install the soft client directly from the management server. When a new version of the client software becomes available, the management server will perform the upgrade without any user intervention. This approach provides software consistency and is the most technically advanced mechanism for soft client deployment in organizations; therefore, this mechanism is supported in the Polycom CMA 5000 management application.

Monitoring

Once the device is operational, it has to be continuously monitored, that is, a central management station must know if the device is working properly, and if not, get information about the type of failure. Today, dashboards are the most popular way to provide an overview of the network. Usually, they neatly aggregate and display information about the number and type of active devices (endpoints and media servers, for example), statistics on scheduled provisioning and software upgrades, the overall network health status (green, yellow, red), and the

number of connected users. If the management application has access to scheduling and conferencing information, the dashboard may also include statistics about scheduled and ad hoc conferences. If the application has access to bandwidth utilization information, it should also be included in the dashboard. Video calls require more bandwidth than data or voice applications and monitoring the bandwidth utilization is critical for smooth system performance. Figure 4 depicts a dashboard of a video management application (Polycom CMA 5000).

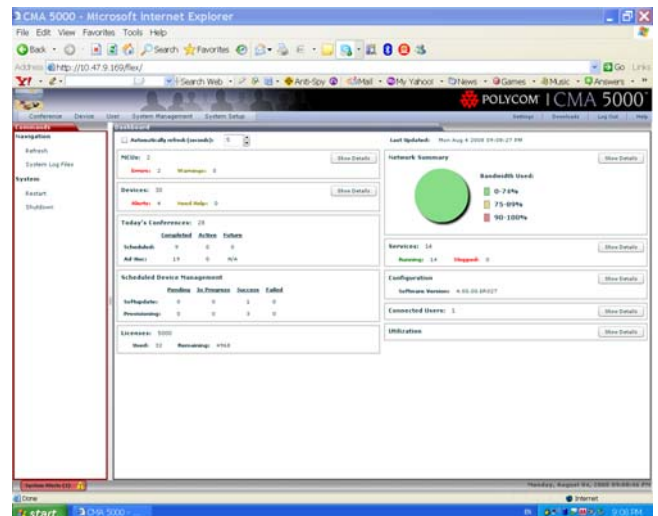


Figure 4: Video Management Dashboard

Once a problem is identified through the dashboard, the administrator can drill down to more detailed information, use troubleshooting tools to investigate the problem, and then fix it. If troubleshooting shows that the problem is within the video device, the remedy may be to upgrade the device software and reload the provisioning data. If this does not fix the problem, it may be a hardware problem that requires replacing the device. If the problem is in the IP network (high packet loss, for example), the remedy may be to limit the amount of bandwidth available to the device. A different set of troubleshooting tools can be used to investigate the failure of IP networking equipment such as routers and switches.

Licensing

More and more functions on video endpoints and servers are being delivered through software, and monitoring the use of software licenses is becoming critical for the deployment of large scale systems. In the past, licenses were delivered on diskettes, dongles, and memory sticks and these mechanisms are still being used for small installations without centralized management. As the size of the video

network grows, applying licenses locally to the devices becomes extremely inefficient.

Modern management applications keep track of licensing and allow for the use of a single licensing file for the entire video network. Centralized license management is especially beneficial to organizations that have numerous and frequent moves, adds, and changes (MAC). When a video endpoint is disconnected, and if the licensing policy allows it, the user license may be registered to another endpoint in another part of the network. Through the dashboard, the administrator can easily see when the network is running out of licenses and purchase additional licenses from the vendor.

Device Control

The device control function allows the management application to invoke particular actions in a video device, such as rebooting the device, resetting (going back to a default state), or placing a call to another device. While “reboot” and “reset” are clearly management functions, “make call” is a call control function that is widely used in the communications industry.

Scheduling

Scheduling is an important capability in video communications because a high percentage of conferences are still scheduled far in advance and for ongoing meetings rather than scheduled on an ad hoc basis. For example, schools typically plan their curriculums months in advance and if classrooms have to be connected through video the conferences are also scheduled well in advance. When the time for the scheduled conference comes, the scheduling function—part of the management application—uses its capabilities to remotely control video devices and establishes a point-to-point or multipoint conference.

Scheduling also may include reserving network resources. For instance, when a conference is scheduled, the management application may reserve resources on a conference media server to make sure the participants will be able to join the conference. If a conference is cancelled, the resources on the media server must be freed immediately, so that they can be reserved for other conferences or used for ad hoc conferences.

THE DEVICE LIFECYCLE

The primary purpose of management systems is to automate processes for remotely managing devices. It

is therefore essential to understand the lifecycle of a device in the network, and identify the management functions required in each of the lifecycle stages.

Figure 5 shows the lifecycle of a device.

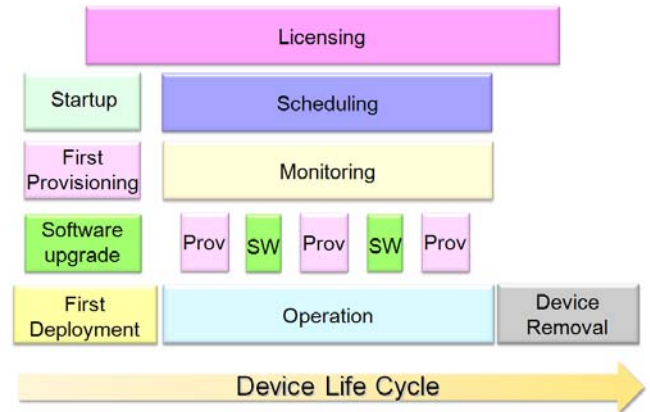


Figure 5: Device Lifecycle

When a new device is connected to the IP network, it must first get access and find its management server. It then receives provisioning information for the first time, registers with all necessary servers in the network, and starts operating. From this moment on, the device is monitored up until it is powered down or disconnected from the network. Monitoring is interrupted every time the device is physically disconnected from the network such as being moved from one location to another, during power outages, or when the cleaning crew accidentally pulls the power cord.

