A Report for

DG Interpretation (SCIC)

Videoconferencing in the Context of Simultaneous Interpretation

24 August 2012

Engagement: 330000646

Gartner.

Table of Contents

Executive Summary

Executive Summary

Study Objectives

Within the EU, many conferences and meetings are organized where representatives from different member states can meet face to face and discuss important topics together. However, the related travel costs are very significant, and sometimes there are no possibilities at all for representatives to travel to these conferences. Videoconferencing solutions address both these issues. Videoconferencing (VC) can be defined as *conducting a real-time two way meeting between two or more participants sitting at different sites, which are connected by telecommunication technologies enabling the participants to see and hear each other*. VC thus allows for meetings with geographically dispersed groups, without the need to travel.

What specifically differentiates conferences in the EU context from more general conferences, is the diversity of the languages used, and therefore also the need for simultaneous interpretation. Interpreters have very strict requirements for providing simultaneous interpretation services, related to e.g., their view on the participants, the quality of the audio and the room acoustics. Only when these are met, the interpreters can perform their task adequately. Therefore, when considering VC for conferences with simultaneous interpretation, the applied VC solution must be able to support all the requirements.

DG Interpretation (also known as SCIC) requested Gartner to answer the key question what VC solutions could be viable for conferences with simultaneous interpretation, and additionally what preconditions have to be met to successfully implement and manage these solutions in an EU context.

Analyses

A baseline assessment conducted by Gartner among the different stakeholder groups (e.g., interpreters, meeting organizers, support staff, management of different DG's, and conference participants) indicated that some stakeholder groups thus far were not involved in the (requirements) process at all. Additionally Gartner found that the ambitions of the different stakeholder groups regarding the use of VC for conferences with simultaneous interpretation were not fully aligned (and sometimes even conflicting). This made it difficult to e.g., agree on an adequate set of requirements and design principles, which was fully supported by all stakeholders. Instead, Gartner could only use an existing set of specific interpreter requirements (formulated in a number of previous studies) as input for identifying best fit VC solutions.

Based on thorough market analysis, Gartner found that there is currently no single commercially available product on the market which meets all requirements 'out of the box', but that the most viable VC solution consist of a combination of different software-based VC technologies. However, in each case significant development and customization effort will be necessary.

In order to ensure that such a customised VC solution will always be of a consistent and repeatable quality, it is also key to have a sufficiently high *level of end-to-end control*[1]. When this is not the case, it is likely that the delivered quality of the VC solution will not be adequate for the interpreters. In practice this means that all parties in the EU will have to use the same VC solution, apply the same standards and processes, etc. This requires strong inter-institutional cooperation.

[1]This refers to the level of control over the interconnecting network (bandwidth, quality), room conditions (lights, acoustics), and the VC technologies used in all sites)

Based on the above, Gartner concludes that the challenge with regards to videoconferencing with simultaneous interpretation is not only a technical challenge, but above all a governance- and change management challenge.

Recommendations

In order to successfully implement an end-to-end VC solution in the EU context, Gartner strongly recommends starting this initiative with establishing a clear and agreed on *governance model,* necessary to ensure adequate inter-institutional cooperation. This governance model should make the responsibilities and mandates of each party fully clear (e.g., responsibility for development, management & support), and should also address the required funding. Additionally, there should be clear and ongoing focus on *change management*, as only this will ensure sufficient involvement of all the different stakeholders, alignment of their requirements and expectations, compliance to the agreed on governance model, and finally acceptance of the to be developed end-to-end VC solution.

After institutionalizing this governance model, Gartner recommends to apply a *three-phased road map* in which is it is critical to first establish a stable VC solution on the internal network. This stable internal solution not only provides the baseline from which to extend, but also serves a critical role as a governor toward more ad hoc demand in later stages.

Phase 1 of the road map should therefore focus on establishing a ‘Room-to-Room based solution’ within a fully controlled environment. For this it is important to first agree on adequate set requirements and design principles with all key stakeholders. The next step then should be selecting, and subsequently customising this VC solution to meet all requirements. Before taking this VC solution into production, it should be tested thoroughly. Additionally, standards for among others the required room settings and processes needs to be developed and prescribed. Only when the internal VC solution proved to be fully successful, Gartner recommends commencing with the next phase of the road map.

Phase 2 of the road map focuses on selecting and customising a software-based ‘Personal desktop VC solution’, again to be used in fully controlled environment only. This Personal desktop VC solution should also be thoroughly tested before rolling it out to distant sites. Latter should be done in step by step tranches. Only when the Personal desktop solution proved to be successful in a sufficient number of distant sites and has been fully internalized, Gartner recommends commencing with the next phase of the road map.

Finally Phase 3 of the road map should focus on trying to facilitate meetings where end-to-end control is not at the highest level. The main goal is to search for the tipping point where the VC solution, even when there is no full end-to-end control, might still be acceptable for the interpreters to work with during simultaneous interpretation. This will allow for e.g., also using the VC solution for ad hoc events. When it has been determined where this tipping point is, a set of ‘operating principles’ should be defined, prescribing under what conditions such meetings can take place. Elements of these operating principles should be among others the required bandwidth, room conditions and specific process agreements.

Introduction

Introduction

Background and Study Objective

Within the EU, many conferences and meetings are organized where representatives from different member states can meet face to face and discuss important topics together. However, the related travel costs are very significant, and sometimes there are no possibilities at all for representatives to travel to these conferences. Videoconferencing solutions address both these issues. Videoconferencing (VC) can be defined as conducting a real-time two way meeting between two or more participants sitting at different sites, which are connected by telecommunication technologies enabling the participants to see and hear each other. VC thus allows for meetings with geographically dispersed groups, without the need to travel.

What specifically differentiates conferences in the EU context from more general conferences, is the diversity of the languages used, and therefore also the need for simultaneous interpretation. Interpreters have very strict requirements for providing simultaneous interpretation services, related to e.g., their view on the participants, the quality of the audio and the room acoustics. Only when these are met, the interpreters can perform their task adequately. Therefore, when considering VC for conferences with simultaneous interpretation, the applied VC solution must be able to support all the requirements.

DG Interpretation (also known as SCIC) requested Gartner to answer the key question what VC solutions could be viable for conferences with simultaneous interpretation, and additionally what preconditions have to be met to successfully implement and manage these solutions in an EU context.

Scope

The scope of the study is the application of VC technologies for the conferences with simultaneous interpretation in an EU context. Therefore the study only addresses VC technologies for the relevant meeting settings like:

- Conference room-to-one participant
- Conference room-to-room (with multiple participants)
- Conference room-to-multiple rooms (with one or multiple participant(s))

Included in the scope are all different parties involved in these conferences with simultaneous interpretation.

Explicitly not in the scope of the study are:

- Hands on evaluation of videoconferencing solution
- Recommendations for improvement of the video conference process
- The integration of video conference material (online or edited afterward) in other applications such as existing websites etc.
- A detailed vendor selection process (proof of concept)

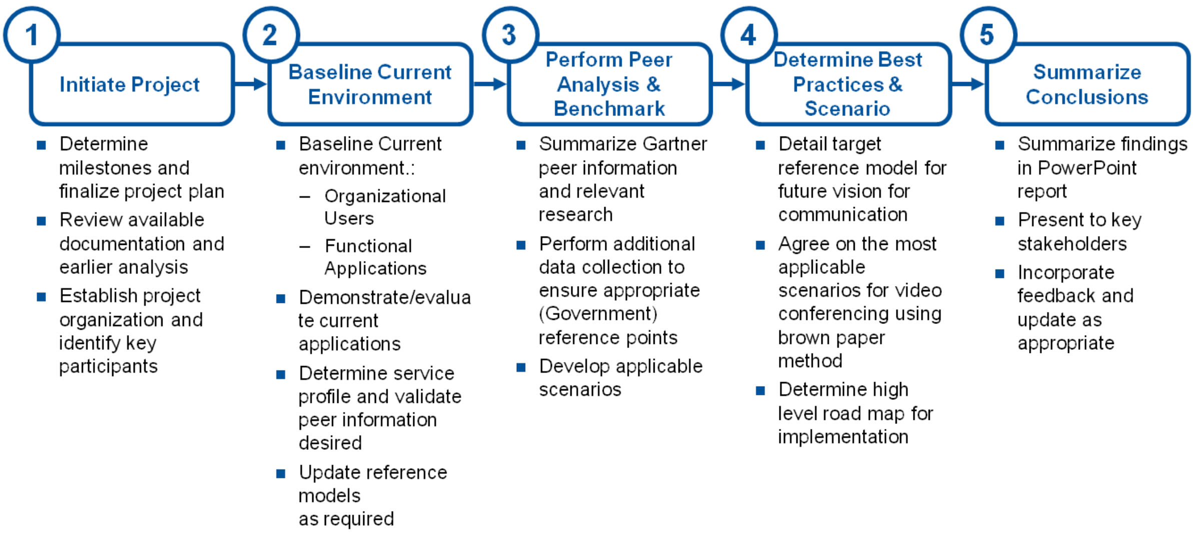
Study Approach

One of the most important criteria for the study was to involve future VC users, and specifically the interpreters, as much as possible. Therefore Gartner applied an approach based on gathering information via analysis of available documentation, interviews, peer analysis and benchmark, and a validation of this information through intensive involvement of major stakeholders. Based on these findings, Gartner was able to draw the major conclusions and to formulate critical recommendations.

The approach consisted of five major steps (see Figure 1).

- Step 1 — Initiate the Project — was to announce the engagement, to discuss, detail and agree the objectives of all stakeholders, to agree on the documentation and interviews and to agree on the detailed planning.
- Step 2 — Baseline Current Environment — the goal of this step was to study relevant documentation, and gather information form key-stakeholders and experts through interviews and through workshops to assure involvement. This step has also included a so-called Findings Workshop, which was used to present, discuss and agree on the consolidated findings.
- Step 3 — Perform Peer Analysis & Benchmark — No solutions have been evaluated in isolation, but rather a combination of separate solutions combined into scenarios that could accommodate the identified multilingual videoconferencing needs. These scenarios, and more important the future vision with regards to multilingual videoconferencing, was shared with the SCIC steering committee.
- Step 4 — Determine Best Practices & Scenario Analyses — The different scenarios and use cases were discussed in an elaborate and interactive manner, ensuring the SCIC steering committee members were involved and are were able to provide input, resulting into preferred scenarios and use cases.
- Step 5 — Summarised Conclusions — the conclusions and recommendations, as well as the input gathered in Step 4 were included in a recommended road map which was presented to the SCIC steering committee in the final presentation.

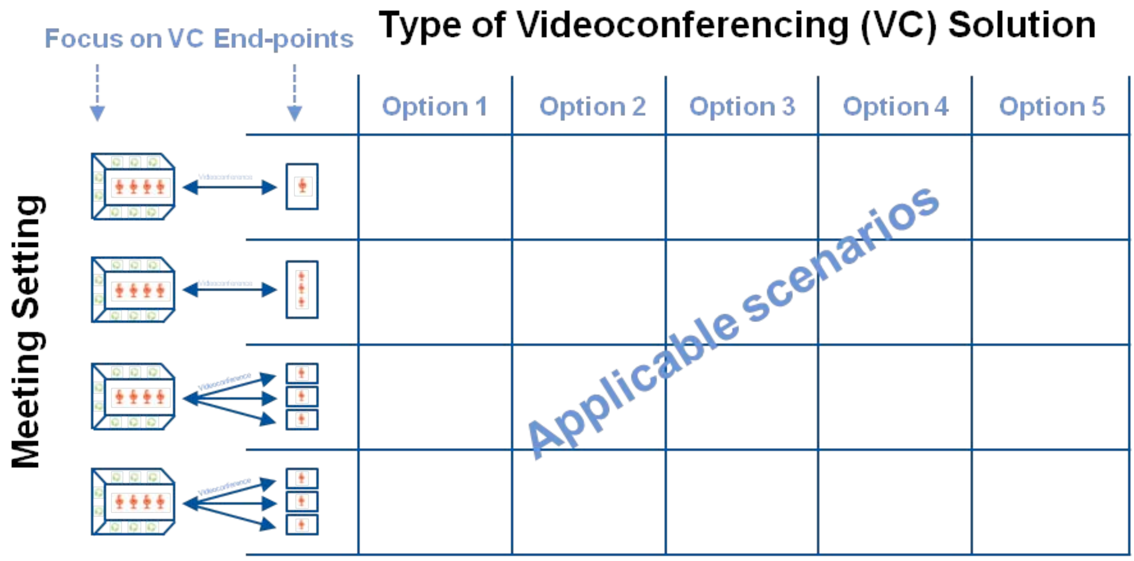
Figure 1. Gartner Approach



Applying this approach resulted in different scenarios which were based on all relevant ‘meeting settings' (or use cases) and the specific VC solution viable for these use cases (see Figure 2). Conforming to best practices, Gartner focused on so called “endpoints[1]" for each of the sites where the videoconferencing solution will be used.

[1] An endpoint is where physical and logical connections to audio systems such as speakers and microphones are made. Additionally, a videoconferencing endpoint must be connected to a display in order for video conference users to see and hear one another.

Figure 2. Scenario



Besides proprietary Gartner Research, Gartner included the results of previous studies, and relevant information sources in the analyses, such as:

- Conclusions of Fraunhofer study and other relevant previous studies conducted on behalf of SCIC
- The recently developed SCIC VC solution (still in test-phase)
- SCIC Interpreters' standards working for the European Commission

Videoconferencing Technology

Videoconferencing Technology

In this chapter the concept of videoconferencing is defined and the core components of videoconferencing technology solutions are briefly described.

Videoconferencing Defined

Videoconferencing involves the use of a collection of interactive telecommunications technologies which allows people in two or more locations to interact real time via two way audio and video transmissions. Videoconferencing is a similar technology to the standard office conference call, except with a videoconference you can actually see the person you are talking to. A videoconference is slightly different to a videophone call because it can involve groups of people — hence the word conferencing, whereas a videophone call usually only involves one person.

Within a business environment a videoconference could be used to set up a meeting between two business people or multiple boardrooms full of many people. Aside from audio and visual data, videoconferences can be employed to share computer files or papers across multiple participants.

Gartner applies the following definition for videoconferencing:

Videoconferencing is conducting a real-time two way meeting between two or more participants sitting at different sites, which are connected by telecommunication technologies enabling the participants to see and hear each other.

Videoconferencing settings ranges from point-to-point conversations between people in two different sites to multipoint videoconferencing, which allows three or more participants at different locations to sit in a virtual conference room and communicate as if they were sitting right next to each other. As videoconferencing provides organisations with the ability to meet and to work with others over a distance, there are potentially multiple benefits for using videoconferencing. Examples of this are reducing travel costs, improve use of executive time, and allowing for meetings via video and data conferencing with geographically dispersed groups at relatively short notice.

Components of Videoconferencing Systems

The fundamental technology at the heart of videoconferencing is real time digital audio- and video compression. The hardware (or software) that performs compression is called a *codec* (coder/decoder), and is regarded as the “brain and heart” of the videoconferencing system. This component takes the video and audio from the camera and microphone or audio sub-system and compresses it down, transmits it over an IP network (or legacy ISDN digital phone circuit) and expands (or decompresses) the incoming video and audio signal so that it can be viewed on a display device.

The videoconferencing industry continues to rely on proprietary technologies for their codecs, although standards-based signalling and encoding protocols are increasingly being adopted within vendor products. Several digital encoding formats are in use, like H.264 and MPEG-4.

The other main components a videoconferencing system includes are:

- Camera: the camera types range from a small camera that sits on top of a computer monitor in an inexpensive desktop system to a High Definition camera that has remote control pan, tilt and zoom features in a conference room-sized system. In addition to the primary conferencing camera optional secondary, document, and specialized camera are available. HD cameras are preferred and offer both the largest images and highest resolutions.

