How does Videoconferencing Work?

PictureTel

videoconferencing 101

- WHAT HAPPENED?
- 1,000,000 PIXELS
- FRESH-SQUEEZED VIDEO
- SOUNDS GOOD
- DIGITAL HIGHWAY
- ACCEPTABLE STANDARDS
- PROPRIETARY SECRETS
- SOME BUZZWORDS
- THE FUTURE
- WHERE DO I GO FROM HERE?

This booklet is part of the PictureTel Videoconferencing 101 series, designed to help you learn about the uses, benefits, and technology of videoconferencing.

PictureTel

100 Minuteman Road, Andover, MA 01810

WHAT HAPPENED?

Suddenly, videoconferencing is a key business tool. Why now? Why not decades ago, when videoconferencing would have been just as valuable as it is today for communicating face to face at a distance?

The answer is easy: The core technologies that make videoconferencing practical simply weren't ready.

The most important of these technologies are

- ❑ **Powerful, inexpensive computers for processing video**
- ❑ **Advanced techniques for compressing live video and audio**
- ❑ **Digital phone lines and computer networks for making connections worldwide**

And just as important as the technologies themselves are the international standards that enable systems everywhere to connect and communicate.

Although the first videoconference actually took place more than 30 years ago, at the 1964 World's Fair in New York, its price tag was millions of dollars. This booklet introduces the technologies that have transformed an expensive, impractical, but captivating demonstration into an affordable, practical, and still captivating business tool.

So...how *does* videoconferencing work?

1,000,000 PIXELS

If someone asked you to describe video-conferencing, you might say it's like combining a phone and a TV. And on the surface, that's not a bad description. But under the surface, videoconferencing sound isn't like telephone sound, and videoconferencing video isn't like television video.

THE DIFFERENCE

In a phone call your voice travels through the line as electrical vibrations, mimicking the air vibrations your voice makes. In a more complex way, a TV picture travels through the air (or cable) as electromagnetic vibrations, mimicking the patterns of dark and light and color that make up the picture.

But in a videoconference, the picture and sound travel from place to place as computer data—electrical pulses representing the 1s and 0s of digital communication. Captured video and audio get converted to computer data, processed by computer circuits, routed through phone lines or networks made for computer data, processed again by more computer circuits, and finally converted back into video and audio.

BUT WHY?

Why go to all the trouble? Why not send the sound and the picture in the more straight-forward ways that phones and TVs do?

The problem is the video, which contains far too much information to squeeze through a dial-up telephone line or network. Such a line can't carry anywhere near the amount of information your TV antenna or cable can.

To push a reasonable likeness of live video through the line, you have to process the dickens out of it at each end. You can do that only with a computer, and that means you have to go to the trouble of converting the video signal into computer data and converting the data back into a signal again.

IN A VIDEO-CONFERENCE, THE PICTURE AND SOUND TRAVEL FROM PLACE TO PLACE AS COMPUTER DATA.

FRAMES AND PIXELS

A videoconferencing system, like a TV, deals with video as a series of still pictures, or frames. Show enough frames per second, and your brain sees live video. TV shows about 30 frames per second and a movie shows 24. A videoconferencing system typically shows 15 or more, even up to 30.

While the frame rate of a TV or movie image is fixed, the frame rate in a videoconference is flexible. The system can reduce the frame rate if there's simply too much information to squeeze through the line. (This typically happens when the picture includes lots of motion, as explained in the next section.)

Each frame consists of a certain number of colored dots. Show enough dots, and your brain sees a smooth picture. TV shows the equivalent of about 120,000 dots (roughly 400 by 300); a videoconferencing system typically shows about 60,000 (256 by 240) or more. These dots are called pixels, short for picture elements.

Multiply 60,000 pixels by 15 frames, and you get the minimum number of pixels a videoconferencing system has to handle each second: about 1,000,000. Multiply that by 2, because the system is both sending and receiving live video. And don't forget that it's processing sound (both coming and going) at the same time.

The problem, then, is how to squeeze at least two million pixels' worth of video data, plus sound data, through a phone line or network each second.