While operational, the device is periodically upgraded with new software. If there are changes in the provisioning information—for example, when the address of a gatekeeper or SIP server changes—all devices using this server will receive updated provisioning data. The device life ends when the device is permanently removed from the network.

De-installation of the soft client application is equivalent to removing the client from the network. The monitoring of soft clients is interrupted more frequently than that of devices because soft clients cannot be monitored when the user has exited the soft client application.

SCALABLE MANAGEMENT

What makes management scalable? Scalability of a management server is basically its ability to manage

more devices. For example, if one server can manage a maximum of 1,000 devices and another server can manage 10,000 devices, the second server is 10 times more scalable than the first.

However, maintaining 1,000 or 10,000 device profiles within the database of a server is usually not an issue. The ultimate test for a management system is a power outage that takes all managed devices down and then, when the power is restored, all devices contact the management server for provisioning data. Therefore, what really limits the scalability of management applications is the inability to supply provisioning data to a lot of devices quickly.

Management has always been a focus for the communications industry; however, the focus traditionally has been configuring Private Branch Exchanges (PBXs) and Central Office switches rather than digital and analog telephones that do not require any special management. Today's IP-based communications industry has been heavily influenced by the Internet and its "intelligent host/dumb network" model. While communications systems in the past were relying on dumb terminals (telephones) fully controlled by the PBX or CO switch, the Internet pushed forward the concept of intelligent endpoints ("hosts," in Internet terms) that deliver significant local functions and better user interfaces.

Intelligent endpoints have a lot of benefits. They also have one substantial drawback—they require extensive management. Managing intelligent endpoints is the largest challenge in the management area today. Other system elements, such as media servers and application servers, are few and far between in the video network. Their management can be manual to a certain extent. At the same time, deploying thousands of endpoints across an organization is a huge task and any manual element in the management process takes more time and can cost a lot of money.

SCALABLE PROVISIONING

Provisioning is a complex process and is therefore discussed in detail in this paper. It includes getting the devices up and running in the IP network, establishing communication channel between the device and the management application, transferring the provisioning data from the management application to the device, and updating the provisioning information.

Getting the Devices Up and Running in the IP Network

To be fully functional hosts in the IP network, all devices must have an IP address, IP address mask, default IP router (called "default gateway," for historic reasons) address, and domain name server (DNS)² address or addresses (most devices support primary and secondary DNS servers). This minimum, starting configuration can be entered manually by the technician who installs a device. Alternatively, an automated and thus more scalable way for initial IP configuration is through the Dynamic Host Configuration Protocol (DHCP).³ See Figure 6.

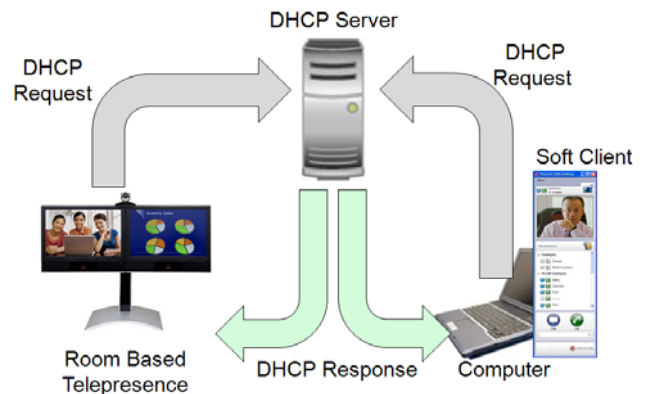


Figure 6: Initial Device Configuration Using DHCP

The configuration of a DNS server address is critical for the operation of a device. Without a DNS server, the device can only contact other network elements through their IP address, such as 98.207.197.221. Once the DNS server is configured, the device can contact network elements based on their names. An example might be server1.ABC.com where ABC.com is the organization's domain. The use of names instead of IP addresses is preferred because servers are often moved or taken down for maintenance and their IP address can change. It is very easy to make the IP address change in the DNS record but very difficult to change an IP address in all devices using this server. Therefore, the use of a DNS server increases scalability of the management solution.

Many IP networks require devices that connect to the network to authenticate before getting access to any network resources. The IEEE 802.1x protocol is the recognized standard for this authentication procedure. The device must have a certificate that identifies it and the authentication server in the network must accept this certificate before granting access, that is, before allowing IP packets to flow from and to the device.

While hard video endpoints, such as the Polycom HDX series, function as IP hosts and receive their own IP parameters from the DHCP server, soft clients like the Polycom CMA Desktop client use the IP configuration of the computer on which the soft client runs. It makes no difference to the DHCP server whether the configuration request comes from a video endpoint or from a computer on the network. The computer has to also take care of the IEEE802.1x authentication because this mechanism authenticates hardware components against the networks rather than a specific application running on this hardware.