- Microphone — (or conference phone): many of the desktop conferencing units come with small microphones fitted for one or two people. On room-based systems a specific conference phones are used which will work with integrated software to enhance the audio capabilities of the system. In larger venues separate audio echo-cancellation systems[1] with numerous microphones can be connected to the system to help with larger group interaction.
- Video Display — (e.g., plasma screen, LCD, or projector): these display devices are options that can be used to show the images received from the videoconferencing codec. Videoconferencing systems can use multiple displaying options. Desktop systems show the video in a small window on the computer monitor. Larger venue systems can have multiple display devices and present multiple endpoint locations and data simultaneously.
- Network Connection: this is the connection that carries data between video systems communicating with one another. The size of the connection and the ability to access it in a consistent manner, determines both video performance and quality of service. Most HD videoconferencing systems require a high symmetric bandwidth.
- Room: this is the environment in which the VC system is being used. It relates to specific settings in a room, e.g., the lighting, acoustics, decor and furniture which all have influence on the perceived quality of a VC solution. In some situations, the room is an integral part of the VC solution in which the room settings are fully prescribed.

In case of a multipoint videoconferencing, a *Multipoint Control Unit* (MCU) is also required. This is a device commonly used to bridge videoconferencing connections. The MCU is an endpoint (typically on the LAN) that provides the capability for 3 or more sites to participate in a multipoint conference.

Videoconferencing systems that users control to make the video calls are generally called '*Videoconferencing endpoints*'. This is important to note because an endpoint is where physical and logical connections to audio systems such as speakers and microphones are made. Additionally, a videoconferencing endpoint must be connected to a display in order for video conference users to see and hear one another.

Endpoints are the 'end of the line' for the videoconferencing connection and this is quite different from other components of a videoconferencing network such as MCU's which are in the middle of videoconferencing connections.

[1] Acoustic echo cancellers apply algorithms which are able to detect when sounds or utterances reenter the audio input of the videoconferencing codec, which came from the audio output of the same system, after some time delay

Findings

Findings

This chapter describes the main findings during the baseline assessment (Step 2) of the project. These findings are based on documentation provided by SCIC, a significant number of interviews, and workshops conducted by Gartner with key stakeholders (e.g., interpreters, meeting organizers, support staff, management of different DG's, and conference participants).

Users

It is important to realize that "users" of VC systems are not limited to interpreters alone, but also needs to include the actual participants (in all locations) of the meeting. They in fact have to really use the VC solutions. Based on a series of interviews, one of the main findings was the participants thus far were not involved in the (requirements) process for VC with simultaneous interpretation at all. It's also important to note that the ambition level of potential participants of videoconferences was actually far lower than what was originally assumed. For instance, many videoconferencing meetings are currently conducted with a relatively low quality (e.g., video-, audio quality), without participants complaining about it. Additionally, most interviewees had no clear view on how a much higher quality videoconferencing solution would improve their current way of working (this might be related to a lack of awareness regarding the capabilities of current videoconferencing solutions).

Design Principles

A basic best practice for selecting a specific technology is discussing and agreeing with all stakeholders on the design principles which help to guideline the selection process. As such, and based on interviews with all stakeholders, Gartner defined an initial set of design principles. However, during the first steps of the project it became clear that the ambitions of the project's steering committee (mostly related to potential simultaneous interpretation services offered), where not fully aligned with the ideas of the interpreters. This made it difficult to agree on a full and adequate set of design principles. However, some of the design principles (partly adapted) were still agreed on by both the steering committee and interpreters, and as such Gartner used these in the succeeding steps of the project frequently:

- SCIC interpreters are in the same location (conference room);
- SCIC interpreters are in the same location where most of the delegates are;
- SCIC interpreters (preferably) are in the same location as the meeting chairman.

Although this is a very limited set of design principles, it could be a starting point in further fine-tuning and agreeing a full set of design principles in Phase 1 of the proposed road map (see page 45). However, it is important that all stakeholders jointly define and agree on this final design principles. And because of the challenge to align all the different ambitions (and requirements) of the different stakeholders, it is fair to conclude that the main challenge with regards to simultaneous interpretation via videoconferencing solution is not so only a technical challenge, but above all a change management challenge.

Service Portfolio

SCIC currently has no service portfolio with regards to interpretation services offered to clients. Due to the lack of a service portfolio it is difficult to determine (and measure) which videoconferencing settings are already applied and demanded by participants.

SCIC Videoconferencing Solution

Gartner assessed SCIC as an organisation which has a vast knowledge of most aspects of end-to-end videoconferencing solutions. SCIC currently applies this internal knowledge for the internal development of a high quality VC solution (see page 15) to support simultaneous interpretation. This VC solution has some unique features not to be found in existing commercial solutions, such as the support of transporting multiple (interpreted) languages upstream and downstream simultaneously. The SCIC VC solution is currently being tested by interpreters. However, one of the main issues with this testing is the fact that interpreters are in the process of determining their general attitude toward using VC for conferences with simultaneous interpretation, as well as using remote interpretation. For them it is of utmost importance to maintain the delivered quality of services and their working conditions. SCIC Management is aware of these hesitations, hence the importance of a strong focus on change management.

Previous Studies

During the workshops and interviews, the stakeholders often referred to the technical requirements that have been formulated some years ago as a result from two previous studies. These studies are labelled as the ‘Benchmark study’ and the ‘Frauenhofer study’. Both these studies agree that *perceived quality*[1] of VC solutions is subjective, and that the suitability of videoconferencing for simultaneous interpretation cannot be captured in technical requirements alone. While Gartner fully agrees with this, Gartner also concluded that a number of the findings in these studies were translated into technical requirements which today are not fully applicable anymore. For example many interpreters refer to the requirement of having ‘at least CD quality audio’, meaning that they require the highest quality of sound possible. However, the CD is a medium which still applies compression techniques which are over 30 years old. As these techniques nowadays have been surpassed by a number of other, far more efficient and effective compression techniques, by applying this specific requirement too literally, these new techniques would potentially be left out. Therefore it will be important to update the requirements.

A very important finding is that Gartner found that, based on thorough market analysis, there is currently no single commercially available product on the market which meets all requirements ‘out of the box’. Analyzing the current capabilities of these solutions in fact implies that significant additional development and customization efforts (specifically with regards to the audio requirements) will be necessary.

[1] Perceived quality in the context of videoconferencing is a combination of image/sound quality, room environment (lights, acoustics), motion capture, picture size, lip sync and ease of use

Benchmark Results and Case Studies

Benchmark Results and Case Studies

This chapter describes the results of the benchmark evaluation and briefly describes two case studies.

Videoconferencing Market

It is evident that the market for videoconferencing systems has grown significantly over the past years and it is anticipated that this market will continue to grown over the next two years. While traditionally the Room systems segment was by far the largest segment, it is now overtaken by the so-called soft clients. This dedicated soft clients segment is currently growing at the fastest rate in comparison with other segments, and it will do so the coming years. Table 1 clearly indicates this trend.

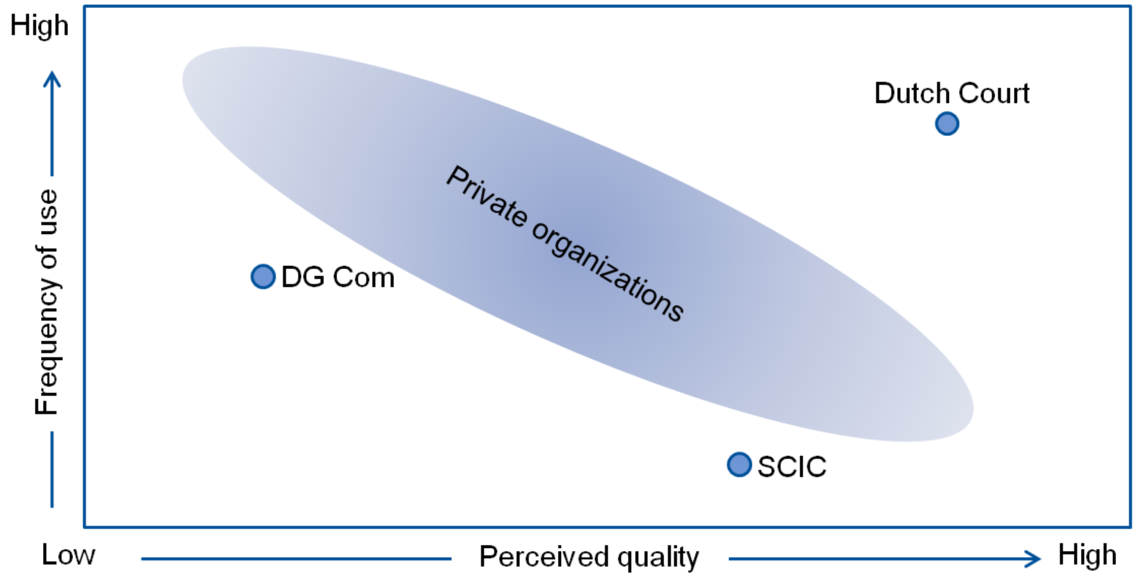
Table 1. Videoconferencing Market

Market	Value	2010	2011	2012	2013	2014	2015	CAGR 2010 – 2015
Video Endpoint Systems	**Units**	**238,251**	**290,562**	**340,350**	**389,194**	**437,478**	**483,707**	**15.2%**
Immersive	Units	2,001	2,571	3,321	4,424	5,327	6,299	25.8%
Room Systems	Units	191,337	229,604	246,045	295,730	325,303	351,327	12.9%
Exec and Personal Systems — Hardware	Units	44,913	58,387	72,984	89,040	106,848	126,081	22.9%
Dedicated Soft Clients — Incremental Licenses		422,222	738,889	1,256,110	2,072,582	3,233,228	4,817,510	62.7%

Benchmark Results

From the findings during the baseline analyses step of the project, it was concluded that SCIC developed a relatively high quality solution for video conferencing (although it is still being tested). This is also acknowledged in the benchmark evaluation. Important aspects of this evaluation include the perceived quality of the applied VC solutions technologies and the frequency of use. When comparing the SCIC VC solution to the VC solutions used by private organisations (for traditional videoconferencing, thus not applying videoconferencing in the context of simultaneous interpretation), it became clear that that the assessed (perceived) quality of the SCIC solution scores somewhat better than (average) perceived quality of the VC solutions of private organisations (as depicted in Figure 3).

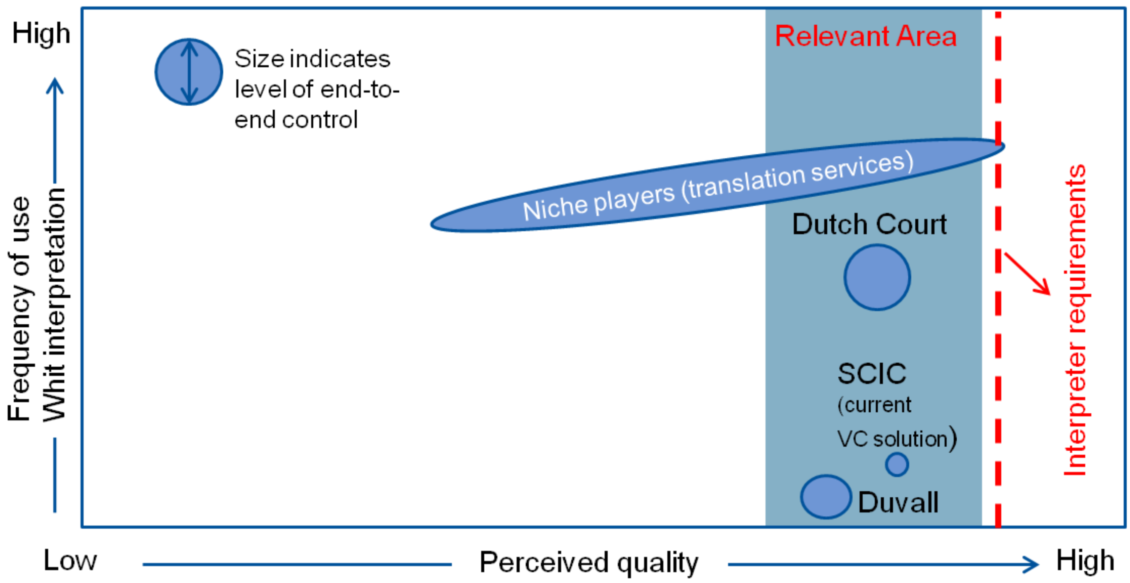
Figure 3. Benchmark Results — Videoconferencing Without Simultaneous Interpretation



The position of the SCIC VC solution on the benchmark further 'improves' when Gartner only looked at references where the VC solutions are used in the context of simultaneous interpretation (depicted in Figure 4 below). Gartner found that there are not many true reference points available (fully comparable to the specific situation of SCIC, that is: applying multiple languages during simultaneous interpretation sessions). In most of the included references, only one language needs to be interpreted, and in some of these references the interpreters did not apply simultaneous interpretation in all cases.

The benchmark sample includes niche players, such as U.S. interpretations. These are relatively small organisations, offering all kinds of interpretation services, including simultaneous interpretation. Such companies (many of them are mainly active in the United States) provide simultaneous interpretation services even using simple VC solutions such as Skype (but only for a single language). The U.S. Interpretation case is described in Case Studies.

Figure 4. Benchmark Results — Videoconferencing With Simultaneous Interpretation



Case Studies

As the benchmark evaluation results in a rather limited set of references with a comparable context, Gartner applied a case study approach to two of the data points in order to get a better understanding of these references. These case studies are described in this paragraph.

Dutch Court Room

The content of the following case was provided by of Peter van Rotterdam[1] and based on the presentation “Experiences with “True-to-Life” Videoconferencing in Courts.” Because of specific copyrights, Gartner only used relevant parts of the presentation and did not alter any of the content.

The Dutch court room applies a videoconferencing solution to their core process. In order to do this in an adequate manner, a number of important principles were applied:

- Each actor in the court room (judge, person to be heard, prosecutor, lawyer, interpreter, clerk, etc.) must feel comfortable to do his/her job. For good quality videoconferencing both VC Facilities must:
  - Have a high quality, uninterrupted connection between the two VC Facilities
  - Capture a good quality image and sound for providing this to the remote VC Facility (what they see is what you provide)
  - Display the image and sound exactly as received from the remote VC Facility (what you see is what they provide) It is about Data Connections and VC Technology … and also about Court Room Design (Lighting, Acoustics, Furniture etc.)

In this case the judge decides whether to use videoconferencing or not and the lawyer chooses his location: in court or next to client. In practice, the VC solution is mainly applied for high volume, low interaction cases such as the Criminal Court (extending preliminary detention). But videoconferencing is also applied to reduce high costs related to specific cases in the Criminal Court such cross-border in/outside EU cases.

Most of the rules for applying VC in the Dutch court room are defined by Dutch law. Most importantly the Dutch law states that the VC solution must provide a ‘True to life experience’:

- In order to assure the interests of each person involved (person to be heard, judge, lawyer and others) the quality of image and sound should be such that each person involved gets a realistic and clear view on what is happening in the other (remote) location
- The view should not be manipulated by e.g., zooming in or out and/or focusing more or less on certain characteristics of a person
- Image and sound should be synchronized
- Looks, facial expression, mouth movements, direction of view, gestures, posture of each person involved should be clearly perceptible
- Other people in the other (remote) location should be visible
- Interactions between people (how they react, when they look at and/or speak to each other) should be clearly noticeable
- Sound should be real-time, audible and allow simultaneous speaking

---

[1] Peter J. A. van Rotterdam is an independent management consultant based in the Netherlands, specialized in VC in Legal Procedures and Transnational VC in Judicial Procedures

- Communication should be possible based on sound only (e.g., while looking through files)
- Objectivity: each person should be represented in the same way to each other person and each person should have the same perception of eye-contact with all other people

The current situation with regards to videoconferencing in the Netherlands can be summarised as follows:

- 15 courts are currently applying video conferencing
- Cases in which VC is applied:
  - 1000 alien detention cases (less growth)
  - 100 extending prelim. detention (pilot)
  - 25/year rogatory committees (since 2007)

Figure 5. Dutch Court Room — Videoconferencing Applied in the Distant Site

Videoconferencing in the Context of Simultaneous Interpretation

Interpretation is frequently required during hearings. The interpreters are then using both simultaneous interpretation and consecutive interpretation, depending on which person is talking. The interpreters use the same videoconferencing system as the participants; no additional modifications have been made. When designing and implementing the VC solution it was assumed that if a Judge would agree with the sound and image quality, it should also be sufficient for the interpreters.