That's where video compression enters the picture.

THE PROBLEM IS HOW TO SQUEEZE AT LEAST TWO MILLION PIXELS' WORTH OF VIDEO DATA, PLUS SOUND DATA, THROUGH A PHONE LINE OR NETWORK EACH SECOND.

FRESH-SQUEEZED VIDEO

PRACTICAL VIDEO-CONFERENCING DEPENDS ON FINDING A WAY TO COMPRESS ALL THAT DATA INTO MUCH LESS SPACE.

The flood of raw data generated by live video and audio could fill hundreds of digital phone lines to capacity. Practical videoconferencing depends on finding a way to compress all that data into much less space.

It's a bit like trying to squeeze a pitcher full of orange juice into a small frozen can—in this case, a very small can indeed.

Video compression is *the* technology that makes dial-up videoconferencing possible. Virtually all systems use a common set of standard techniques, so they can communicate with systems made by other companies.

In addition, companies like PictureTel have invented advanced compression techniques that improve the video and audio when their own systems communicate with each other.

SHRINKING THE PICTURE

Two techniques work to reduce the amount of computer data needed to represent an individual still image. These aren't unique to videoconferencing; they're commonly used whenever computers handle pictures.

One is to represent "sameness" in a compact way. For example, suppose 50 pixels in a row are the same color. The computer can handle these pixels individually, as 50 pieces of data. Or the computer can handle the same information as just 2 pieces of data: a single pixel, plus a note saying that the next 49 pixels are all the same.

A second technique is to throw away information the eye won't miss.

Think of a color image as being made of two parts: a black-and-white image, plus color information. The black-and-white image consists of many dots, each one having a certain brightness. And the color information consists of a certain hue assigned to each dot.

But what if you don't assign a separate hue to every single dot? What if you assign a hue to each group of four dots instead, representing their "average" color? You'd cut the amount of color information by 75 percent. Yet, because your eye sees sharpness more in terms of the black-and-white part of the image than the color part, you would hardly notice the difference.

Using techniques like these, compression can reduce the data needed to represent an image by 90 percent or more.

THE MOTION'S THE THING

It helps to compress each individual frame of a live video image. But with 15 or more frames coming every second, even saving 90 percent isn't enough. So other techniques attack the problem by looking not at isolated frames but at the differences between one frame and the next.

Suppose you send live video of a bowl of fruit. What's the difference between the first frame you send and the second?

There isn't any, of course; nothing in the picture has changed. So, since the two frames are the same, why even send the second one? Why not just tell the system at the far end to keep displaying the same thing it was displaying $\frac{1}{15}$ second ago? "Just beam those same pixels up there again." After the system squeezes the first frame through the phone line, it can sit there idling indefinitely. "Same pixels." "Again."

Until something changes.

Suppose a ladybug, who was resting quietly at one end of the banana in the bowl of fruit, starts walking toward the other end. Suddenly a few pixels in the current frame are different from the ones in the previous frame.

What does the system do to show this change? It could send a whole new frame, but it doesn't really need to. It simply needs to figure out which pixels have changed, and send just those pixels.

Or it can do something cleverer. It can figure out that a certain oval group of pixels has moved from one place to a slightly different place. And instead of sending the individual changed pixels, it can send simple instructions: "Move this group of red and black pixels just a bit to the northwest, and fill in yellow behind it."

Such cleverness is more than the system needs just to show a moving ladybug. But real meetings between people are another story. When a dozen people gather in front of the camera and start talking, gesturing, shuffling papers, and making presentations at the whiteboard, a videoconferencing system needs all the cleverness it can muster.

OTHER TECHNIQUES ATTACK THE PROBLEM BY LOOKING NOT AT ISOLATED FRAMES BUT AT THE DIFFERENCES BETWEEN ONE FRAME AND THE NEXT.

SOUNDS GOOD

POOR SOUND MAKES THE PICTURE SEEM WORSE, WHILE EXCELLENT SOUND MAKES IT SEEM BETTER.

Although the picture is the feature attraction of videoconferencing, good sound is just as important for good communication. In fact, poor sound makes the picture seem worse, while excellent sound makes it seem better.