Finding the Management Server

Since intelligent endpoints operate relatively independently, they have to be able to perform functions such as registration, authentication, encryption, signaling, and directory access, therefore they have to know the addresses for many servers in the network. These server addresses are used by devices network-wide and the logical place for this information is the management server that can distribute it to all devices through its provisioning capabilities.

There are four ways to inform the device where its management server is:

First, the user or the administrator can enter the management server address through the user interface. This approach can be used in soft client deployments during which the client runs on a computer and entering the address through keyboard is not difficult. The management server address can be distributed through e-mail or other means to soft client users. Organizations that roll out the same software image to all computers can include the soft client and the management server address in the standard image. While this method seems applicable to soft client users, our experience has shown that any end user involvement in the configuration of the soft client creates problems. Thus, the latest generation of the soft client Polycom CMA Desktop does not rely on this mechanism.

Secondly, personal and room-based telepresence systems also allow entering the management server address through the on-screen keyboard and remote control. If users do not feel comfortable doing it, the initial configuration information, including the management server address, could theoretically be delivered using a memory stick since the most recent generations of video devices have USB ports. Once the device knows the management server address, it

will contact it and obtain all necessary provisioning information.

The third way to tell a device the location of its management server is through DHCP which delivers parameters over what are termed "options." The use of most options with lower numbers (under 100) is standardized, that is, option 1 always delivers an IP address for the IP device while option 6 always delivers the IP address of the DNS server in the network. Some options above 100 are still not standardized and can be used to deliver the domain name or the IP address of the management server. Yet another approach is to use the vendor-specific option 43 that allows vendor-specific extensions to DHCP. The device vendor can define an extension and assign a number to it; the extension can then be used to transmit any kind of information, including the management server address. The device then requests DHCP option 43 plus extension X in order to receive the correct value. Nevertheless, DHCP is designed to configure local parameters in the IP network subnet while the management server is used in large video networks that may span the globe. There are also many DHCP servers and adding the management server address to all of them is not an easy task.

This leads to the fourth method: using DNS SRV⁴ to configure the management server address. DNS is a global system and is well-designed to locate servers and other hosts. DNS SRV is a standard DNS extension that allows clients to ask for a specific service or protocol for a specific domain (the "domain" is defined in DNS²), and to receive the names of any available servers. For example, if the organization deploying the video network is ABC and owns the domain ABC.com, the video device that needs management sends a SRV request for the management service (let's call it "management") in the domain ABC.com to the DNS system. DNS responds with a list of servers that provide management functions. The list can include multiple servers and each server has a weight or priority associated with it. Figure 7 shows a DNS SRV flow.

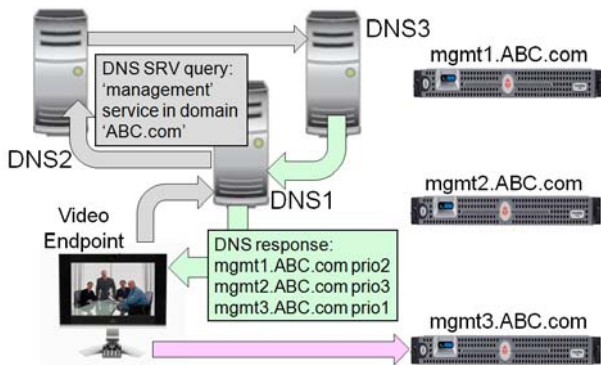


Figure 7: Using DNS SRV to Find a Management Server

The device decides which management server to contact based on this information. In the scenario described in Figure 7, DNS responds with a list of three management servers with priority 1 associated with server 3. The device then connects to management server mgmt3.ABC.com. Only if this server fails will the device try the priority 2 server (mgmt1.ABC.com).

DNS SRV therefore allows for load balancing, that is, if DNS detects that too many devices are already registered with one management server, it can change the weights or priorities of the servers in the response, and thus make new devices contact another management server. More about server redundancy and load balancing are discussed in the paper.¹

To be complete, network scanning and an alternative method for establishing a connection between a management application and managed devices should be included. With this pooling approach, the management application scans a range of IP addresses and determines if these IP addresses belong to video devices that require management. The address range is set by the management application administrator. If the range is large, the scanning process may take hours and even days, so limiting the address range (based on prior knowledge of devices location) is recommended. The management application must periodically or continuously scan the network to find new devices.

This method has two huge drawbacks. While the management application is busy scanning the network, its performance and scalability suffer. Even more importantly, scanning is considered a security threat by IT administrators because many hackers use scanning techniques to find weak network elements and break into the network. Standard security software

identifies scanning network elements and shuts down their access to the network.

Since security and scalability are at the core of the Polycom VC2 vision, Polycom management applications, such as Polycom CMA 5000, do not deploy scanning and use the secure and scalable methods based on DNS described above.

Communication between Device and Management Application

Once the device has found the appropriate management server, it connects to it to retrieve detailed provisioning information. Different protocols can be used, including Telnet, Trivial File Transfer Protocol (TFTP),⁵ File Transfer Protocol (FTP),⁶ and Hypertext Transfer Protocol (HTTP)⁷ or its secure version HTTPS. TFTP is the least secure mechanism since the device does not need a user ID and password to register with the management application. FTP provides authentication through an FTP username and FTP password, but is relatively slow and provides no encryption. HTTP is faster and provides authentication but no encryption. Modern management applications, such as the Polycom CMA 5000 application and the new generation of endpoints such as the Polycom HDX series, use HTTPS, the most secure protocol for exchanging information in IP networks. HTTPS uses Transport Level Security (TLS)⁸ to encrypt the data between the device and the management server.