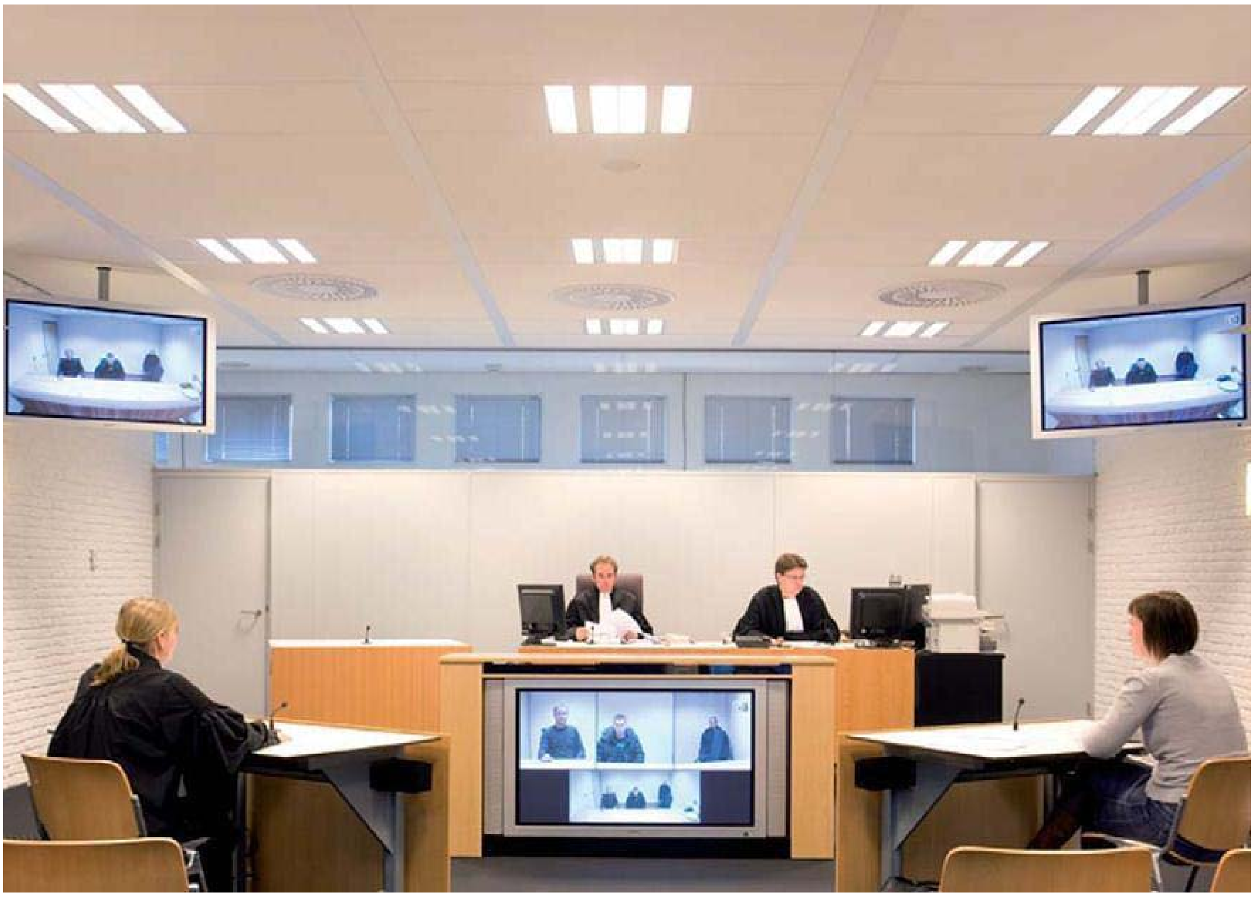
Lessons Learned

The most important finding from this case study is that consequently applying the ‘True-to-Life’ criterion really proved to be the key success factor. When this criterion would not have been applied during the hearings there would have been a lack of interaction and participants will find it difficult to:

- Interact with one or more people in your room or in the other room
- Make eye contact to one or more people
- Notice being spoken to, looked at, pointed at or addressed to
- Determine the direction from where a sound comes
- See how each person comports him/herself
- See how people react to one another, to whom they are looking/speaking
- See what is on the desk of each person and what they do with their hands
- See each person’s external appearance, facial expressions, gestures, posture
- See each person’s lip movements and direction of gaze

Often the true to live criterion during videoconferencing is not achieved because of lacking image and sound quality. Next to poor audio and image quality, the true to live experience can also be impacted by an unusual look and feel of the room. For example positioning/appearance of equipment floor, walls, ceiling or furniture makes participants feel not being in a court room. Or lights, acoustics, background noise makes you feel not being in court room. In this case it is important to state that perceived positioning of or distances to other people on your screens differ from actual positioning and distances as if they were in face-to-face hearing.

Figure 6. Dutch Court Room — Videoconferencing Applied in the Court Room



US Translation Company

The content of the following case was partially obtained via the website "UStranslation.com." Because of specific copyrights, Gartner did not alter any of the content unless otherwise indicated.

The U.S. interpretation company represents a group of interpretations services companies that operate on a low cost bases and provide remote interpretation though video conferencing.

About U.S. Translation Company (USTC)

In 1998 USTC was awarded a contract to provide interpretation services with one of the world's biggest suppliers of nutritional supplements, Nu Skin. The job was to provide professional interpretation services for an event that brings in tens of thousands of attendees annually, many from different countries around the world. Focus on quality and processes were recognized by third parties.

2009 proved to be a breakout year for the company as it doubled its growth amid economic recession. Helping businesses in the U.S. to reach foreign markets through professional-grade translation and interpretation for product information, instruction manuals, equipment, conferences, conventions and events, has allowed companies facing sliding domestic sales launch products and services in new economic arenas.

Remote Interpretation Services

USTC scour professional associations and organisations searching for interpretation services experts that have completed secondary education and training in their language pairs and provide impeccable references with a minimum of five years professional, paid experience. Today's fast-paced business climate requires on-the-fly solutions. That's one of the reasons video/Web conferencing has become a time-saving, inexpensive alternative for meeting with clients, shareholders, or regional management. In this context, U.S. translation offers simultaneous interpretation through videoconferencing solution with the following benefits:

- **Cost-Effective** — Travel time, per diem, and equipment costs are eliminated.
- **Timely** — from initial request to project completion, video/Web conference interpreting can be accomplished in a fraction of the time traditional conferences require.
- **Accurate** — Third Party Certification of our services, linguists and processes guarantee interpretation accuracy and observance of cultural nuances, eliminating embarrassing faux pas and/or miscommunication.
- **Professional** — Not only in linguist training and educational background, but also appearance and demeanour, ensuring that your company's image and reputation are well-represented.
- With over 2,000 consecutive interpretation experts in a vast array of subject fields including medical, business, legal, technical, healthcare, military, creative marketing, cosmetics, mining, software localization, and so much more — USTC are able to provide not only experts in the language but also interpreters that understand a certain field of expertise.

USTC Approach[1]

In their approach USTC is very flexible. They prefer to sit next to the client to assure a high quality (which is mostly the case), but USTC employees can also participate remotely in videoconferences. In principle they claim that there are no requirement regarding minimum standard for the VC technology use, although they emphasize that specifically the quality of the audio should be good. USTC currently even considers using Skype (in combination with standard telephony) to allow for (simultaneous) interpretation of multiple languages.

[1] *Based on an interview between Gartner and Jentry Reinhard (employee of US translation)*

Videoconferencing Scenarios

Videoconferencing Scenarios

Best practice for determining the best-fit VC solution(s) is to start with identifying the main use cases. The next step then analyzes for each of these use case what the most viable VC endpoints are for each of the sites where the videoconferencing solution will be used. For this study, this analysis is based on the requirements of the interpreters with regards to simultaneous interpretation. Following from the set of a-priori viable videoconferencing endpoints, it is determined what the preferred end-to-end VC solutions (combinations of VC-endpoints) are. Each of these steps is described in this chapter.

Use Cases

Use cases in the context of this study define in what setting the main ‘actors’ will use the videoconferencing system. Main actors here are the interpreters, and the participants.

The main use cases refer to the most commonly expected *event*s (meeting or conference) where SCIC provides simultaneous interpretation services to its clients.

Assumptions

A number of assumptions have be applied to formulate the different use cases:

- An event is a meeting or conference where one or multiple participant(s) join(s) this meeting or conference via a videoconference solution.
- An event always take place in a ‘conference room’
  - In this context, a conference room is a room which is able to accommodate at least 10 delegates, and has dedicated interpreter booths. It is assumed smaller sized rooms (or rooms without booths) are out of scope.
- For all use cases, the interpreters are located in the booths in the conference room, and not in a distant site
  - This conference room is referred to as the ‘main site’
- The total number of distant sites is maximized (at four)
  - Currently, most videoconferencing calls with multiple distant sites/participants, involve a display screen being divided into four quadrants, with additional distant sites beyond the original four being layered into the display quadrants as they become active. As such, these four quadrants can be seen as the natural maximum number of distant sites caused by current technical limitations. However, it is to be expected that even without technical limitations the maximum number of distant sites involved in a videoconferencing call will still be four, as from Gartner experience a higher number of distant sites/participants simply will not be manageable anymore.

With regards to the simultaneous interpretation services, and taking into account the above assumptions, three main use cases are distinguished that need to be supported with a videoconferencing solution.

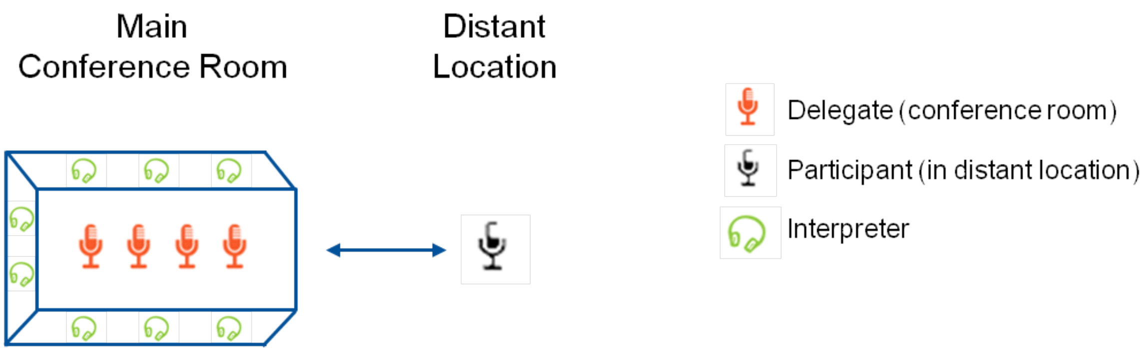
- Use Case 1 — One distant location with one participant
- Use Case 2 — One distant location with multiple participants
- Use Case 3 — Multiple distant locations, one (or multiple) participants per distant location

These use cases are described in more detail below. Of course there other use cases that could be determined, however these have not been taken into consideration for this study at present.

Use Case 1 — One Distant Location with One Participant

Use Case 1 refers to an event that takes place in a conference room, where one participant in a distant location joins the event via a videoconference solution, and the interpreters and their delegates are sitting in the conference room. It is likely that this also is the situation that most often will occur. An example of this could be when a single representative of a member state cannot travel to the main conference, but still wants to give his input. This use case typically requires a point to point VC solution.

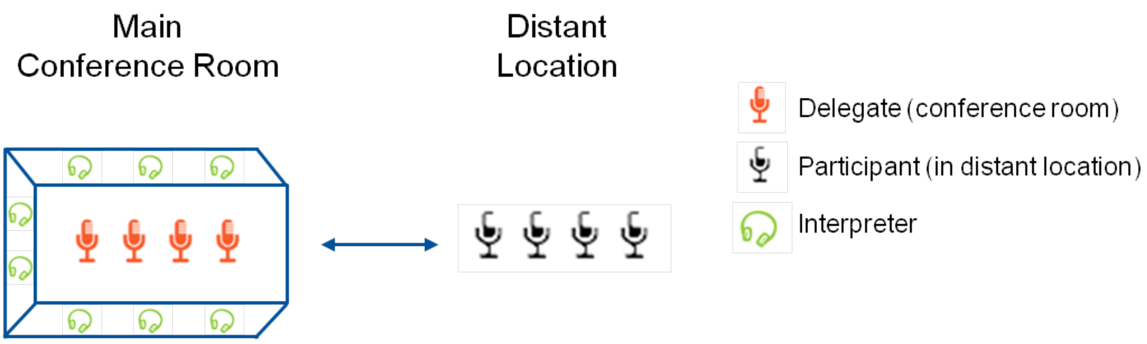
Figure 7. Use Case 1



Use Case 2 — One Distant Location with Multiple Participants

Use Case 2 refers to an event that takes place in a conference room, where *multiple* participants in one distant location join the event via a videoconference solution. The interpreters and their delegates are sitting in the conference room. An example of this could be when a delegation of one member state cannot travel to the main conference, but still wants to give their input. This use case typically requires a point to point VC solution.

Figure 8. Use Case 2



Use Case 3 — Multiple Distant Locations, One (Or Multiple) Participants per Distant Location

Use Case 3 refers to an event that takes place in a conference room, where one (or multiple) participants in multiple distant locations join the event via a videoconference solution. The interpreters and their delegates are sitting in the conference room. An example of this could be when multiple representatives from multiple member states cannot travel to the main conference (for instance due to a natural disaster or a general strike), but still wants to give their input to the meeting- or conference participants. This use case typically requires a multipoint VC solution.

Figure 9. Use Case 3

Videoconferencing Endpoints

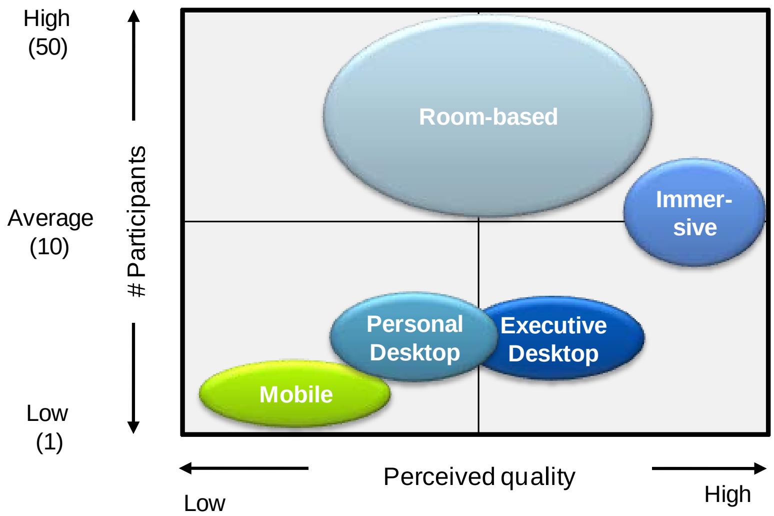
As ‘videoconferencing endpoints’ are the only part of the videoconferencing solution that the user sees, it is best practice to directly align the endpoints with use cases. This ensures the right videoconferencing endpoints in the right locations and that expectations can be met.

Videoconferencing endpoints come in many forms, intended to suit a range of use cases and price points, as well as network connectivity. Due to technical advancements, the differences between the different VC endpoints tend to blur over time. However, five main segments of videoconferencing endpoints can be distinguished:

- Immersive VC
- Room-based VC
- Executive desktop VC
- Personal desktop VC
- Mobile VC

These different segments are described in more detail below. The segmentation is mainly based on (1) the number of participants that typically can use such an endpoint simultaneously and (2) the general ‘perceived quality’ of this specific endpoint as depicted in Figure 10 below.