Good sound depends on competent microphones and speakers, of course. But it also depends on what happens to the sound between the microphone and the speaker. Does it survive intact?

NARROWBAND, WIDEBAND

A hi-fi music system sounds good because it can handle a wide range of frequencies, from low bass to high treble. Even sounds that are too high to hear, up to 20 kilohertz (kHz), survive its circuits pretty much intact. Another way of saying this is that the system's bandwidth is 20 kHz.

By contrast, a telephone sounds flat, because it can handle only a narrow range of frequencies. Its circuits carry sounds only up to about 3 kHz.

Basic videoconferencing sound has the same limits as telephone sound—its bandwidth is 3 kHz. As a result, it sounds flatter and less natural than you might like.

Many systems can switch to an improved sound mode with more than twice the bandwidth, 7 kHz. The resulting sound has more sparkle and clarity, a big improvement. So, why not use this wideband sound all the time? In fact, why not increase its bandwidth to hi-fi levels?

The problem is that both sound and video still have to squeeze through that phone line. Wideband sound uses more of the line's capacity, leaving less for video.

Standard narrowband sound takes 16 kilobits per second (kbps) of the line's capacity. By contrast, standard wideband sound takes 48 to 64 kbps.

So the advantages of wideband sound come at a cost. But if you have the line capacity to spare, the improved sound makes your meetings more natural and more effective.

And PictureTel systems have their own version of wideband sound that gives you more for less. More about that later.

ECHO CANCELLATION

After you master the science of getting crisp, clear sound in a videoconference, you want to make sure you don't hear it more than once. You need to eliminate the echoes.

Why do video calls have a problem with echoes, when ordinary phone calls don't?

In an ordinary phone call, the mouthpiece can't pick up the faint sound from the earpiece. But in a video call, the microphone does pick up sound from the speaker. Left unchecked, that sound would travel back to the other system, and the person speaking would hear his or her own voice echoed back.

An ordinary speakerphone sidesteps this problem by muting your microphone while the other party is talking. You can listen or you can talk, but not both at the same time. If both parties talk at once, some of your words get clipped off, ruining the flow of the conversation. And that's simply unacceptable in a videoconference.

A well-designed videoconferencing system solves the echo problem in a way that lets you maintain two-way (or "full-duplex") audio all the time. It actually cancels the echo.

The idea is fairly simple. Your system monitors the sound of your voice as you speak, and remembers that sound just long enough for its echo to return from the far end. As sound comes in from the far end, your system mathematically subtracts from it the original sound of your voice. The incoming echo is canceled out.

This process works even when people at both sites are talking at the same time. The system cancels only the echo, leaving other voices and sounds intact.

AFTER YOU MASTER THE SCIENCE OF GETTING CRISP, CLEAR SOUND IN A VIDEOCONFERENCE, YOU WANT TO MAKE SURE YOU DON'T HEAR IT MORE THAN ONCE.

DIGITAL HIGHWAY

ALONG WITH DIGITAL PHONE LINES, COMPUTER NETWORKS PROVIDE ANOTHER WAY TO LINK VIDEOCONFERENCING SYSTEMS.

All this cleverly compressed video and wideband, echo-free sound is no good without a way to move it from place to place.

Videoconferencing has become practical because of the growing network of dial-up digital phone lines. "Dial-up" means you don't have to string a special wire between one site and another. And "digital" means you're not limited by analog lines meant to carry voices.

BASIC...

The most common kind of dial-up digital service is called ISDN BRI (Integrated Services Digital Network, Basic Rate Interface). It gives you two data channels, each with a capacity of 64 kbps.

Only recently have computers been powerful enough, and compression technology good enough, to use such low-capacity lines for practical videoconferencing. They're a boon for two reasons: they're inexpensive, and they form a worldwide network that's growing daily.

...AND BETTER

Of course, you do get better-quality picture and sound if you use higher-capacity lines.

You can get higher capacity by using more than one line. A device called a multiplexer connects the lines and the videoconferencing system. The multiplexer makes the system think it's connected to a single higher-capacity line, while distributing the total flow of information over all the available lines.