Provisioning data comes in many formats from simple comma-separated text files to cryptic encoding schemes, but the Extensible Markup Language (XML)⁹ has become the industry standard, mainly because it is flexible and easy to understand by both machines and humans. XML provisioning data is basically a text file (see example in Appendix A).

There are two ways to handle provisioning: "push" and "pull." With push, the management application sends (pushes) the file to the device. Most of the work is done by the management application, and this has impact on its scalability.

With pull, the device retrieves (pulls) the XML provisioning file from the management application. Since the device does most of the work, the management application can support more devices and scale. The pull approach also leans towards the Web browsing paradigm where the client/browser retrieves information from the Web servers on the Internet. Thus, the pull method is more suitable for complex (multidomain) networks and works well

across firewalls and Network Address Translation (NAT) devices.

Not surprisingly, there is a trend in the industry to move from push to pull as way to increase scalability of management applications. Figure 8 shows the interaction between device and management server.

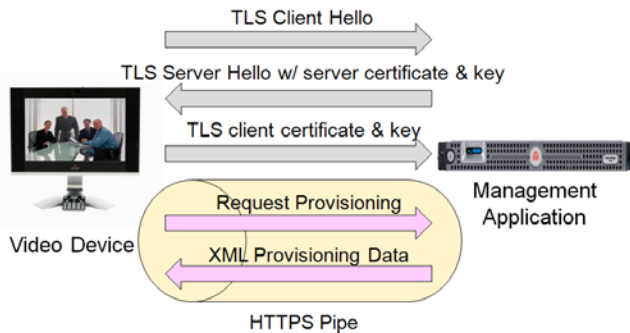


Figure 8: Pulling XML Provisioning Data over HTTPS

When the XML file is transferred to the managed device, the device's XML parser goes through the XML file line-by-line and applies the values to the specified parameters. If the parser does not understand a parameter or a value, it logs an error in the device log file and continues through the remainder of the XML file. This mechanism is robust and scalable. Moreover, it has been widely deployed and field-tested with the Polycom SoundPoint® IP desktop phones and Polycom SoundStation® IP conference phones and is now implemented in the Polycom HDX series of video endpoints and supported by the Polycom CMA 5000 management application. The mechanism is extensible, that is, when new devices with more functions are deployed, the XML provisioning file can be extended with new parameters. Newer devices will interpret the parameters correctly while older devices, which may not support these parameters, will simply ignore them.

Provisioning Key Server Addresses in the Device

The XML provisioning data may include hundreds of parameters, including IP network parameters, quality-of-service parameters (for example, DiffServ and priority values), timeouts for the supported communication protocols, time and language settings, home screen settings (for video endpoints), and, most importantly, the addresses of the network servers. Depending on the functionality offered in the video network, the list of must-know servers may be shorter or longer, but it usually includes the call control, directory, and presence servers.

In H.323-based networks, the device must know the H.323 Gatekeeper address. In SIP based networks, the device will need to know the addresses of the SIP Registrar, SIP Proxy, and (optionally) SIP Redirect servers. More details on the functions of these servers are in the paper *Migrating Visual Communications from H.323 to SIP*.¹⁰

Device users need to be able to find their communication partners in some form of an organization directory (often called “phone book” or “address book” for historic reasons). Directory access has changed significantly over the years, and moved away from proprietary protocols and towards the use of the standard Lightweight Directory Access Protocol (LDAP).¹¹ The protocol allows devices and servers to contact a directory, authenticate for access, and then search, based on certain parameters. For example, the request may be for the e-mail addresses of a person—using first and last name—or of a group of people—using department name like “service” or “marketing.” The directory executes the search and then sends a LDAP response with all found entries (e-mail addresses in this example).

With the wider availability of video, directories have evolved to include parameters for video devices, for instance, H.323 aliases for H.323 endpoints. A special schema was standardized for retrieving video-related data from directories: H.350.¹² More information about LDAP and H.350 can be found in the paper *Scalable Infrastructure for Distributed Video*.¹ Figure 9 shows the most important servers with which a device has to communicate.

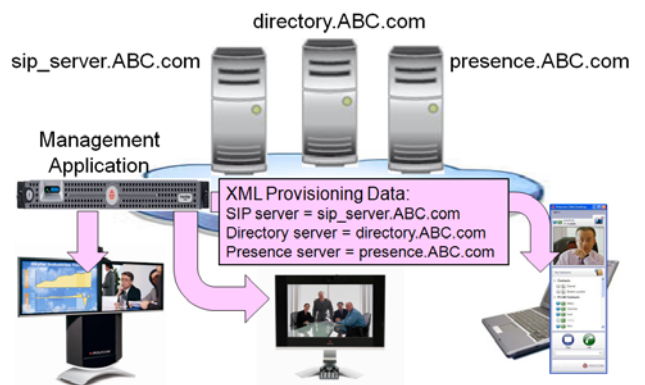


Figure 9: Key servers for video devices and soft clients

Presence is becoming more important in communication systems and is now supported in many devices, for example, the Polycom HDX series endpoints and the Polycom CMA Desktop soft clients support presence. Two protocols are used for

presence—XMPP¹³ or SIMPLE¹⁴—and the presence server address must be provisioned to the device. The Polycom CMA 5000 application supports the XMPP protocol for exchanging presence information with Polycom CMA Desktop soft clients and Polycom HDX and Polycom VSX® endpoints. The management application also uses the XMPP to enable instant messaging functionality in the soft clients.