Figure 10. Videoconferencing Endpoint Segmentation



It is important to note that in this context perceived quality is a combination of image/sound quality, room environment (lights, acoustics), motion capture, picture size, lip sync and ease of use of a videoconferencing endpoint. It is a result of several related factors, including video size, frame rate, microphone/camera quality, the room, and network. In real life, the network performance is usually having the greatest influence.

Immersive

Immersive systems are at the high end of the VC market, typically featuring multiple codecs, multiple display HD screens for video and content sharing, and directional high-quality microphones. Immersive VC systems are tailored environments in which the lighting, acoustics, decor and furniture on both sides of the video are tightly controlled (and often identical), giving the participants the appearance of being in the same room as participants in other locations (hence the 'immersive' description). Such immersive systems often cost over €150,000 per endpoint.

Figure 11. Example of an Immersive Videoconferencing Endpoint

In order to limit these costs somewhat, several vendors also offer 'adaptive' immersive systems, often using the same underlying hardware as immersive systems, but delivered with more limited environmental components, such as furniture, lighting and wall panels.

It is important to note that immersive systems make very high demands on the network, with high-quality, three-screen HD video and collaboration consuming as much as 20 Mbps of dedicated bandwidth. Therefore, immersive systems are almost exclusively run over dedicated Internet Protocol (IP) networks to ensure the highest possible quality (the private IP implementations far outpace Internet-based connectivity).

Immersive videoconferencing systems are mostly used for small group sessions, with the typical number of 6-8 users.

Room-based systems are regarded as 'traditional' videoconferencing systems, and come in a wide variety of formats and price points. The most common room-based videoconferencing endpoints are self-contained room-based systems, which have a television set-top form factor. Typically, these endpoints have been used in shared (multipurpose) rooms, which have to be prescheduled for use.

Figure 12. Example of a Room-based Videoconferencing Endpoint

As display technology has evolved, room-based systems have become more modular, and have added support for more-flexible configurations of displays (e.g., projectors for display on large screens), microphones and cameras. The quality of room-based systems has currently moved to HD-screens and much improved audio quality. The room-based videoconferencing systems are often referred to as appliances, since they do not require a separate desktop.

Room-based videoconferencing systems are used for both small and large group sessions, with the typical number of users ranging from 10 to 50.

Executive Desktop

Executive desktop videoconferencing systems offer a stand-alone, generally (though not always) higher quality experience than a Personal desktop (see next paragraph). Executive desktops are often sold as flexible (movable) solutions; however in practice these systems are not really that portable. Due to the relatively high cost, the segment of executive systems is more or less a niche market, mainly used by c-level managers of commercial organisations.

Figure 13. Example of an Executive Desktop Videoconferencing Endpoint

Executive desktops are mostly used for one-to-one user sessions.

Personal Desktop

The personal desktop category includes all-software videoconferencing solutions that run on laptop and desktop computers, typically with a webcam and sometimes with external microphone. Thanks to advances in hardware, personal desktops can currently run powerful codecs. The combination with specific software and the ability to connect almost any peripheral in the quality the user requires makes the solution rather customizable. Therefore Personal desktop videoconferencing solutions have become increasingly popular over time.

Figure 14. Example of a Personal Desktop Videoconferencing Endpoint

Personal Desktop videoconferencing is mostly used for one-to-one user sessions.

Mobile

Mobile videoconferencing uses handheld devices to conduct video meetings in the field. Typical mobile devices are smartphones and tablets. Popular examples in this segment are Skype and Facetime (Apple). Mobile videoconferencing allows for high flexibility, as attendees do not have to be in any particular location to participate in the call and the call does not need to be scheduled upfront. The main disadvantage of mobile devices is that the smaller form-factor will generally not lend themselves to two-way video communication, both from a screen real-estate perspective and delivered audio quality. And additionally, it is still very difficult to provide a stable video image from a handheld device.

Mobile videoconferencing solutions are almost exclusively used for one-to-one user sessions.

Endpoint Summary

Table 2 summarises the main aspects which should be used to compare the applicability of the different VC endpoints for a specific situation.

Table 2. Comparison of Different Videoconferencing Endpoints

	Immersive	Room-based	Executive Desktop	Personal Desktop	Mobile
Typical Usage	Average Group	Average Group/Large Group	Single User/Small Group	Single User/Small Group	Single User
Perceived Quality	Very High	Medium High	Medium High	Medium	Low
Typical Location	Dedicated Room	Shared Purpose Room	Office	Office/Home	Home/Field
Flexibility (Location Related)	Low	Low	Medium	High	Very High
Network Requirements	Very High	High	High	High	Medium
Installation Time	High	Medium High	Low	Low	Low
Relative Cost	Very High	High	Average	Low	Very Low

As products continue to proliferate, there is a high probability of increasing overlap between different VC segment further blurring the product distinction lines.

Endpoint Applicability per Use Case

As a third step, the applicability of 'VC endpoints' for both the main site (conference room) and the distant site(s) needs to be determined. The analysis per site per use case is described below. It should be noted that the analysis primarily is based on the technical capability of the specific endpoints with regards to the requirements of the interpreters[1], not taking into account criteria like e.g., cost, network requirements, and flexibility.

Interpreter Requirements

In the previous years, two studies have been conducted related to the interpreter requirements on behalf of DG Interpretation: a benchmark study[2] and the Fraunhofer study. While the latter study primarily focused on the way that the perceived quality of videoconferencing can be measured and captured, in the former study a number of general requirements for the videoconferencing solution with simultaneous interpretation where formulated. These general requirements (nr. 1 — 13) are listed in Table 3. Gartner combined these individual requirements into four consolidated groups of requirements, which were then used to determine the applicability of a specific VC endpoint.

Table 3. Interpreter Requirements

No.	General Requirements for Simultaneous Interpretation	Consolidated Requirements
1	Minimum frequency bandwidth at least 125 — 12.500 Hz	Requires high quality microphones, headphones, and optimized room conditions (acoustics) in all sites
2	Good acoustics, sound proof (no interference of other sound)	
3	No sound reverberation	
4	No overlapping sounds	

[1] The requirements of the participants of the videoconference (not being the interpreters) are not yet formulated.

[2] Benchmarking tests carried out by DG Interpretation between July 2005 and December 2007

No.	General Requirements for Simultaneous Interpretation	Consolidated Requirements
5	Minimum image resolution at least 1280x720	Requires high quality camera(s) and optimized room conditions (light) in all sites
6	The image quality needs to be of the highest possible standard	
7	The image quality needs to be stable (no drop outs or pixilation)	Requires sufficient bandwidth (guaranteed quality), active network management, and sophisticated codecs
8	Image and sound need to be synchronized	
9	Image and sound should not be delayed	
10	Sound needs to be full duplex, allowing sound from both locations to be transmitted at the same time without sound “cutting out”	
11	Interpreters preferably see all the participants at the distant location(s) at the same time	Requires specific camera settings, a number of screens, screen size (layout), sophisticated codecs and, in some cases, an MCU
12	Interpreters have a frontal view of the distant participant who is speaking (talking head)	
13	Interpreters are able to see the documents (when available) displayed simultaneously on the screen	

Main Site — Use Case 1

In Use Case 1 (one distant location with one participant), the group of interpreters are sitting in the main site require to have the front view of the participant in the distant site displayed in a high quality. The audio quality delivered to the main site needs to be of a very high level. This implies that for Use Case 1, the VC solution in the main site needs to include:

- Screen with a resolution of 1280x720 (or better)
  - Screen size needs to accommodate a larger group
  - When documentation needs to displayed, it is assumed a second screen is required
- Sophisticated codec, based on the H.264 standard (or better)
  - Depending on quality of the network, a coded based on the H.264 SVC[1] standard might be preferable

Based on the above, Gartner does not consider the Mobile, Personal desktop and Executive desktop as viable alternatives for the main site, as these VC endpoints, specifically from a

[1] Scalable Video Coding (SVC) an annex to the H.264 standard for video encoding that allows the encoder to build layers of enhanced data on top of a base layer video stream. The base layer can be read by any H.264-compatible decoder. The enhanced layers provide additional temporal information (i.e., intermediate frame data) or additional spatial information (i.e., higher resolution data). Each layer of data can be independently addressed, so that the encoding endpoint need only send out a single stream (consisting of base and enhanced layers) while the decoding endpoint receives only those enhanced layers pertinent to its situation (which may be limited by screen resolution, decoding frame rate or the overall bandwidth available to deliver the video stream). As such, SVC lends itself to the delivery of video over best-effort networks, such as the Internet, and to the delivery of video to multiple user formats (Room based systems, and Personal desktop systems) without the need for transcoding in an MCU. This can reduce both the cost of deployment and the end-to-end processing latency involved in delivering video streams between users.

screen size perspective, are intended to be applied by single users (i.e.) and not by groups or large rooms.

Immersive VC endpoints have limited viability for the main site, as key environmental components such as furniture and wall panels are not fit for purpose in a large conference room. In general, immersive systems typically are not fit to accommodate groups larger than 10 participants. Only 'adaptive' immersive systems have some viability, as they offer the same underlying hardware as immersive systems but not the furniture.

Room-based VC is the most viable segment for the main site in this use case, as these endpoints offer support for a large HD quality screen (via e.g., a projector). However, still there will be customization needed to deliver audio in the required quality.

Main Site — Use Case 2

In Use Case 2 (one distant location with multiple participants), the group of interpreters (and their delegates) sitting in the main site require to have a front view of the speaking participant and simultaneously an overview of all the participants in the distant site, each in high quality. The audio quality delivered to the main site needs to be of a very high level. This implies that for Use Case 2, the VC solution in the main site needs to include the same elements as in Use Case 1, plus:

- One additional screen (similar as for Use Case 1) to show an overview of all participants at the distant location
  - When documentation needs to displayed, it is assumed a third screen is required
- One additional sophisticated codec (similar as for Use Case 1) to display to the overview of the participants at the distant location simultaneously on the additional screen

Based on the above, Gartner again does not consider the Mobile, Personal desktop, Executive desktop and as viable alternatives for the main site in Use Case 2.

Like for Use Case 1, immersive VC endpoints also have limited viability for the main site in Use Case 2, because of exactly the same arguments.

Similarly, Gartner considers room-based VC is the most viable segment for the main site in this use case, as these endpoints offer support for multiple large HD quality screens (via e.g., a projector). However, besides customization needed to deliver audio in the required quality, the solution will be somewhat more complex due to the need of two codecs instead of one.

Main Site — Use Case 3

In Use Case 3 (multiple distant locations with multiple participants), the group of interpreters (and their delegates) sitting in the main site require to have a front view of the speaking participant and simultaneously an overview of all the participants in all distant sites, each in high quality. The audio quality delivered to the main site needs to be of a very high level. This implies that for Use Case 2, the VC solution in the main site needs to include the same as in Use Case 1, plus:

- Multiple additional screens (similar as for Use Case 1) to show an overview of all participants at the distant locations, depending on the number of sites.
- Multiple additional high quality codecs (similar as for Use Case 1) to display to the overview of the participants at the distant location simultaneously on the additional screens, depending on the number of sites
- MCU to bridge all distant site connections to the main site.

As formulated in the assumptions for the use cases, the natural maximum number of distant sites is four. Based on the above, Gartner again does not consider the Mobile, Personal desktop, Executive desktop and Immersive VC endpoints as viable alternatives for the main site in Use Case 3.

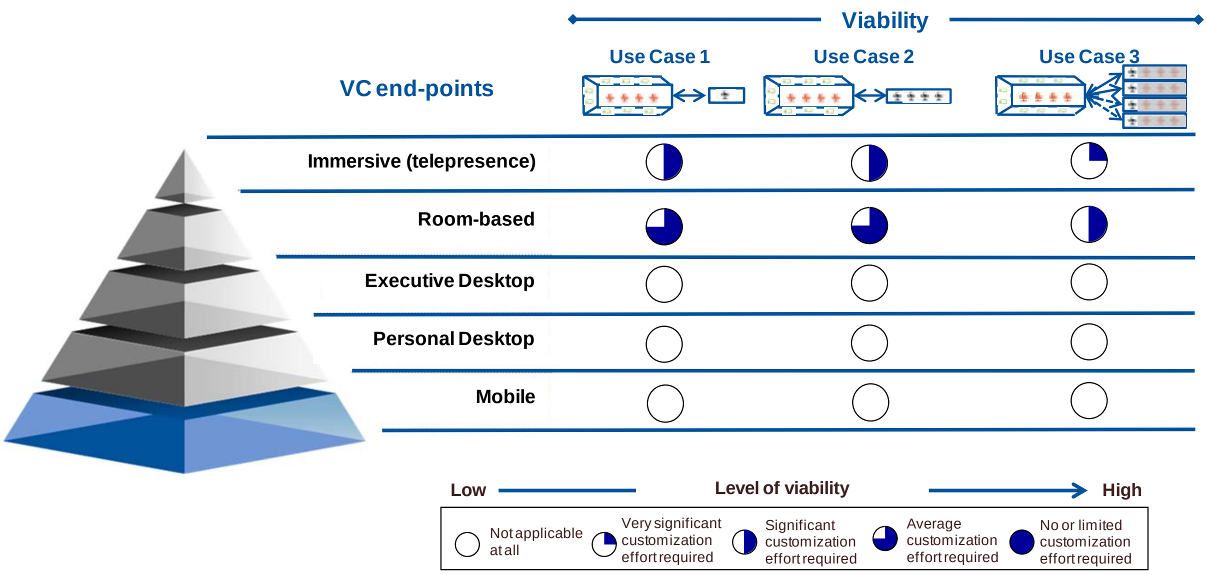
Gartner considers room-based VC as the most viable segment for the main site in this use case, as these endpoints offer support for multiple large HD quality screens (via e.g., a projector). However, besides customization needed to deliver audio in the required quality, and the complexity due to the need of two codecs instead of one in this situation, it should be noted that adding an MCU will increase the complexity of the VC solution significantly. Not only does an MCU increase the latency in the end-to-end path significantly, it also increased the complexity of actually managing the conferencing. What also adds to the complexity of multipoint VC is the manageability of the conference (e.g., to manage which participant can say something when), which increases significantly when the number of active speakers in the distant sites increases. The total number of distant sites/participants therefore is maximized at four.

Main Site — Summary

As depicted in Figure 15 Gartner only considers VC endpoints in the Room-based segment as viable for the different use cases. However, in each case significant customization will be required, specifically with regards to the sound.

All other VC endpoints are not (or only very limited) viable for the main site.

Figure 15. Videoconferencing Endpoint Viability Per Use Case For The Main Site



Distant Site — Use Case 1

In Use Case 1 (one distant location with one participant), the group of interpreters (and their delegates) sitting in the main site require to have a front view of the participant in the distant site in a high quality. The audio quality delivered to the main site needs to be of a very high level. This implies that for Use Case 1, the VC solution in the distant site needs to include:

- High quality camera
- High quality microphone
- Optimized room conditions (acoustics and light)
- Sophisticated codec

- In order to avoid transcoding (which introduces latency and potentially lip-sync issues), this codec ideally should be the based on the same as the one used in the main site

Based on the above, Gartner does not consider the Mobile VC segment as viable alternative for the distant site because the quality of the internal microphone and camera, as well as the internal codec which is typically sub-par.

The Personal desktop and can potentially be a viable alternative for the distant site. Not only because the current hardware can run high quality codecs, but these solutions also allow for being equipped with high quality peripherals, which can deliver the sound and image in the required quality. However, the Personal desktop should be used in a room with optimized conditions, which will limit the flexibility advantage of the solution somewhat.

The Executive desktop, like the Personal desktop, can potentially be a viable alternative as endpoint in the distant site. The out-of-the-box 'perceived quality' it delivers in general is higher than the Personal desktop. However, unlike the Personal desktop, the Executive desktop mostly does not allow for customization of the peripherals (most importantly the microphone and camera). The specifications of these peripherals are determined by the vendors, and although the quality in general is rather high, they are most likely not of the sound and image quality required for simultaneous interpretation. Additionally, the Executive desktop should be used in a room with optimized conditions, which will limit the flexibility of the solution as well.