You can also get higher capacity by leasing a different kind of line, such as ISDN PRI (Primary Rate Interface). This offers you a single channel with 384-kbps capacity, plus lower-capacity channels for additional digital connections.

NOT TO MENTION...

Along with digital phone lines, computer networks provide another way to link videoconferencing systems. As company intranets and the worldwide Internet grow and get faster, they are connecting more and more people in video meetings. This growth is likely to be very strong in the next few years.

To learn more about the ins and outs of network services and equipment, see the booklet *Network Planning—Connecting the World*.

ACCEPTABLE STANDARDS

When you make an ordinary phone call, you don't have to worry about whether you and the other party have the same brand of phone. You know you'll be able to connect, no matter who made your phones.

That's pretty much true of video calls, too. Practically everyone who makes video-conferencing systems follows international standards for connecting and communicating. In the jargon of the industry, such "standards-based" systems can "interoperate" with one another.

What exactly are these standards? Who comes up with them, and what do they standardize?

ITU-T ALPHABET SOUP

"Who" is mostly the International Telecommunication Union—Telecommunication Standardization Sector, or ITU-T.

This United Nations group has approved a variety of standards that affect video-conferencing technology. The standards have names that only an engineer could love: H.320, H.261, G.722, T.120, and many others just as descriptive.

(When people say these names, they may or may not pronounce the "." character. When they do, they say "dot"—as in "H dot three twenty.")

H.320 VIDEOCONFERENCING

One of the most important names in videoconferencing is H.320. This is a suite of standards that covers videoconferencing over ISDN and other digital phone lines. It deals with data rates between 56 kbps and 2048 kbps.

This suite of standards has a nickname, too: p×64 (pronounced "p times 64"). The name comes from the data capacity of ISDN lines, which is measured in multiples of 64 kbps.

Along with H.320, a set of standards called H.323 is gaining importance. It covers videoconferencing over local computer networks and the Internet.

VIDEO STANDARDS

One part of the H.320 standard, called H.261, describes how to process the video signal—by coding it as digital information to send, and by decoding the received information.

This standard defines some parts of the coding and decoding precisely, so different systems can interoperate. But it leaves other parts open

PRACTICALLY EVERYONE WHO MAKES VIDEO-CONFERENCING SYSTEMS FOLLOWS INTERNATIONAL STANDARDS FOR CONNECTING AND COMMUNICATING.

to creative engineering, with the result that one system's video can look quite different from another's.

It's again like frozen orange juice. Every producer condenses and freezes orange juice ("codes" it) into the same standard cans. And every consumer makes the standard concentrate back into juice ("decodes" it) by adding three cans of water.

But each producer uses a different process. One process gives you juice that tastes like fresh oranges, and another gives you… well, quite a bit less.

VIDEO CODING

The most critical engineering is in the coding. The standard defines the final "package" for the coded video but not the whole process for squeezing the video into that package. So the picture you receive depends mostly on how well the coder did its job.

The video coding process involves:

1. Segmenting each video frame into small, manageable pieces, similar to mosaic tiles
2. Evaluating which tiles have changed since the previous frame
3. Calculating the most efficient way to send each change—whether to send a whole new tile or to send instructions for tweaking the previous tile
4. Summing up all the changes to be sent, and then figuring out how to send as many of them as possible through a limited-capacity line

Step 4 is crucial. Any system can do a decent job of coding video when there's little difference from frame to frame (little motion). The best systems shine in the way they handle lots of motion—the way they figure out how to send big changes through a small line.

One effect of heavy motion is that you may see the edges between the tiles that make up the picture. This happens when not all the changes can be sent at once, and so the tiles no longer blend seamlessly. Systems that do a good job with step 4 can minimize this tiling effect.

THE BEST SYSTEMS SHINE IN THE WAY THEY HANDLE LOTS OF MOTION.

VIDEO DECODING

Decoding affects picture quality, too, although not as much as coding does.