Provisioning of Groups and Sites

While the server’s addresses discussed above must be provisioned in all devices in the video network, there are certain parameters and capabilities which only apply to a subset of devices. For example, home screen makes sense for a room-based telepresence system but not for a soft client; therefore, it is a good idea to create a group of room-based telepresence systems in the organization and provision them with the same home screen information.

Another example is instant messaging (IM). While it is natural for soft clients, using IM is not a good idea on room-based telepresence systems. The management application should have the capability to create a group of soft clients and provision the entire group with an instant messaging server address. In Figure 10, all soft clients, independent of their location, are combined in a group and get the same group provisioning information (as illustrated by the pink arrows).



Figure 10: Provisioning groups and sites

Provisioning parameter values may also differ by site. For example, if the devices in the Los Angeles office have to use an Outbound Border Proxy, the OBP address must be provisioned to all devices at this site (as depicted by the light green arrows in Figure 10).

Devices may be members of multiple groups/sites, for example in Figure 10, the soft client in the Los Angeles office is a member of the Los Angeles site and the soft client group. The management application should provide clear rules regarding which provisioning information has higher priority and what setting should be applied if there is a conflict. The Polycom CMA 5000 application gives administrators full flexibility with the provisioning of groups and sites, and provides an intuitive and logical user interface.

SCALABLE SOFTWARE UPGRADES

The first issue for the administrator is how to identify the software version of each device in the network. When the device finds its management application, it sends information about its software version and this information must appear in the management screen next to the user name, address, and other aliases of the device. The administrator should be able to select the device, select the software version (to be used in the upgrade), and start the process.

Not all software versions are created equal and not all users are the same, so it is prudent to match software version and user type. Newer software that has not been extensively field tested is probably more appropriate for devices used in a pilot environment before rolling them out to all users. New software may not be appropriate for installation on a device used by C-level executives in an organization. The administrator may want to keep them one or two versions behind the “bleeding edge” to make sure that executives do not have to struggle with bugs.

As a result, the administrator may need to maintain at least three versions of the software for the same device type on the management server. Moreover, different types of devices need different software versions, as we earlier depicted in Figure 3. In addition, the number of software versions grows if there are older devices in the network that do not have the necessary hardware resources to run the latest software versions.

In short, dozens of different software versions may reside on the management server, and are downloaded by various devices across the network. For increased scalability, the device software can be stored on a separate software download server that is controlled by the management application. When the application receives a request for software upgrade, it provides the address of the software download server to the device and lets it complete the download without using any more management server resources.

Scalable video networks have thousands of devices thus upgrading them one by one is not efficient. Therefore, modern management applications, such as the Polycom CMA 5000 application, allow the administrator to upgrade groups of devices.

When the group is large, there is a risk that all devices in the group will contact the download server at the same time which can lead to errors and can slow down the upgrade process. There is always a limit to the number of simultaneous connections that the download servers can support. It is therefore a best practice to cascade the requests, that is, the management application initiates the upgrade for the first device, waits for a second or two, then initiates the upgrade of the second device, and so on. Figure 11 illustrates this concept.

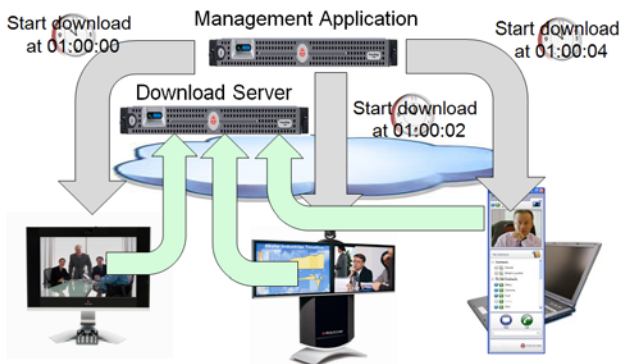


Figure 11: Scalable cascaded software deployment

An additional level of flexibility for the administrator can be provided through creating batches (scripts with commands) in the management application, and setting date and time for their execution. This assures the least interruption of the video network operation due to software upgrade, and makes an administrator's presence during the upgrade unnecessary.

By setting up rules, the administrator can also define under which conditions a particular device, group, or site should be upgraded to a particular version of the software. This level of automation is available in the Polycom CMA 5000 application and allows an administrator to manage large networks of thousands of devices and soft clients.

MONITORING

Monitoring capabilities are critical for a management application. The level of monitoring can be adjusted based on how much the administrator wants to know

about the network. This is usually correlated with how mission-critical the video network is for the organization. Monitoring in IP networks has moved to the Simple Network Management Protocol (SNMP) standard,¹⁵ s protocol used to monitor all types of IP network elements, including routers, switches, and servers.

There are three SNMP protocol versions. The original V1 is only used by very old IP network elements. V2 enhances protocol operations and changes the message formats. It is the most widely deployed version today. The latest version—V3—adds security and remote configuration capabilities to the protocol. The major benefit of SNMP is its compactness. The messages are short and simple and the data is transmitted in an encoded/compressed format, so that SNMP messages can be transported in a single IP packet and without fragmentation. This limits the impact on the IP network and allows the management application to scale.