Room-based VC is a viable endpoint in the distant site for this specific use case. Such an endpoint also allows for being equipped with high quality peripherals, and therefore it can deliver the required quality with regards to image and sound. However, the room itself, which is mostly likely a shared purpose room, will have to be optimized (acoustics and lights) in order for the sound and image to be captured in the required level.

Immersive VC endpoints are also viable endpoints for the distant site in this use case, as they are already equipped with high quality peripherals and incorporate a dedicated and optimized room. Still the peripherals of an immersive endpoint have to be tested in order to validate if they indeed meet the requirements for performing simultaneous interpretation.

Distant Site — Use Case 2

In Use Case 2 (one distant location with multiple participants), the group of interpreters (and their delegates) sitting in the main site require to have a front view of the speaking participant and simultaneously an overview of all the participants in the distant site, each in high quality. The audio quality delivered to the main site needs to be of a very high level. This implies that for Use Case 2, the VC solution in the distant site needs to include the same as in Use Case 1, plus:

- An additional high quality wide angle camera to capture an overview of all participants
- Microphone(s) that support multiple speakers
- An additional sophisticated codec
  - Similar as in Use Case 1, however this additional codec is required to support the additional wide angle camera

Based on the above, Gartner does not consider the Mobile, Personal desktop, Executive desktop and Immersive VC endpoints as viable alternatives for the distant site in Use Case 2, as these VC endpoints, specifically from a screen size perspective, are intended to be applied by single users (i.e.) and not by groups or large rooms.

Room-based VC is a viable endpoint for the distant site in this use case. Such an endpoint also allows for being equipped with high quality peripherals, and therefore it can deliver the required quality with regards to image and sound. However, the room itself, which is mostly likely a shared purpose room, will have to be optimized (acoustics and lights) in order for the sound and image to be captured in the required level.

Immersive VC endpoints are viable endpoints for the distant site in this use case, as they are already equipped with high quality peripherals, including multi-codecs, and incorporate a dedicated and optimized room. Still the peripherals of an immersive endpoint have to be tested in order to validate if they indeed meet the requirements for performing simultaneous interpretation, and it is most likely that modifications have to be made to assure the required audio quality.

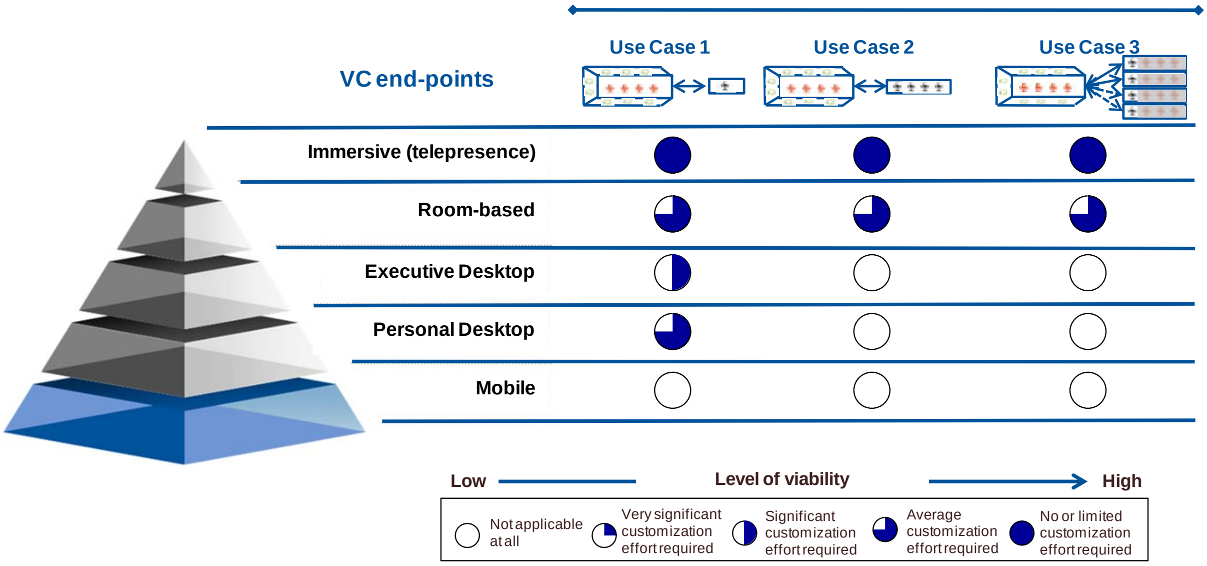
Distant Site — Use Case 3

In Use Case 3 (multiple distant location with multiple participants), the group of interpreters (and their delegates) sitting in the main site require to have a front view of the speaking participant and simultaneously an overview of all the participants in all distant sites, each in high quality. The audio quality delivered to the main site needs to be of a very high level. This implies that for Use Case 3, the VC solution in the distant site needs to include the same as in Use Case 2. Therefore the viability of the different VC endpoints for the distant sites in Use Case 3 is exactly the same as in Use Case 2

Main site — Summary

As depicted in Figure 16, Gartner only considers specifically the Room-based and Immersive solutions as viable endpoints for the distant sites in all the use cases. The Personal desktop, and (to a lesser degree) the Executive desktop are regarded as viable endpoint for the distant site in Use Case 1 only. Mobile endpoints are not viable as endpoint for the distant site. However, it is to be expected that Use Case 1 will be the most important use case for SCIC, and therefore should carry the most weight in the analysis of viable end-to-end solutions. It should also be noted that for each solution customization will be required, specifically with regards to the audio-related aspects.

Figure 16. Videoconferencing Endpoint Viability Per Use Case For The Distant Site(S)



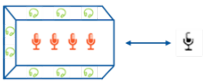
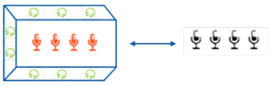
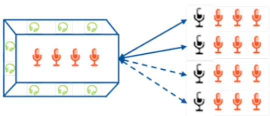
Viable End to-End Videoconferencing Solution(s)

Strictly based on the requirements of the interpreters, three combinations of endpoints for the main sites and distant sites are possible. For new situations however (sites without an existing VC endpoint), Gartner considers the combination Room-based to Immersive to be less viable. Implementing Immersive endpoints in a significant number of sites is extremely costly, and additionally, it will require a very high amount of bandwidth, while the true added value of Immersive systems will not be delivered to the users when the endpoint in the main site is not Immersive. Additionally, unlike the Room-based and Personal Desktop solution where the room can be multipurpose, the room for an Immersive system is always dedicated. While this not only makes the end-to-end solution highly inflexible, it is also questionable from a real estate perspective if such dedicated rooms can be made available in all potential distant sites in the EU. Therefore Gartner only sees the Room-based to Personal desktop and the Room-based to Room-based solutions as potentially viable.

While the Personal desktop is not suited as endpoint for the distant site in Use Case 2 and Use Case 3, as the Personal Desktop does not support groups of participant, it is the preferred solution for Use Case 1, as both cost, flexibility, and bandwidth requirements are favourable compared to the Room-based endpoint for the distant site. This is summarised in Figure 17 below.

Figure 17. Viability of End-to-End Videoconferencing Solutions Per Use Case

Viable Solution

Use Case	Room-based (Main Site)— Immersive (Distant Site)	Room-based (Main Site)— Room-based (Distant Site)	Room-based (Main Site)— Personal Desktop (Distant Site)
1 	■ Limited flexibility ■ Very high cost ■ High bandwidth required	■ Limited flexibility ■ High cost ■ Average bandwidth required ■ Room to be optimized	■ High flexibility ■ Low cost ■ Average bandwidth required ■ Room to be optimized
2 	■ Limited flexibility ■ Very high cost ■ Very high bandwidth required	■ Limited flexibility ■ High cost ■ High bandwidth required ■ Room to be optimized	■ Not suitable for larger groups
3 	■ Limited flexibility ■ Very high cost ■ Very high bandwidth required	■ Limited flexibility ■ High cost ■ Very high bandwidth required ■ Room to be optimized	■ Not suitable for larger groups

Gartner

From “AAA Solution” to “A Solution”

From "AAA Solution" to "A Solution"

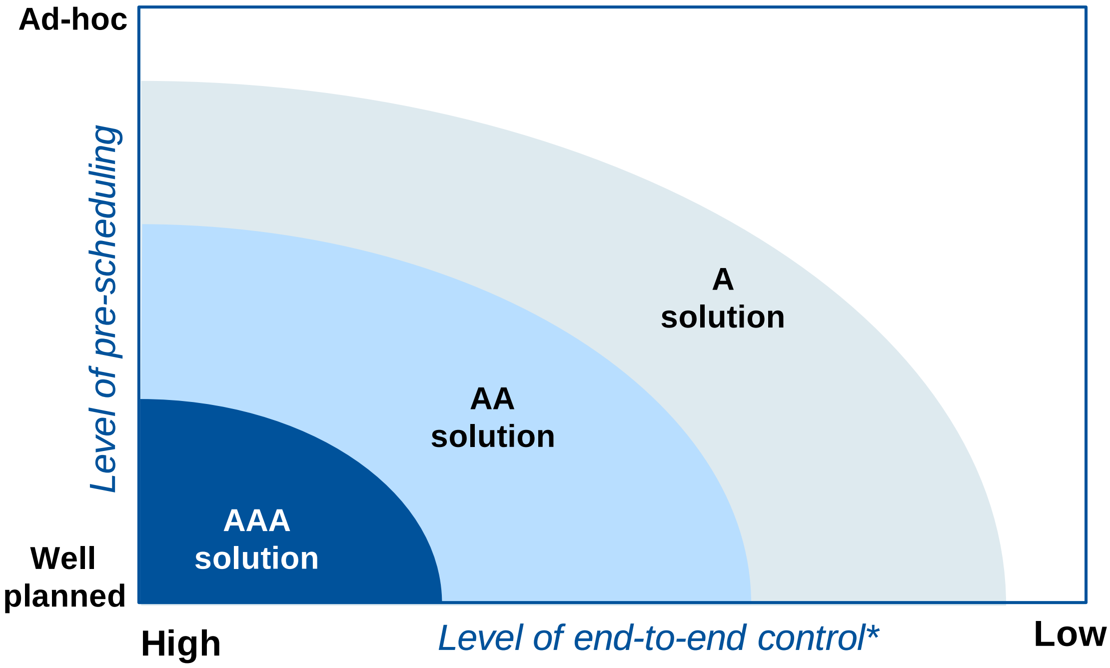
In the previous chapter it was concluded that two end-to-end VC solutions are viable for simultaneous interpretation, namely the Room-based to Personal desktop solution for Use Case 1, and the Room-based to Room-based solutions for Use Case 2 and Use Case 3.

In order for SCIC to provide an 'AAA solution'[1] for VC with simultaneous interpretation, that is a consistent and repeatable quality of experience which meets the user (and specifically the interpreters) requirements and expectations, two aspects are very important: Level of end-to-end control and the level of pre-scheduling.

- Level of end-to-end control refers to the level of control and support over the interconnecting network (bandwidth, quality), room conditions (lights, acoustics), and standards of VC endpoints used in both main sites and distant sites
- Level of pre-scheduling refers to the time available for VC management & support to test, optimize and validate the end-to-end VC solution between all the different sites involved in the specific meeting or conference

As depicted in Figure 18, this 'AAA' solution for a meeting can (only) be provided when the level of end-to-end control is high (meaning full control) *and* the meeting is scheduled well upfront (meaning planned according to SCIC standards)

Figure 18. Preconditions for the "AAA Solution"



In other situations, that is when the level of end-to-end control is not high and/or the meeting is not well planned in advance (allowing e.g., for pretesting), it is to be expected that the level of perceived quality of the end-to-end VC solution will drop significantly (depicted by the 'AA' and 'A' solutions). In these situations, the perceived quality of the end-to-end VC solution will most likely not be sufficient for the interpreters.

[1] The 'AAA solution' refers to situation where SCIC provides a VC solution with the perceived quality as required by the interpreters for simultaneous interpretation.

The next paragraphs describe what the specific preconditions are for providing the 'AAA solution', related to the connecting networks, room conditions, standardization and management & support.

Network

Network Requirements

Video calling differs significantly from telephony, when considering the per-call bandwidth as a percentage of the available network capacity. While Internet Protocol (IP) telephony can require only 30 Kbps to 70 Kbps to sustain a conversation, a high-quality video call can require 10-20 times that amount or more. For example, the per-call bandwidth suggested for a video call at a resolution of 720p at 30fps requires about 1Mbps (with aggressive compression) or 1.5Mbps (with better compression) with the use of a new generation H.264 codec. For a resolution of 1080p at 30fps this increases to 1.5-2Mbps, and 60fps will even require 4Mbps. And, this is per screen/codec, so a multi-screen endpoint will require this per screen.

This per-call bandwidth means that while an incremental voice call can usually be squeezed over a branch office WAN (wide-area network) connection, an incremental video call requires nearly one-third the capacity of a branch T1, capacity that is less likely to be available at the time the call is placed.

In addition to requiring less bandwidth, voice calls are more resilient, because there is less perceived impact on the quality of experience by the user when packets are dropped. Video coding is far more sophisticated than voice coding, and packets dropped in the wrong place can create objectionable video artefacts like frozen video.

Network Options

Because of the specific (and rather high) network requirements, selecting the right network for videoconferencing is essential for meeting the video call quality expectations, while also being able to deliver a repeatable experience that is consistent with the capabilities of the participating endpoints. Options for videoconferencing networking include ISDN, Internet, fully converged transport, and dedicated video "overlay" networks, all of which have different attributes. The factors that drive network selection include the volume of video calling, the experience of the organisation in managing real-time traffic, the desired quality of experience and the capabilities of the current WAN provider. Taking these into consideration, network selection for videoconferencing shows the following options:

- **ISDN** — this approach is least preferable due to the limitations in bandwidth and the challenges managing end-to-end call quality, especially for international calls. However, ISDN is still a preferred approach for organisations that must adopt predefined ISDN connectivity as part of a supply chain relationship. ISDN is also still preferable for room-focused enterprises that have very low calling volume (fewer than two hours per week per endpoint) associated with installed ISDN devices. Video connectivity using ISDN is also preferred for organisations that call extensively outside the enterprise to a range of different destinations like public rooms. Choosing an ISDN network means that a distinct Internet Protocol (IP) connection will be required for endpoint management, and that the organisation will be limited from practically scaling to meet HD bandwidth. Over 85% of ISDN video calling is at 384 kbps or 256 kbps.
- **Internet** — Using the Internet for videoconferencing connectivity can deliver a satisfactory experience for organisations that are without private network alternatives, but is not recommended as a primary network alternative. Most organisations have regional demilitarized-zone designs that introduce bandwidth chokepoints and additional latency that diminish an interactive video experience. Internet connectivity

can be effective if delivered directly to the site, but it still needs to be from the same ISP and delivered in-region for sufficient performance. Internet transport is preferred over ISDN for calls to IP endpoints outside the firewall, since it avoids transcoding and can scale more easily for high-speed calls.

- **Converged** — Combining videoconferencing traffic onto an existing IP network is preferable when the organisation has a limited number of WAN providers and has experience managing real-time traffic, especially voice over IP (VoIP). This approach is more common for organisations interested in managing their own videoconferencing infrastructures, especially if they have a high percentage of internal video calling. Transitioning to a converged network for videoconferencing is also preferred by organisations that anticipate growth in Personal Desktop videoconferencing, since the desktops are already clients on the internal network. Selecting a converged network for transport does not obviate the possibility of managed services, which can still be delivered by a virtual network operations centre (VNOC) provider over a management virtual private network (VPN).
- **Overlay Network** — a dedicated IP network that is not converged with the existing WAN effectively sidesteps convergence issues like multiple WAN providers, class-of-service administration and more-intricate capacity planning. The overlay approach can involve transport-only, where a videoconferencing IP network is built in parallel, but the organisation runs its own infrastructure. Alternatively, the overlay network may be a parallel connection with a service provider, offering both connectivity and access to value-added services like bridging, gateway and exchange. As a general rule, dedicated IP networks for video are most preferred when quality requirements are paramount (as in an immersive solution), or when organisations lack the experience in running real-time traffic over the WAN. Overlay networks can consist of private lines that concentrate a small number of systems into a central multipoint control unit (MCU), but are most likely to be Multiprotocol Label Switching (MPLS). An overlay approach is preferred if the organisation has limited experience managing real-time IP traffic, has a fragmented WAN architecture, has the need to build out videoconferencing capabilities quickly, or simply desires the highest possible video quality with the lowest amount of risk.