The decoding process involves:

1. Reassembling the picture from the information received through the line
2. Trying to eliminate tiling effects ("artifacts") from the picture before displaying it

Step 1 is a straightforward process defined by the H.261 standard. But step 2, called postprocessing, is another opportunity for creative engineering.

The postprocessor's job is to find mismatches between each tile and its neighbors, and blend the edges together. The trick is to blend just enough but not too much. Too little blending means the tile edges will be visible; too much means the picture will lose its sharpness.

Postprocessing can't fully correct the results of a mediocre coding job. But the best systems can reduce artifacts to a minimum while keeping the picture as sharp as possible.

AUDIO STANDARDS

The H.320 standard covers audio as well as video. It includes two options for sending narrowband (3 kHz) sound and one for wideband (7 kHz) sound:

Name of Standard	Sound Quality	Line Capacity Required
G.711	Narrowband	64 kbps
G.722	Wideband	48–64 kbps
G.728	Narrowband	16 kbps

As you can see, the G.711 standard isn't very desirable—requiring a big chunk of the line's capacity but giving only narrowband sound. The other two options make much more efficient use of limited line capacity.

Better videoconferencing systems automatically choose the best audio option for each call, so you don't need to worry about it.

DATA STANDARDS, TOO

One more big name in standards-based videoconferencing is T.120. While H.320 covers the video and audio parts of a videoconference, T.120 covers the data part.

More and more videoconferences include computer-based meeting tools—such as computerized slide displays, Windows™ application sharing, and digital whiteboards. These tools work by sending and receiving data through the same line that carries the video and audio signals.

In addition, conference control features—such as far-end camera control and multipoint director control—also work by sending data through the line.

The T.120 standard defines how such data communication should take place. And it's a standard that can grow to accommodate new features such as multiple video windows, bridge reservation systems, and electronic voting.

All these technical standards—for video, audio, and data—are designed to let you call any videoconferencing system anywhere, without even knowing what "interoperate" means.

ALL THESE TECHNICAL STANDARDS-- FOR VIDEO, AUDIO, AND DATA-- ARE DESIGNED TO LET YOU CALL ANY VIDEO-CONFERENCING SYSTEM ANYWHERE.

PROPRIETARY SECRETS

The best videoconferencing systems capitalize on every opportunity for creative engineering.

In standards-based calls, the top systems send better pictures by using superior coders. And they display better pictures by using advanced postprocessors.

But standards lag behind technology. So an innovative company can design its systems to communicate with each other on an even higher plane, without sacrificing standards-based compatibility.

Even though the proprietary approach works only in calls between same-brand systems, that happens to be how most videoconferencing takes place. And a well-designed system automatically switches between standards-based and proprietary methods, as needed for each call. You never even have to think about it.

EVEN THOUGH THE PROPRIETARY APPROACH WORKS ONLY IN CALLS BETWEEN SAME-BRAND SYSTEMS, THAT HAPPENS TO BE HOW MOST VIDEOCONFERENCING TAKES PLACE.

SHARPER VIDEO

Standards-based video can be quite good. But engineers are constantly working on video compression technology that yields higher frame rates, smoother motion, and sharper pictures.

A system that uses a proprietary video coding method can clearly stand out from others. PictureTel's SG4™ is one such method, resulting from more than 10 years of research originating with the company's founders at MIT. It uses a technique that avoids tiling effects altogether and results in much better motion handling.

If you'd like a one-sentence engineer's summary: "The frame encoder is implemented by a motion-compensated, recursive, hierarchical vector quantization algorithm, operating at successively finer levels of resolution, and optimized using background prediction."

CLEARER, CLEANER SOUND

Like standards-based video, standards-based sound can be quite good—if it's the wideband variety. But that requires 48–64 kbps of the line's capacity.

A hallmark of PictureTel systems is their top-notch sound, which adds so much to the natural feel of a videoconference. An important part of that excellent sound is a proprietary audio coding technique called PT724™.

This technique reduces the line requirements for wideband sound to only 24 kbps, freeing up an additional 24 to 40 kbps of the line's capacity for video. So everything looks and sounds better.