While SNMP can request data (GET command) and change data (SET command) in devices, the alarm function is most widely used. It allows the device to spontaneously send a notification to the management server if certain events out of the ordinary happen on the device. Figure 12 depicts this concept.

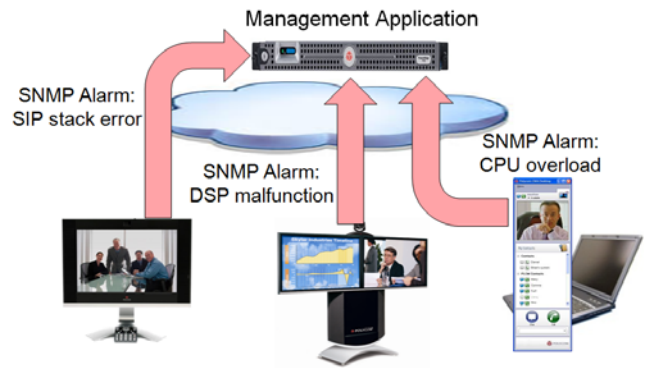


Figure 12: Using SNMP for Monitoring

In a very basic scenario, the monitored device reboots due to a failure and sends an SNMP alarm with information about the reboot to the management application. More sophisticated monitoring covers dozens of critical events (memory overflow, DSP failure, corrupted configuration data, for example), and organizes these events in levels. The administrator can then set the monitoring level in the device through a parameter in the provisioning data, and only receive notification (an SNMP alarm) when an event belonging to this level or higher occurs.

SNMP alarms can also be used to monitor quality of service (QoS) for video calls. To enable that, devices must support an extension of the Management Information Base¹⁶ used by SNMP. The extension should define the QoS parameters to be monitored, such as the average round trip delay, maximum packet loss, maximum consecutive lost packets, and maximum jitter. The management application then must apply threshold values to each of these QoS parameters. Devices and soft clients will continuously measure the QoS parameters and generate SNMP alarms if one or more of the parameters exceed the configured threshold values. Similar techniques can be used to monitor when conferencing servers, recording servers, and gateways are close to running out and have run out of resources.

The management application receives the alert, recognizes the error condition, and reacts to remedy the problem. If there is a problem with the protocol stack, the management application may initiate a software upgrade of the device to make sure the correct software is used. If there is a problem with the configuration data on the device, the management application may initiate new provisioning for the device. If the error indicates a severe hardware problem, the management application may notify the administrator and/or local support personnel that the device has to be replaced. More sophisticated applications will block scheduling of malfunctioning devices in conferences.

SNMP was designed for monitoring hardware network elements and is therefore not widely supported in soft clients. If the user exited the soft client application, the SNMP monitoring application would lose connectivity to the soft client and would report that as a network element failure. The last thing a network manager wants to see in the dashboard is red flashing icons for all soft clients to which the monitoring application lost its connection.

The Polycom CMA Desktop client, for example, generates a report file after a call is completed; The Polycom CMA 5000 application can then parse the file. This is in line with the Polycom CMA Desktop use of the server model in which the server can aggregate and report on events and statistics.

LICENSE MANAGEMENT

The traditional approach to licensing is to load individual license files into each of the devices in the network. This can be accomplished manually, through a dongle or memory stick. However, a more scalable way is to download the device license file from the

vendor's Web site. Based on the sales contract, the vendor knows what type and how many licenses the organization has purchased, and the vendor's licensing server can issue individual licenses for the devices deployed in the organization. The license file is associated with a particular hardware attribute, and MAC addresses are widely used in licensing of IP devices because of their uniqueness. Each IP interface for a device has a worldwide unique MAC address, for example, the MAC address of my laptop is 00 18 DE 1E 02 5F and there is no other device in the world that has the same MAC address.

Scalable networks with thousands of devices quickly uncover the weaknesses of these licensing models. When a device is down or is permanently removed from the network the license cannot be used. The network administrator has to request from the vendor to remove the association of the license with the MAC address of the broken device and instead associate it with the MAC address of a new device. This process has been automated but still requires a lot of interactions and creates opportunities for errors. The most advanced approach to centralized licensing is CAL, or Client Access License. CAL does not associate the license with a particular device but assigns a pool of licenses dynamically to the devices in the network. The model also works very well for soft clients. Figure 13 illustrates the concept.

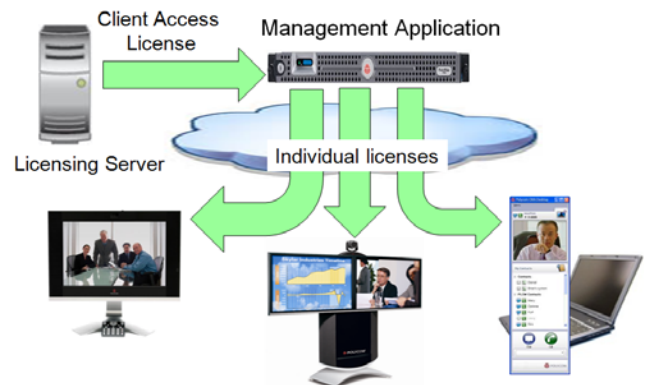


Figure 13: Centralized Licensing

Centralized license management is especially beneficial to organizations that have numerous and frequent device moves. If a device is disconnected, a user license may become available - based on policy - and can be used for another device in another part of the network. Through the dashboard, the administrator can easily see when the network is running out of licenses and purchase additional CAL licenses. Modern management applications such as Polycom

CMA 5000 deploy CAL and keep track of licensing throughout large video networks.