Taking into account the requirements of the interpreters, the possible end-to-end VC solutions and the bandwidth this requires, SCIC ideally should opt for an overlay network, with sufficient guaranteed bandwidth and reach to support the number of potential distant sites. However, if there is sufficient experience within SCIC (or the parties managing the networks) regarding the management of real-time traffic over IP networks, a fully converged network can also be viable.

When the cost of having private networks between all potential distant sites are too high, it should be tested under what conditions best effort networks like the Internet could also provide sufficient bandwidth and quality. In this case it is most likely that a VC solution with a coded based on the H.264 CVC standard will be required for this.

Table 4. Viability of Network Types for Videoconferencing

Network	Capability for Videoconferencing	Comments
Wireless, Satellite, ISDN	Capable in very limited cases	■ Does not provide sufficient bandwidth or quality
Internet	Capable in some cases	■ Only effective if delivered directly to the site and needs to be from the same ISP and delivered in-region for sufficient performance ■ Not recommended as primary network when organisation has regional demilitarized-zone designs
Fully Converged Network (Combining VC traffic with existing private IP networks)	Capable in the majority of cases	■ Dependent on quality and bandwidth of existing networks ■ Very complex ■ Requires experience in managing real time traffic
Overlay Network (VC traffic on separate IP networks)	Capable in almost all cases	■ Highly scalable ■ Very expensive ■ Overlay networks can be provided by third parties

Network Quality of Service

Fitting real-time traffic over a fixed pool of bandwidth has long been the role of Quality of Service (QoS) techniques. These techniques honour gradations of packet markings to prioritize delay-sensitive traffic over network connections. However, dense deployments of VC endpoints are like the proverbial ambulance on the crowded highway, that is, no amount of sirens and flashing lights can speed the ambulance to its destination. Further complicating the problem of relying on QoS is the situational-specific behaviour of video that is more nuanced than just completing a phone call. Video calling often has tiers of quality, ranging from simple video chat to immersive VC. As a result, some video calls are simply more important that others, with the highest-quality Room-based calls at the highest level of preference. This however does not imply that for instance Personal desktop endpoints are unimportant, especially when they are participating in calls with high-quality Room-based systems, or when the Personal desktop systems calling the Room-based systems belong to very important politicians. And, while QoS policy is a static design that gets incremental modifications as needed, the decisions that need to be made about video calling are dynamic. Therefore the organisation also needs the internal capabilities to constantly manage QoS at the actual time of the call.

Room Conditions

Immersive videoconferencing is sometimes termed as ‘videoconferencing with rules’ to capture the notion that, while the codecs and cameras in immersive solutions are sophisticated, the solutions also deliver *highly standardized lighting and room design*. Taking note of this is very important, specifically when taking into account the requirements of the interpreters regarding the perceived quality of the end-to-end VC solution. For instance, the quality of lighting in standard rooms is often insufficient for high-quality video capture. As immersive systems include side lighting, they can deliver a more consistent image, especially in conditions of low ambient light, which makes seeing the correct facial expressions easier and thus improves interpretation. If systems without integrated lighting are considered, it should be sure that adequate room lighting is available. Endpoint placements are also more likely to be impacted by sunlight from an adjacent window, which can play havoc with even the best cameras. Excessive ambient light should be addressed through endpoint placement and, where necessary, additional window treatments.

But based on the interviews Gartner conducted with them, even more important for the interpreters than the quality of the lightning is the quality of the audio. Perceived high audio quality is a prerequisite for high speech intelligibility and interpretation. Low audio quality leads to communication quality impairments. Human speech normally has a high degree of redundancy. This means that humans are able to work out the missing information even when the audio suffers from being of less than optimal quality. However, this is not always the case, as in some cases ambiguity due to low audio quality emerges: low audio quality may lead to misunderstandings and most importantly to wrong interpretations during a meeting. Lowered speech intelligibility may also cause more fatigue of the interpreters.

The main contributor to reduced audio quality is suboptimal acoustical conditions in the meeting room, typically when sound is reflected too much by the walls, ceiling and floor. This condition is perceived as having too much reverberation in the room.

It is not the acoustical conditions in terms of reverberation alone that may lead to reduced perceived audio quality. Noise sources present will lead to higher background noise, which in turn may mask the audio and disturb the interpretation. Traffic noise heard in the meeting room, due to poor sound insulation, can also be disturbing. Noisy heating ventilation and air conditioning installations may also make interpretation more difficult.

Based on the above, it is a precondition for delivering simultaneous interpretation services via VC, that the room’s conditions for both the main site and the distant sites are fully adequate.

Standardization on Codecs and Endpoints

Videoconferencing has long placed an emphasis on interoperability, based on the notion that the increased reach associated with more endpoints would drive a “network effect.” This assumption has proved flawed, with complexity and loss of quality far outweighing any benefits of additional reach. The decision to focus on IP or ISDN has helped simplify MCU planning and drive more consistent call quality. The modern videoconferencing environment, however, adds multivendor complexity. A growing number of video codec implementations are arriving to satisfy different requirements, and despite the seeming ‘standard’ of h.264 video coding, organisations now face over half-dozen implementations of h.264 that cannot natively interconnect, and a range of proprietary VC solutions that is optimized for closed deployments. Standardization on specific codecs and endpoints therefore is critical for interoperability in EU context. A standardized endpoint portfolio is also easier to manage, simplifies the process of pushing software updates and addresses book modifications, and offers more opportunities to control network utilization. Many organisations are surprised too late in the deployment that their endpoint investment can be dwarfed by media processing infrastructure required for interoperability.

Management and Support

Gartner clients consistently reflect that video should be “as easy as a phone call,” a goal that equipment vendors have continued to approach and, in some cases, have exceeded. But ease of use is only one key attribute for the ongoing adoption of video calling in the organisation. Research by Gartner reveals that there are several factors beyond setting up and connecting which are important for the acceptance of videoconferencing. Interpreters have expectations of video quality which, while more difficult to quantify than audio quality, must be adequate for all main use cases.

In addition to video quality, repeatability of the experience is important. Repeatability is whether the video- and audio quality are consistent from call to call and driven by a consistent process. While network quality plays a role in repeatability, it is also a by-product of the design and processes associated with the video environment, and how they are managed over time. So quality, repeatability and ease of use are all consistent themes for successful video deployments.

Delivering ‘like a phone call’ performance to a video environment means paying careful attention to design and management & support practices. Successful management & support of video means understanding the design triggers for service quality, identifying appropriate staffing levels and expertise specific to video, deciding what tools to leverage and knowing where service providers can best complement the end-to-end video service.

Road Map and Recommendations

Road Map and Recommendations

This chapter describes the most important recommendations related to Governance and to Change Management. Additionally, a specific road map for successful implementation of viable VC solution is given.

Governance

In order to successfully implement an end-to-end VC solution in the EU context, and to ensure that such a VC solution will always be of a consistent and repeatable quality, Gartner strongly recommends to start such an initiative with establishing a clear and agreed on *governance model.* Strong governance is required to have sufficient inter-institutional cooperation (as this will better ensure parties in the EU use the same VC solution, apply the same standards and processes, etc). This governance model should make the responsibilities and mandates of each party fully clear, including responsibility for development, and management & support (ideally under the responsibility of one party) of the VC solution, and the required funding to support this imitative.

Change Management

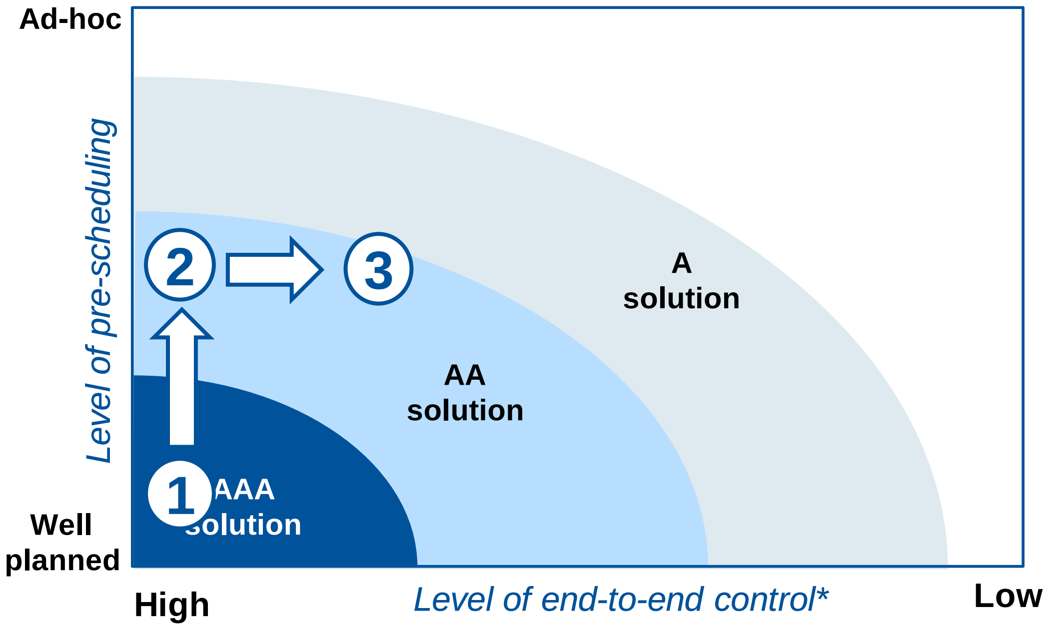
There should be clear and ongoing focus on *change management*, as only this will ensure sufficient involvement of all the different stakeholders, alignment of their requirements and expectations, compliance to the agreed on governance model, and finally acceptance of the to be developed end-to-end VC solution. As Gartner considers change management to be a critical factor for success of this initiative, it is recommended to:

- Ensure continuous communication regarding VC to all key stakeholders
- Involve all stakeholders (and specifically the interpreters) during the whole process of selecting the right VC solution
- Define clear acceptance criteria for the VC solution and ensure these are agreed on by all stakeholders
- Thoroughly test the VC solution before taking it into production and ensure involvement of the interpreters during testing
- Develop appropriate training materials and ensure all stakeholders are adequately trained
- Create and share a service portfolio for simultaneous interpretation and include VC (only the viable solutions)

Road Map

After institutionalizing this governance model and change management, Gartner recommends to apply a *three-phased road map* in which is it is critical to first establish a stable VC solution on the internal network that meets the expectations of all stakeholders (that is: an AAA solution). This stable internal solution not only provides the baseline from which to extend, but also serves a critical role as a governor toward more ad hoc demand in later stages. Therefore Gartner recommends to apply a road map which consists of three phases (see Table 4).

Figure 19. Road Map



Road Map — Phase 1

Phase 1 should start with formulating and agreeing on (with all stakeholders) a full set of requirements and the associated design principles. These should then be applied to guideline the process of selecting an adequate end-to-end VC solution for conferences with simultaneous interpretation. This will be then become the *de facto baseline VC solution*. A starting point for the design principles could be the initial set of design principles as previously formulated by Gartner. It should be clear that formulation of the design principles is not a onetime event, because it is expected that the design principles will evolve over time, allowing to adapt to additional use cases and/or changed requirements.

Gartner would like to stress that it is important not to be overly ambitious in this first phase, but rather to be pragmatic. For instance limiting the amount of participants in the distant sites will reduce the complexity of the internal VC solution significantly.

When possible, all VC endpoints should be converged to one strategic vendor. The portfolio of this strategic vendor should ideally not be limited to endpoints that only support the VC solution for Phase 1, but should also cover the endpoints required for Phase 2 (that is to extend the baseline video capability with a compliant Personal desktop solution). This better ensures compatibility and vendor support for the different end-to-end VC solutions. It is advisable to focus on endpoints which are based on new generation codecs (e.g., H.264 CVC), as these can relax the demands on bandwidth capacity and quality (see Appendix for more details on new generation codecs).

The selected Room-based solution will then have to be customised to support VC in multiple languages. Gartner recommends to initially start with broadcasting only one language to the distant site. This will strongly decrease the complexity of the baseline VC solution and will increase the chance of success.

Additionally, standards for the room conditions (lights and acoustics) should be developed, tested and prescribed for all the (distant) sites where the baseline solution will be used. These standards also need to be prescribed and applied in Phase 2, for the sites where the Personal desktop solution will be utilized.

In order to ensure that the VC experience will meet the interpreter expectations, ideally a converged or (preferably) an overlay network with sufficient bandwidth should be applied. In case only converged networks can be used, QoS will be required. And finally, before taking the VC solution into production, the Room-based solution should be tested thoroughly in a fully controlled environment.

Road Map — Phase 2

Phase 2 should focus on extending the baseline videoconferencing solution with Personal desktops. This will allow for being more flexible regarding the planning of meetings. In this case, it is most effective to build on the experience of the provisioned Room-based systems by systematically extending to a specific (to be determined) target number of distant sites. Only when this endpoint population is optimized with incremental WAN capacity (and appropriate QoS when required), additional endpoints are added. The number of Personal desktop endpoints supported should thus gradually be extended in successive stages, with an assessment of capacity and performance at each iteration. Key learning's should be captured at each stage, which also enables the adjustment of assumptions at each successive step.

What should absolutely be avoided is opening up the Personal desktop VC solution from the start to a large number of distant sites, specifically when SCIC doesn't have sufficient control over these sites. This approach will result in high "novelty" use of video in early stages, saturating links and degrading quality.

Taking a step by step approach also delivers the most consistent return on investment (ROI), because emphasis is initially placed on connecting to high-quality, managed room-based video systems, establishing a base of utilization for incremental investment. An added benefit is that the first wave of desktop endpoints is still likely to be in EU locations with the largest WAN ports.

Key actions for Phase 2 are:

- Select a Personal Desktop solution (including hardware) best meeting the current requirements, ideally from the same vendor providing the baseline VC solution
- In order to avoid transcoding (which causes delays and lip-sync issues), the codec(s) used in the endpoint in the main site should be leveraged (matching up different video codecs is expensive and adds latency to every call that has to be processed because of required transcoding).
- Customise the Personal Desktop solution such that is accommodates high quality audio and video and standardize on this. When VC endpoints are standardized, it is easier to manage, it simplifies the process of pushing software updates and address book modifications, and offers more opportunities to control network utilization
- Test and roll out (step by step) the Personal desktop solution in a fully controlled environment (related to network, room conditions, standardization, and support processes)
- Also test a range of peripherals such as microphones and camera's (do not only limit this to only one model) which are of sufficient quality to be used for the Personal desktop solution. The outcome of this should be a list of preferred hardware, which will aid in the roll-out and adoption process at a later stage. This list should include (not exhaustive):
  - ❑ A list of hardware requirements (which might included software considerations)
  - ❑ A list of microphones that can be used
  - ❑ A list of camera's that can be used
  - ❑ Operating manual

Phase 2 can already partly be carried out in parallel with Phase 1. However, Gartner strongly recommends to only start Phase 2 when the set of requirements and design principles is fully agreed on, and to not roll out the Personal desktop solution before the baseline VC solution is fully accepted.

Road Map — Phase 3

When Personal desktop is successfully added to the portfolio of VC endpoints (in a fully controlled environment), Phase 3 should focus on trying to facilitate meetings where the level of end-to-end control is not so high. For instance meetings with participants outside the private networks via the Internet. The main goal of this phase is to search for the tipping point where instead of only the 'AAA solution', perhaps also an 'AA solution' could be acceptable for the interpreters to work with during simultaneous interpretation. This will dramatically increase the reach of the end-to-end VC solution and flexibility to facilitate such meetings.