ECHO CANCELLATION

All videoconferencing systems must cope with echoes. Although echo cancellation is fairly easy to think about, it's hard to do, and some systems do it better than others. PictureTel systems use a refined approach called Integrated Dynamic Echo Cancellation, or IDEC™, that does a virtually perfect job.

It's "dynamic" because it automatically adjusts as needed if the acoustics change—for example, if people walk around or if someone moves the microphone. Less adept systems need manual retraining if the acoustics change.

SMART MICROPHONES

Problem: A microphone sensitive enough to pick up a voice across the room also picks up noises like shuffling papers and whirring fans.

Solution: Make the microphone directionally sensitive, and point it at the person speaking.

Problem: Who's going to point it?

Solution: Make the microphone smart enough to point itself—electronically, with no moving parts.

PictureTel's PowerMic™ does just that. It contains not one but several microphones, each one sensitive to sound from a different direction. The circuits in the PowerMic figure out which direction the strongest sound is coming from, and send only that sound to the system.

In this way the microphone "points" itself at the person speaking and mutes other sounds. It even mutes the reflections of the speaker's voice that bounce off the walls, effectively improving the room's acoustics.

SMART CAMERAS TOO

The pan, tilt, and zoom controls are easy enough to use. But what if you could forget the controls altogether, and simply tell your camera to turn automatically to the current speaker?

That's what PictureTel's LimeLight™ does. Not only does it find the person speaking and turn the camera in that direction—it's even smart enough to zoom in on a single speaker or zoom out to show several people conversing.

THE CIRCUITS IN THE POWERMIC FIGURE OUT WHICH DIRECTION THE STRONGEST SOUND IS COMING FROM, AND SEND ONLY THAT SOUND TO THE SYSTEM.

SOME BUZZWORDS

Artifact An undesirable glitch in the video you see during a videoconference. One example is a mosaic-tile effect, called tiling.

Bandwidth The range of signal frequencies that a system or line can handle. In the case of audio bandwidth, it's a measure of sound fidelity.

Codec Coder-decoder. The coder transforms live video and audio into computer data for sending to another videoconferencing system. The decoder converts received data back into live video and audio.

Compression A process for squeezing live video and audio through low-capacity networks, using advanced mathematical techniques called compression algorithms.

Data communication Communication between computers and other meeting tools during a videoconference, enabling a richer and more natural meeting.

Dial-up connection The kind of connection you establish by simply dialing a number, as opposed to leasing a dedicated line. Also called a switched connection.

Digital phone line A line that carries information in digital form (the 1s and 0s of computer data), in contrast to an ordinary phone line that carries analog signals (electrical vibrations).

Echo The result of your voice being picked up by the microphone at the far end and transmitted back to you. Echo cancellation reduces or eliminates this echo.

Frame A single still picture in a stream of live video. Videoconferencing typically uses a frame rate of 15 or more frames per second.

G.711, G.722, G.728 Standards for coding and decoding narrowband (G.711 and G.728) and wideband (G.722) audio; part of H.320.

H.261 A standard for coding and decoding video; part of H.320.

H.320 A suite of standards that covers videoconferencing over ISDN and other kinds of digital phone lines. Also called p×64.

IDEC Integrated Dynamic Echo Cancellation. A PictureTel audio technology.

Interoperability The ability of one videoconferencing system to conduct a video call with another system by conforming to international standards.

ISDN Integrated Services Digital Network. A common kind of dial-up digital line with two 64-kbps data channels.

LimeLight PictureTel's automatic camera-pointing device, which points the camera at the current speaker.

Motion handling The ability of a videoconferencing system to code and decode video so that the resulting picture shows smooth motion.

Narrowband audio Audio similar to telephone sound, with a bandwidth of 3 kHz.

Pixel Picture element. One of the thousands of dots that make up a frame.

Postprocessor A device for removing artifacts from decoded video before displaying it.

PowerMic PictureTel's "smart" microphone.

PT724 PictureTel's technique for efficiently transmitting wideband sound.