DEVICE CONTROL

Through device control, the management application 'tells' the managed device what to do. Today the function is most frequently used to initiate point-to-point calls between endpoints. When the scheduling function on the management server detects that a scheduled call is point-to-point (scheduling cannot therefore use the conference server to set up the calls to the endpoints), the server uses the device control function to tell endpoint A to call endpoint B.

The only way to control legacy devices is through setting up a Telnet session (actually logging into the device), and running a command in the command line interface. The process is automated and made invisible to the management application administrator, but it is slow and inefficient.

Modern device control mechanisms rely on lightweight transport protocols such as HTTP, and use lightweight formats such as XML. The "make call" functionality, discussed above, is a very simple scenario that can be implemented in variety of ways. The more general problem of "telling" the device to do complex things has been addressed by call center vendors and there are standards in place that define hundreds of functions that can be invoked on a controlled device. For example, Ecma International standard TR/87¹⁷ describes the use of the CSTA protocol over a SIP session to control and monitor SIP user agents embedded in devices. The uaCSTA specification defines the transport of CSTA commands in SIP INFO messages (carrying content-type application/csta+xml). SIP INFO messages can only be used in the context of a SIP dialog, that is, after the exchange of INVITE, 200OK, and ACK messages establishes the communication between two parties. Once the SIP dialog (call) is set up, SIP INFO messages can be exchanged by leveraging the same SIP authentication, encryption, and firewall and NAT traversal that other SIP messages use. The SIP dialog is terminated with a SIP BYE message, and no CSTA/INFO messages can be exchanged after that. More details about the SIP call flow can be found in the Polycom white paper *Migrating Visual Communications from H.323 to SIP*.¹⁰

SCHEDULING

Scheduling is used for planning conferences ahead of time, and allows for both point-to-point calls (two

endpoints are connected) and multipoint calls (three or more endpoints are connected). Scheduling can be handled by a separate application running on a separate server in the network but the synergies between scheduling, provisioning, monitoring, and device control lead to substantial benefits when the scheduling function is collocated.

The scheduling function uses device control ("make call") to set up point-to-point calls. It uses the conferencing server to schedule multipoint calls, essentially sending the list with conference participants (and their associated endpoints) to the conferencing server and asking it to dial all of them and to connect them in a conference. Figure 14 is an example of a scheduling screen from the Polycom CMA 5000 application.

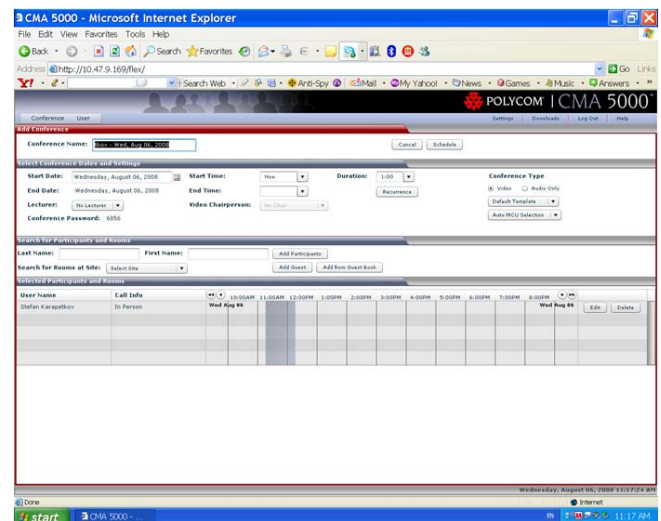


Figure 14: Scheduling Screen Example

A simple scheduling function will prompt the conferencing server the second the meeting must start. More sophisticated scheduling includes resource reservation, at least in the bottleneck areas, including conferencing resources and network bandwidth. Reserving conferencing resources is straightforward. The scheduling function asks the conferencing server (or the management server controlling it) to block the necessary number of resources. The resource calculation is based on the number of participants and required video quality—CIF, Standard Definition, or High Definition. Once the resources are reserved, they cannot be used for the duration of the scheduled conference by other scheduled or ad hoc conferences.

Reserving network bandwidth in the IP network is more complex and requires communication between the scheduling function and a QoS policy engine in the IP network.

SPECIAL CONSIDERATIONS

The last section of this paper addresses the special scenarios of a power outage, backup and restore, and management of legacy devices.

The Power Outage Scenario

The power outage scenario puts a lot of strain on the management application, especially if the organization's offices are concentrated in the affected geographical area. Once power is restored, all video devices will try to contact the management application, and with thousands of devices in the network, this may overwhelm the management server. There are two approaches to address this scenario.

First, multiple management servers can be installed in a cluster, designed to handle the maximum (worst case) load. The use of DNS for load balancing across a pool of management servers was discussed earlier in this paper and is depicted in Figure 7. This is an expansive approach that is more appropriate for service providers. A more cost effective approach is to use cascading in which each device generates a random number equal to the number of seconds the device must wait until contacting the management application.

Backup and Restore

The power outage scenario described above may also have an impact on the management server. It is therefore a best practice to keep the server in a data center with backup power generators or batteries. If the organization does not have access to such a facility, it is vital to regularly backup the server information. The backup procedure should include all device provisioning files, group and site definitions, any batches with commands for software upgrades, and rules defined by the administrator.