When it is determined where this tipping point is, a number of 'operating principles' should be defined, prescribing under what conditions such a meeting can be organized and take place. Elements of these operating principles should be among others the required bandwidth, room conditions and specific process agreements.

Appendix

Appendix

Digital Encoding Formats

The videoconferencing industry continues to rely on proprietary technologies for their codecs, although standards-based signalling and encoding protocols are increasingly being adopted within vendor products. The most common standards are:

- H.263: This standard is the predecessor to H.264 for video compression developed by the ITU-T Video Coding Experts Group (VCEG)
- H.264/MPEG-4 Advanced Video Coding: This is a standard for video compression developed by the VCEG with the International Organisation for Standardization (ISO)/International Electro technical Commission (IEC) Moving Picture Experts Group (MPEG)
- MPEG-2: This standard defines the coding of video and associated audio. It was developed by MPEG and is the predecessor to MPEG-4. It is an ISO/IEC standard — ISO/IEC 13818
- MPEG-4: This standard defines the compression of audio and visual digital data and was developed by MPEG. It is an ISO/IEC standard — ISO/IEC 14496

Videoconferencing Infrastructure

Dedicated Videoconferencing Infrastructure

Dedicated infrastructure is often already embedded in large enterprise video deployments, creating asset inertia that is exaggerated by existing staffing models and mature processes to insure reliable call production. This installed base of enterprise video networks is dominated by room based appliances that, as a result of their room associations are used on a prescheduled basis. This allows the aggregate demand set for video infrastructure to be well characterized and offers elegant resource smoothing, since room availability provides a natural governor for video infrastructure utilization. Since demand in this case is not highly variable, it is straightforward to size infrastructure resources like MCU ports with little risk of overbooking.

Security also is a key decision criterion for enterprises supporting videoconferencing. Since well over 90% of video calls involve internal participants, transport security is simply a matter of provisioning the WAN, typically a trusted network for production data or an equally secure overlay network. And while a private WAN can be extended to an external video service provider, this topology usually requires multiple landing points, with the need to hub traffic before handing off to the service provider. This adds cost and complexity, moreover, the added dedicated security equipment at the provider edge makes it more difficult to change video providers. Large organisations are also concerned about more than just video and audio security in-flight to the provider. Since room-based meetings are scheduled, enterprises also do not want to expose meeting topics or participants to a third party that could draw inferences in classified areas.

Large organisations are not solely dependent on data centre infrastructure for video. As organisations have engaged in immersive videoconferences, these exchange services are delivered over a private network, often from the incumbent Multiprotocol Label Switching (MPLS) supplier, but rely on service provider infrastructure to interconnect private video networks. The tie-in between video exchange services and the private networks that interconnect them have made it easier for immersive VC to rely on a hybrid solution of internal and service provider solutions that all reside on-net. As some large organisations have adopted this approach for managed immersive VC, the same conduit is being extended to other non-immersive endpoints. Immersive VC continues to validate the private hybrid approach in large organisation video, where an internal service is selectively extended to a community of interest, while the majority of call requirements remain satisfied by dedicated infrastructure in the data centre.

New Forms of Videoconferencing Infrastructure

Transcoderless and software-based video infrastructure solutions are two relatively new mechanisms for dramatically reducing the cost of deploying videoconferencing infrastructure used for multiparty calls. At a time when the cost per port of a typical high-definition multipoint control unit (MCU) runs into thousands of dollars, such cost reductions are a prerequisite for mass adoption of personal video. A software-based MCU (or softMCU) has the same architecture and provides the same basic functionality as a traditional MCU, but does so using software running on off-the-shelf server hardware. Transcoderless solutions rely on the use of different architectures to link endpoints, and switch or route video traffic across the network, rather than bridging through an MCU, removing the need for an MCU altogether.

Emerging video infrastructure architectures use mechanisms for dramatically reducing the cost of deploying central infrastructure in videoconferencing environments, a key component in enabling multiparty calls. At a time when the cost per port of a typical high-definition MCU runs into thousands of dollars, such cost reductions are a prerequisite for mass adoption of personal video.

Two principle forms of alternative architecture are starting to appear: softMCUs, which use an architecture that has the same configuration as a traditional MCU-based video architecture, but replaces dedicated hardware with software running on generic servers; and transcoderless architectures, which use alternative mechanisms (implemented in hardware or software) to remove the need for an MCU altogether. Transcoderless solutions may also use software-based infrastructure for the routing and switching elements, and both transcoderless and softMCU architectures could implement other video infrastructure technologies, such as firewall traversal in software, too.

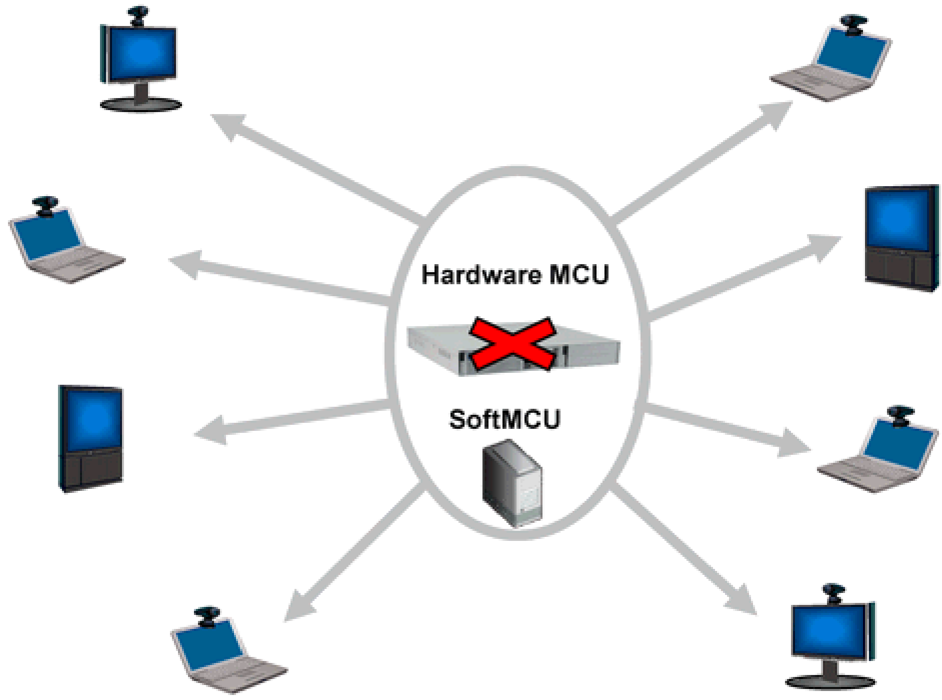
SoftMCUs

Architecturally, softMCUs can be a drop-in replacement for hardware-based MCUs in many implementations, as shown in Figure 19. With sufficiently powerful generic servers, most codecs (including the Scalable Video Coding [SVC] and Advanced Video Coding [AVC] implementations of H.264 (described in the next paragraph) can be implemented in software. Traditionally, it has been common practice for all endpoints to interconnect with a single MCU instance, which creates something of a bottleneck. However, newer implementations of both hardware and softMCUs use signalling to allow cascading across multiple MCU instances in larger deployments.

SoftMCUs use software-based encoding and decoding, coupled with generic compute power and virtualization, to maximize the utilization of low-cost CPU cycles, so reducing the cost of deploying video infrastructure by up to one order of magnitude, compared with existing architectures. Vendors include Siemens and Blue Jeans Networks.

By using an infrastructure-as-a-service model to charge such MCUs on a peak-utilization or on-demand basis, providers can deliver “just enough” port capacity to satisfy user needs. Such environments can handle transcoding between different video encoding standards, but they perform better in single codec calls that avoid the need for more complex computation. The most common encoding standard for such systems currently is H.264 AVC, the video encoding standard used in the majority of Internet Protocol (IP) videoconferencing deployments, although one of the benefits of a software-based MCU is that the cost of producing a new codec implementation is generally much lower than if specialized silicon had to be developed, as is the case for most hardware implementations.

Figure 20. SoftMCU



Transcoderless Infrastructures

In transcoderless architectures, a mixture of codec selection and endpoint signalling ensures that the infrastructure needs only to be able to forward video traffic, not process it, so doing away with the need for an MCU altogether, as shown in Figure 21. Such solutions manage this by having all endpoints use the same codec.

Two principal architectures exist: router-based implementations of H.264 SVC and switched implementations of H.264 AVC. Multiple switches or routers may be used to transport traffic across the network, which spreads the traffic load and enables traffic to remain geographically local, where feasible.

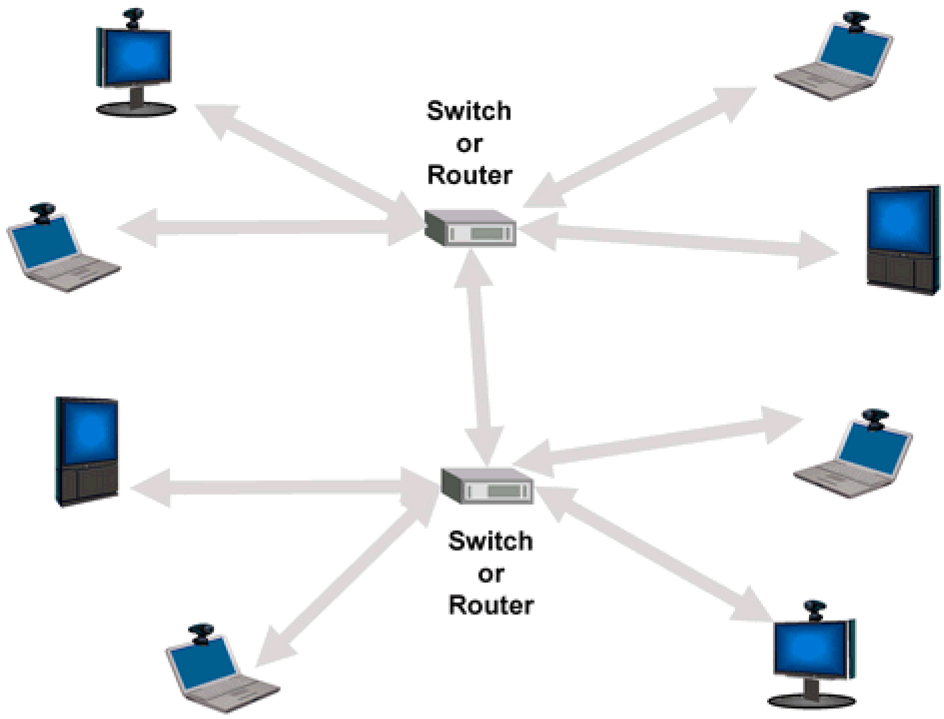
The router-based implementation of H.264 SVC, an architecture introduced by Vidyo, and now also sold by various Vidyo channel partners, lets each endpoint send and receive video at a frame rate and resolution that suits its constraints (due to camera or screen resolution, processing power, or available network bandwidth). SVC builds layers of enhanced data on top of a base-layer video stream. The base layer can be read by any H.264-compatible decoder. The enhanced layers provide additional temporal information (intermediate frame data) or additional spatial information (higher-resolution data). Each layer of data can be independently addressed, so the encoding endpoint needs only to send out a single stream (consisting of base and enhanced layers), while the decoding endpoint receives only those enhanced layers pertinent to its situation (which may be limited by screen resolution, decoding frame rate or the overall bandwidth available to deliver the video stream).

By replacing dedicated MCU hardware with a more generic network router, to discard any tagged layers that a particular endpoint cannot receive, infrastructure costs can be reduced by up to two orders of magnitude. SVC deployments also confer other significant cost-to-performance benefits. They are, for one thing, better at dealing with the vagaries of best-effort networks such as the Internet, meaning that endpoints can be connected over lower-cost bandwidth than real-time traffic on a Multiprotocol Label Switching (MPLS) network (the type generally used for current IP videoconferencing deployments). Additionally, they remove the latency associated with transcoding in the MCU, which either improves end-to-end latency, or frees additional latency-related budget for traversing a best-effort network like the Internet. Furthermore, in addition to supporting voice-activated switching, implementations of this approach can increasingly blend lower-resolution feeds from multiple users at the same time, in order to enable continuous presence of participants.

Switched implementations of H.264 AVC use Session Initiation Protocol (SIP) signalling to provide endpoint monitoring, and are more geared toward networks that provide guaranteed bandwidth. Vendors include Cisco.

Each endpoint can choose to switch in or out the video coming from another endpoint, based on controls that the user implements or automated processes such as voice-activated switching (so that the video for the active speaker is always being received). Such implementations benefit from the lower cost and lower latency also seen in a router-based H.264 SVC environment, but they cannot translate — all endpoints must send and receive at the same resolution and frame rate, as a switch cannot process individually addressed layers of video information in the way a router can. Extensions to this approach are beginning to appear that overcome this shortcoming, however, by using SIP signalling between endpoints to send and receive lower-quality images in real time, which can then be blended at the far end to create a composite continuous-presence image of multiple participants, for example.

Figure 21. Typical Transcoderless Infrastructure



Architectural Constraints and Combinations

Table 5 outlines key differences in the way each of these architectures handles multiple codec standards, resolution and frame rate variations, and future-proofing for pure implementations of each type of environment that do not use gateways or other devices to connect dissimilar architectures in a hybrid manner.

Table 5. Key Differences Between Video Architectures

Video Architectures				
	Hardware MCU	SoftMCU	Router-Based Transcoderless	Switched Transcoderless
What range of endpoint codec standards are supported in a single call?	Any to any*	Any to any*	All are identical	All are identical
What range of endpoint resolutions and frame rates are supported?	Any to any*	Any to any*	Any to any*	All must be identical
How can endpoints using other or future standards be connected?	Limited or no options*	Software upgrade*	Requires a gateway	Requires a gateway
* Within the limitations of available hardware, firmware and software performance. MCU = multipoint control unit				

Gateways will play an increasingly important role as offload points for codecs or designs that are not handled natively in a given architecture. They could also extend the life of individual investments by potentially allowing users to develop multiple video domains of distinct video architectures. However, this approach will add to the complexity of the environment and reintroduce many of the bottlenecks that newer architectures are designed to avoid.

The choice of codec used plays a fundamental role in the interplay between different implementations of each of these architectures. Both H.264 SVC and H.264 AVC can be deployed in either a transcoderless or softMCU architecture (and different vendors are exploiting each of these scenarios). AVC and SVC are not directly compatible with each other above a common base layer (352x288 resolution, Common Intermediate Format). However, these architectures may well develop in hybrid environments, with, for example, transcoderless H.264 SVC environments being linked to softMCUs running regular H.264 AVC via a gateway. Both softMCUs and transcoderless environments will increasingly be sold on a SaaS basis. Both are suited to cloud deployments by service providers.

Factors That Will Drive Adoption of New Videoconferencing Infrastructures

The biggest single driver of adoption is the growing demand for pervasive video to desktops, and other personal devices. In a typical large enterprise, the deployment of a few hundred group video systems, each costing thousands of dollars, may warrant investment in hardware ports on an MCU (each costing thousands of dollars). But this MCU cost cannot be justified when scaling up to deploy tens of thousands of soft clients across the organisation. Gartner forecasts that, by 2015, there will be 132 million users of enterprise-supplied personal videoconferencing globally, up from just 9 million in 2010. To put this in context, only around 2 million hardware videoconferencing endpoints have been sold since the industry started in the 1980s, and fewer than 1 million of these are still in active use. To support such a massive increase in connected endpoints will require orders of magnitude more call-handling capacity.

Increased use, not just supply, of video technology is a necessary part of this transformation, of course. Many people have become used to cloud-based video delivery in their private lives, through the use of Skype and similar solutions. Many others remain “video-sceptics” and are therefore not natural adopters of personal video solutions. But pressure to conform to new ways of communicating will come from management, and from external users in the core value chains of the organisation. Video is at a similar stage in its cultural adoption to mobile telephony 20 years ago. At that time, many felt it was socially unacceptable to talk on a mobile phone in a public place. Now, it is considered awkward not to use mobile communications.