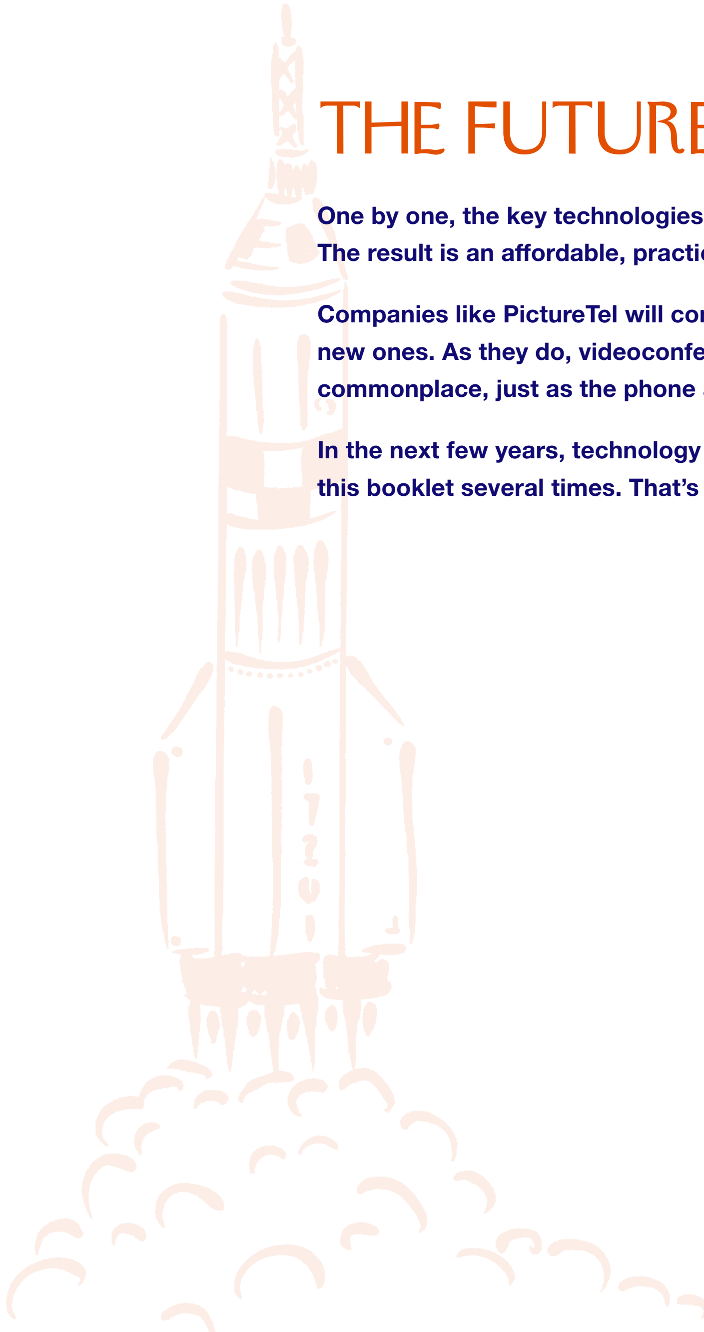
P×64 A nickname for H.320. Pronounced "p times 64."

SG4 Software Generation 4. PictureTel's advanced video compression algorithm.

Standards Technical specifications that enable systems from different manufacturers to interoperate.

T.120 A standard for data communication during a videoconference; part of H.320.

Wideband audio Audio superior to telephone sound, with a bandwidth of 7 kHz.



THE FUTURE

One by one, the key technologies of videoconferencing have fallen into place. The result is an affordable, practical business tool.

Companies like PictureTel will continue to refine existing technologies and invent new ones. As they do, videoconferencing will become more sophisticated and more commonplace, just as the phone and the fax machine have done.

In the next few years, technology may advance so much that we'll need to rewrite this booklet several times. That's OK—we're the ones who can do it.

WHERE DO I GO FROM HERE?