The backup and restore capability is also very important during hardware upgrade, when the organization moves to more powerful server hardware, for example. It allows the administrator to export the entire management data (including provisioning, software upgrade, and monitoring settings) from the old server and then import them into the new server without a long downtime for users.

Management of Legacy Video Endpoints

Legacy video endpoints use proprietary management protocols and can only be supported by management applications that implement these proprietary protocols. Since every endpoint has its own proprietary management interface, the management server ends up supporting numerous plug-ins. This limits its scalability. There is a tradeoff between scalability of management servers and the ability to support legacy endpoints for backward compatibility.

Polycom took a balanced approach to supporting legacy video endpoints. The Polycom CMA 5000 application manages legacy Polycom VSX series endpoints and non-Polycom endpoints based on their proprietary protocols. It manages the Polycom HDX series video endpoints and Polycom CMA Desktop clients using the new set of standards-based mechanisms and protocols described in this paper. Future devices—endpoints and other—will be supported using standard protocols only.

CONCLUSION

Polycom's VC2 vision stresses the importance of scalability and management for transforming today's video conferencing into tomorrow's visual communication. Since the management application touches all devices in the video network, including endpoints, media servers, gateways, and recording servers, the scalability of the management application is a key prerequisite for building large video networks.

Polycom has developed deep expertise in management applications and created an architecture that meets the high-performance requirements of future visual communication networks. This architecture deploys new flexible and standards-compliant mechanisms to manage new devices while continuing to support legacy equipment, thus providing a bridge between new and legacy systems.

In summary, Polycom is working to protect its customers' investments in video devices, while leading them towards its VC2 vision to bring video into the mainstream and make visual communication essential in both personal and professional lives.

APPENDIX A: XML PROVISIONING FILE EXAMPLE

```
<?xml version="1.0" encoding="UTF-8"
standalone="yes" ?>
= <ProvisionResponseMessage
  xmlns:ns2="http://polycom.com/polaris">
  <protocolVersion>1.0</protocolVersion>
  <status>OK</status>
= <provItems>
  <provItemsVersion>1.0</provItemsVersion>
= <XMPP>

  <enableXMPPDirectory>TRUE</enableXMPPDirect
  ory>
  ...
</XMPP>
= <H323>
  ...
</H323>
= <SIP>
  ...
</SIP>
= <LDAP>

  <enableLDAPDirectory>TRUE</enableLDAPDirecto
  ry>
  ...
</LDAP>
= <CONFIG>
  ...
</CONFIG>
</provItems>
</ProvisionResponseMessage>
```

REFERENCES

1. Karapetkov, S., 2008. *Scalable Infrastructure for Distributed Video*. Polycom Whitepaper, June 2008.
2. Mockapetris, P., 1987. *RFC 1035 Domain Names Implementation and Specification*. IETF Document, November 1987.
3. Droms, R., 1997. *RFC 2131 Dynamic Host Configuration Protocol*. IETF Document, March 1997.
4. Gulbrandsen, A. et al, 2000. *A DNS RR for specifying the location of services (DNS SRV)*. IETF Document, February 2000.
5. Sollins, K., 1992. *RFC 1350 Trivial File Transfer Protocol*. IETF Document, July 1992.
6. Postel, J., 1985. *RFC 959 File Transfer Protocol*. IETF Document, October 1985.
7. Fielding, R. et al, 1999. *RFC 2616 Hypertext Transfer Protocol (HTTP)*. IETF Document, June 1999.
8. Dierks, T. et al, 2006. *RFC 2246 Transport Layer Security (TLS)*. IETF Document, April 2006.
9. Bray, T. et al, 2006. *Extensible Markup Language (XML) 1.1 (Second Edition)*. W3C Recommendation, September 2006.
10. Karapetkov, S., 2008. *Migrating Visual Communications from H.323 to SIP*. Polycom Whitepaper, April 2008.
11. Zeilenga, K., 2006. *RFC 4510 Lightweight Directory Access Protocol: Technical Specification Road Map*. IETF Document, June 2006.
12. H.350 *Directory Services Architecture for Multimedia Conferencing – Base Architecture*. ITU-T Recommendation, August 2003.
13. Saint-Andre, P., 2004. *RFC 3921 Extensible Messaging and Presence Protocol (XMPP): Instant Messaging and Presence*. IETF Document, October 2004.
14. Campbell, B., 2002. *RFC 3428 Session Initiation Protocol (SIP) Extension for Instant Messaging*. IETF Document, December 2002.
15. Case, J. et al, 1990. *RFC 1157 Simple Network Management Protocol*. IETF Document, May 1990.
16. McCloghrie, K. et al, 1991. *RFC 1213 Management Information Base for Network Management of TCP/IP-based internets: MIB-II*, March 1991.
17. ECMA-TR/87 *Using CSTA for SIP Phone User Agents (uaCSTA)*. Ecma International Document, June 2004.

ABOUT THE AUTHOR

Stefan Karapetkov is Emerging Technologies Director at Polycom, Inc. where he focuses on visual communications market and technology. He has MBA from Santa Clara University (USA) and an MS degree in Engineering from the University of Chemnitz (Germany).

ACKNOWLEDGEMENTS

I would like to thank my colleague Kendall Myers for his contributions to this work. Special thanks to Dean Schoen and Brian Kerns for their review comments.