SoftMCUs make for an easier transition for many organisations with existing deployments, because they use the same hub-and-spoke model of endpoints connecting via an MCU, and can be more readily mixed and matched with existing investments. Transcoderless solutions have initially proved more suitable to greenfield deployments or those where pervasive video and room systems are being kept apart. Over time, however, integration will improve here, too, which will make it easier for organisations to adopt these architectures alongside their existing environments.

If both these approaches have their attractions, the choice of which to use in a given situation largely depends on the mix of networks and services deployed alongside. Transcoding is important where value-added services are envisaged (for example, audio-to-text transcription) and acceptable where other hubbed services form part of the mix (for example, with Web conferencing). Transcoderless solutions help to reduce latency and so can work well with best-effort or high-latency environments (Internet connectivity, for example) and where a homogeneous environment is envisaged (in the sense that all endpoints use the same encoding algorithm). Cost, of course, also plays an important role, with softMCU architectures costing in some cases five to 10 times as much as transcoderless architectures to implement.

Factors That May Inhibit Adoption of New VC Infrastructures

SoftMCUs are, by nature, less efficient than a well-constructed, built-for-purpose hardware device. They are likely to introduce more transcoding-related latency, which will limit their appeal for certain applications where latency is critical.

H.264 SVC is more processor-intensive for endpoints than its predecessors, as during encoding, the data stream needs to be split into layers, and these layers must be requested and regrouped and recoded. They therefore require relatively powerful PCs, which means that a significant proportion of desktop computing resources in low-end PCs will be eaten up in the process, and that mobile devices such as tablets may be unable to operate at maximum potential performance, unless there is a graphics engine to which they can offload some of the computing effort.

Due to the additional overhead of layering the data, H.264 SVC requires around 10% to 15% more bandwidth than H.264 AVC at an equivalent resolution. So, in a steady-state environment, that is to say, where all endpoints share the same codec, and there is sufficient bandwidth to allow them to run at the same resolution, SVC will consume more bandwidth than AVC: it really comes into its own when best-effort or contended bandwidth is being used, as it can use up “just enough” bandwidth to maximize the user experience within a given set of constraints.

To date, there are multiple implementations of SVC, which have yet to be tested properly for interoperability. This will take time and almost certainly cause version control problems for users as software gets updated to reflect post-testing requirements. SVC interoperability is, of course, essential for SVC to take off as a globally accepted standard. And this means different architectural models being able to interoperate too. For example, while Polycom and Radvision both support their own SVC implementations, neither has yet introduced a transcoderless variant. These will still need to interoperate with the model used by Vidyo, HP and Teliris, which does not need an MCU.

Impact of New VC Infrastructures

The move to new forms of infrastructure delivery will have several impacts on the market. First and foremost it will be a significant contributor to the transition of videoconferencing from a low-volume, high-unit-price market to a mass-deployed, high-volume, low-unit-price market. It is inconceivable that organisation wide personal video clients could be deployed unless the cost of the associated infrastructure falls by orders of magnitude per user. But this is not the only requirement for this change: mechanisms to lower the cost of bandwidth associated with video (through a combination of greater use of lower-cost bandwidth, such as the Internet, and more efficient codecs, such as the soon-to-be-ratified H.265) are also essential.

Smaller providers such as Vidyo already have pure software and transcoderless architectures. Some major providers, such as Polycom, are actively embracing the software concept. Others, such as LifeSize, are seeking to introduce as-a-service implementations that blend hardware and software.

The move to software lowers the barriers to market entry, making it more feasible for new market entrants, both technology startups and more mature technology and service providers, to implement their own solutions. Service providers with their own software architectures, such as Blue Jeans Network, will seek to develop either direct sales or partnership models to extend the reach of their solutions.

However, different codecs, architectures and the constant need to test for interoperability will ensure the survival of traditional transcoder-based environments. Cisco is betting that requirements for the highest-quality user experiences will continue to be fulfilled by hardware-based infrastructure.

Service providers, from carriers such as BT, AT&T, Tata, Orange and Telefonica, to independent providers like Glowpoint, BCS Global, InterCall, PGi and Arkadin, all need to make technology choices that give them the flexibility to support different customer needs at different prices. So it is likely that these providers will be at the forefront of blending different mixtures of hardware, software and transcoderless environments into product sets, to hit different price points and meet differing performance and scalability requirements.

Multipoint Conferencing

Video bridging, continuous presence in particular, is resource-intensive if minimal additional latency is to be added to the call path, and is usually performed on dedicated hardware. The simplest option for adding bridging to the environment is to simply bridge the calls directly on the endpoint. This approach requires an augmented video endpoint with integral multipoint control unit (MCU), and is most practical for all-IP calls. Video endpoints have historically been available with a four-way MCU option, but are now also available in eight-way and nine-way configurations that can conference endpoints, including HD endpoints, in a wide range of video layouts. Recurring conferences can even be programmed directly into the address book to create click-to-meet capabilities from the endpoint remote control.

Multipoint bridging directly on an endpoint(s) is most effective for small endpoint populations (20 or fewer systems), since large calls cannot be accommodated. In this approach, it also is required that the endpoint with bridging functionality have sufficient bandwidth to account for each leg of a multipoint call. For example, a six-way call on a multipoint system will require approximately 2.5 Mbps of network capacity. It is a longstanding misconception that all the endpoints in a multipoint call require this additional bandwidth; this is not the case, as they are only transmitting their own video, and receiving a single video stream created by the bridge.

Organisations with larger endpoint populations (greater than 20 endpoints) that have outgrown endpoint-based bridging can leverage a VNOC provider for bridging services. Bridging services are accessible via public and private IP connections, as well as ISDN, and currently cost about €25 per hour for each video participant in the call. Prices for service-based bridging tend to rise in linear fashion as call bandwidth increases. Service provider bridging is preferred for environments that have large or complex calls that require transcoding, speed-matching (sometimes called “transrating”) or high levels of in-call monitoring. This is also a preferred approach for organisations that wish to outsource videoconferencing, since scheduling, reservations and call production are best provided by the same entity. Service provider bridging is also the only viable alternative for large environments that are unable to converge on IP transport as a result of endpoint or network considerations. Immersive VC exchanges also provide bridging and switching in a more dedicated fashion, and carry higher units rates of closer to €100 per hour.

Organisations with larger endpoint populations that have transitioned video traffic to their WANs should consider dedicated MCUs to service their conferencing requirements. These appliances can be deployed in a data centre or other effective point of network concentration, and collect large numbers of IP or ISDN systems together into multipoint calls. Dedicated MCUs can support a combination of SD and HD (including telepresence) systems, and can add VoIP and public switched telephone network (PSTN) audio participants. As in the case of multipoint-capable endpoints, MCUs require enough bandwidth to handle the anticipated concurrent call volume. As global requirement grow, enterprises should consider regionalizing their MCU architectures, and should provide capacity in-region to avoid excessive backhaul and the associated latency.

Quality of Service (QoS)

Quality of service (QoS) is a measure of the performance of a communications network. There are many ways to achieve an intended QoS in a communications network and we refer to the collective set of tools as QoS technology. The need for QoS technology increases as enterprises combine multiple networks with different characteristics and priorities (e.g., transaction traffic, real-time multimedia, Voice over Internet Protocol [VoIP], and Internet Web browser traffic) to create a single, enterprisewide, multiservice utility network.

Although the networks are combined into one infrastructure, the different categories of traffic still have different requirements for network performance. Different applications also have different importance to the enterprise, and they should be given different priorities when the network is congested. QoS technologies assist with this situation, providing the different levels of performance and different priorities to applications when needed on the single multiservice network.

Some typical cases requiring QoS are:

- Mixing transaction processing, Web services, or multimedia with low-priority Web browsing and bulk file transfer on low-bandwidth lines
- Sharing of expensive narrow-bandwidth links by different organisations that want to apportion bandwidth
- Mixing VoIP, which requires short latencies, with transaction traffic and bulk file transfers

Network Traffic

Network traffic can be characterized by the presented workload, or throughput, and by three classic network performance requirements: latency, jitter, and error rate:

- Latency is the time needed for transit across the network; it is critical for real-time services. Excessive latency interferes with applications that are latency-dependent and degrades the quality of websites and of interactive sound and video.
- Jitter measures the deviation from ideal, evenly spaced packet arrival. It is created whenever there are queues, buffering, and path rerouting in a system. The de-jitter buffer incorporated into most receivers is an important contributor to total system latency. Bursts of jitter automatically increase the size of these adaptive de-jitter buffers, adding latency to all packets for a considerable time after the burst.

- Error rate is the result of packet loss, corruption, or, in some cases, delay beyond the point at which the data can be used. It has different effects on user experience, depending on the transport and application using the service. Some transports, such as Transmission Control Protocol (TCP), recover from lost or corrupted packets, but they increase the effective latency because they must wait for retransmitted data to arrive. Other transports and applications, such as VoIP and some User Datagram Protocol (UDP) real-time video applications, cannot wait for retransmission and attempt to conceal or ignore errors, resulting in increased bandwidth demand (for the error-concealing redundancy) or degraded user experience.

Traffic within each traffic category may be further subdivided by priority. Priorities are not substitutes for traffic categories, which are based on absolute requirements. Unlike a traffic category that supports an application that will fail or perform unacceptably if sufficient communications service is not provided, priority traffic receives preferential treatment because of its importance to the enterprise. Because no absolute network measurements are associated with a priority, the meaning of a particular priority (in terms of measurable network performance) varies from one network to the next. Examples of priority traffic are order-entry transaction processing having priority over other types of interactive traffic, and intranet Web browsing of internal documentation by order-entry personnel having priority over Web browsing of sports scores.

The capacity available within each traffic category is adjusted to meet the agreed-on service levels for each user group. Prioritization becomes necessary only if a category doesn’t provide sufficient capacity to accommodate all the presented traffic; priorities can be used to differentiate among user groups, and among applications and users within a group. If there’s not sufficient capacity, the network can either refuse to admit additional traffic in the affected traffic categories (admission control), or it can selectively discard or delay packets in the affected traffic categories (congestion control). Prioritization assists with these decisions.

With QoS service levels usually comes a demand from enterprise managers for service-level agreements (SLAs), combined with the resulting need to measure the presented workload and to enforce those service levels.

Note that many situations may not require QoS, which can be complex and expensive to design, implement, and manage. Before creating an overly complex set of requirements that lead to the implementation of an overly complex QoS infrastructure, ensure that complexity is actually necessary and will provide business results that are worth the cost and effort. For example, QoS is not useful when the great majority of traffic is typically of the same type and of approximately equal importance to the organisation.

QoS may also not be needed if it’s less expensive to “throw bandwidth at the problem” by over provisioning. That’s often the approach used by large-scale Internet service providers (ISPs), which have huge amounts of optical bandwidth available over which they carry traffic flows that have been smoothed by massive amounts of aggregation[1].

Management of End-to-End Videoconferencing

One of the key lessons learned from successful videoconferencing implementations is that while video quality is important, high levels of adoption are driven by investing in proactive VC management that delivers a consistent experience on which participants can depend. The following are the key elements of managing a videoconferencing environment for success:

[1] Aggregation of moderate or low amounts of IP or Ethernet traffic does not, surprisingly, smooth the aggregated flow to any major degree. That's because of some unusual mathematical characteristics of such flows. On high-bandwidth paths, as are found on major corporate and ISP backbones, traffic smoothing by aggregation does occur

- Endpoint management
- VC infrastructure management
- VC staffing

Endpoint Management

The most critical element of managed video is the endpoint, the element that directly dictates the user experience. One of the most important reasons to manage endpoints is that they are distributed and inherently under less control. End users can accidentally or intentionally reconfigure endpoints for the purpose of a specific meeting, leaving the succeeding meeting participants unable to connect. Users can also struggle with connecting peripherals, leaving video equipment physically reconfigured and in disarray for the meeting that follows. In simple cases, equipment may just be inadvertently unplugged or moved to another location where IP addressing can cause calls to fail. Video endpoints also require frequent software updates and need to have consistent and current directory services. This may be as simple as an e.164 alias or Session Initiation Protocol (SIP) uniform resource identifier (URI), or as elaborate as a listing in a video exchange directory for connectivity outside the firewall.

Hardware-based video endpoints have SNMP Management Information Bases (MIBs), but usually expose only rudimentary information about the state of the endpoint. Larger deployments (more than 10 distributed endpoints) are best served with an endpoint-specific management platform like Cisco TelePresence Management Suite (TMS), Polycom Global Management System (GMS) or LifeSize Control. These systems are available as appliances or software that can be run on an existing server, and provide a substantial functional uplift to what is available via Simple Network Management Protocol (SNMP) only. In addition to visibility into the state of the endpoint population, endpoint management platforms can push software updates and global phonebook entries, ensuring a consistent look and feel. New endpoints added to the network can be centrally configured and controlled, and endpoint utilization can be captured for capacity planning, as well as for tracking the ROI of incremental endpoint capacity. In addition, endpoint management platforms can provide raw call detail records (CDRs), as well as dashboard reporting to help gauge performance of endpoints and infrastructure.

Managing Videoconferencing Infrastructure

The network impact of video traffic needs to be managed as early in the process as possible to avoid impacting production data, as well as coincident real-time traffic. Since videoconferencing has often been loosely managed by the facilities group, organisations need to establish a trigger for shifting this responsibility to the network staff. A straightforward approach is to begin managing and controlling video traffic as soon as it transitions to IP. Once the video traffic moves to IP transport, it become critical to size, plan and manage it. Since there is no mean opinion score (MOS) for video, the forensics are more complicated, and network staff are often forced to rely on network quality as a proxy for video quality. As a result, granular visibility into network performance is required, along with the ability to associate network performance with call records. The element management platforms mentioned provides visibility into network performance like packet loss, and offer forensics that allow operators to dig deeper into video quality issues on a per-call basis. Video-specific test and monitoring tools like Applied Global Technologies (AGT) Fathom have also been introduced that can perform link testing in advance of a video call with synthetic video traffic to validate network performance. Network performance is not synonymous with video quality, but has the greatest correlation with video quality and can be measured explicitly, usually with existing tools such as those in the Measuring Results section below.

Staffing

Organisations transitioning from a service provider model to an internal service should carefully consider staffing ratios to deliver the anticipated level of service. For infrastructure, specifically multipoint control units (MCUs), even large environments can be run with a limited pool of resources. Video MCUs and attendant infrastructure are usually run by one to two full-time technicians, even when multiple MCUs are involved. This limited staffing for the infrastructure is the result of equipment and process standardization, as well as the proactive use of management tools to triage potential call-impacting problems. Bridge technicians can manage a large number of active calls through a combination of MCU tools and visual monitoring of active calls.

Head count for endpoint management is more varied, depending on the amount of site-level support required. In heavily managed video environments, enterprises can have a video technician dedicated to each large site, resulting in a ratio of one technician for every 10 to 20 endpoints. For organisations where video calls are self-launched or remotely launched on the user's behalf, field-level technician staffing can be reduced to only the most critical locations that require concierge support.

Scheduling and reservations can also be resource-intensive, but only for enterprises that rely on phone and fax confirmation, or those that have a complex hierarchy of room rights (who gets to schedule which rooms in which locations). Basic video scheduling can be moved online, but high-touch scheduling and reservations can involve additional head count for every 50- to 100-room system. Staffing levels are highly dependent on standardization of video endpoints and the infrastructure, and remain lowest in environments that are highly standardized.

Any questions regarding this Report should be addressed to:

Danny Starre
Senior Consultant
Gartner UK Limited
Gartner Netherlands - Amsterdam
Toren A De Entrée 79
16th and 17th Floor
Amsterdam South East 1101 BH
NETHERLANDS

Telephone: +31 20 3 144 144
Email: danny.starre@gartner.com

This Report was prepared for DG Interpretation (SCIC):

Lode De Raedt
DG Interpretation (SCIC)
Rue de la Loi 200
B-1049
Brussels BELGIUM

Telephone: +32 2 295 9142
Email: lode.deraedt@ec.europa.eu