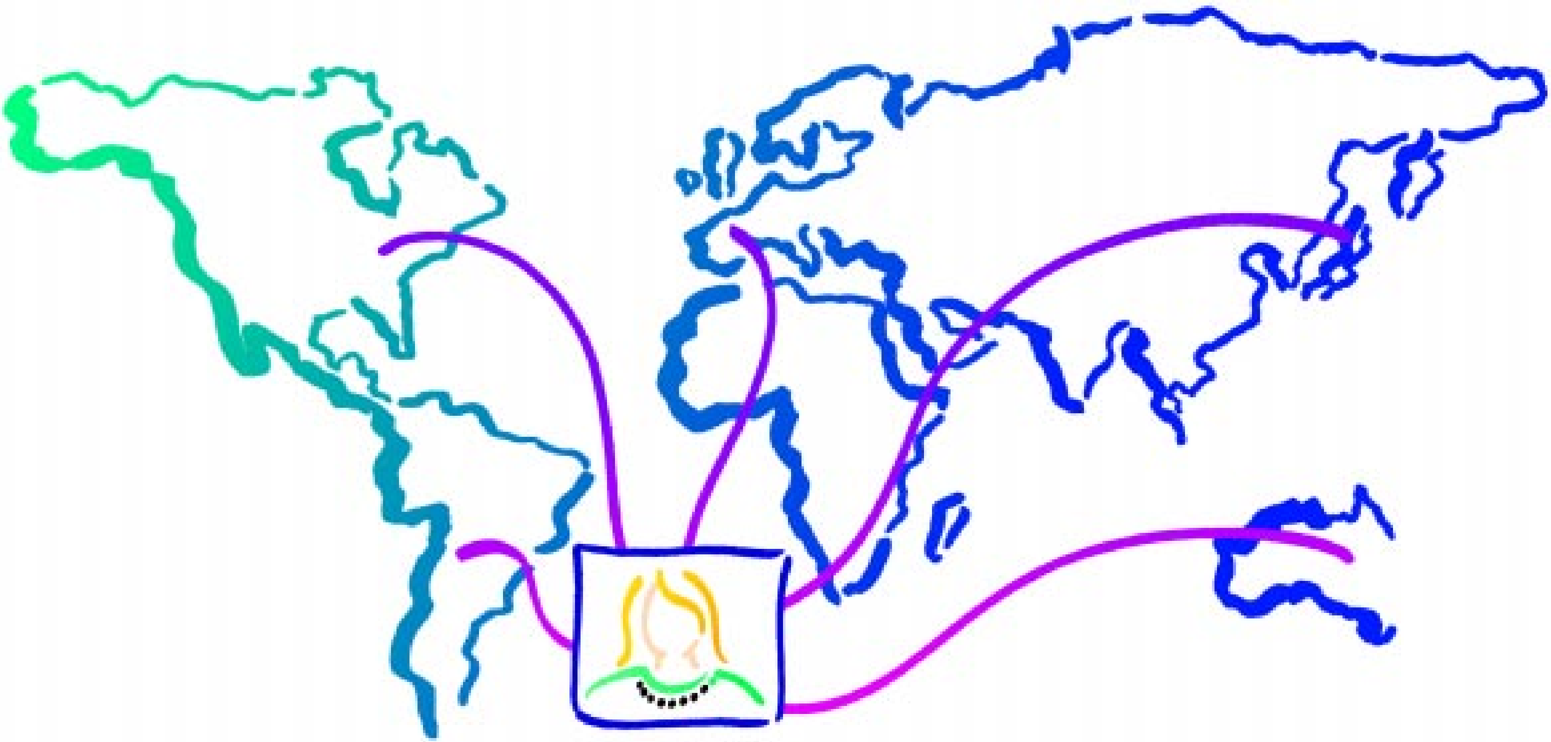
For more information, talk to your local PictureTel representative, or call:

Worldwide (to USA): 1-978-292-5000
Within USA (toll-free): 1-800-716-6000

You can also visit the PictureTel web site at:

www.picturetel.com

How does videoconferencing work?



PictureTel

videoconferencing 101

PictureTel Corporation · 100 Minuteman Road, Andover, MA 01810 · www.picturetel.com

VC101-7/0198/0k Edition 800-0463-02/A