

H.323

M. Di Donfrancesco

H.323 with respect to H.320

H.323 transmission protocols

H.323 Entities

Audio, video and data codings

The connection protocol

Packet Based Networks, also known simply as IP Networks, were initially introduced for data transmission, however the increasing success of technologies like **Voice over IP** (*private networks with audio and data integration*) and **IP-Telephony** (*public audio networks*), made them the *de facto* standard for Telecommunication applications.

Way back in January 1996, a group of computer and digital network manufacturers, at the time deeply involved in the development of applications and multimedia audio-video products for IP networks, brought to the fore the urgent need for an instrument capable of solving the growing interoperability problems. It became evident that a standardisation was required for the various (*proprietary*) techniques adopted to solve the intrinsic limitations in the Quality of Service (QoS) that affected the architecture of PBN networks, when used for real time applications.

Initially the attention was focused exclusively on LANs (Recommendation H.322), as due to their the reduced dimensions, they were easier to control; but before long in a short time the tremendous diffusion of the Internet increased the urgency for a unique standard for covering all the PBNs.

In May 1997, the Working Group XV of the International Telecommunication Union (ITU) published the ITU-T H.323 Recommendation, which was followed by a second revision just one year later (1998).

Although often called the “H.323 standard” or “H.323 protocol”, H.323 is neither a standard nor a protocol. H.323 indicates a collection of devices, procedures and protocols (some mandatory, others

optional), which guarantees interoperability between different products and applications over different networks, comprising analogic Switched Circuit Networks (SCN) and digital Integrated Service Digital Networks (ISDN).

H.323 with respect to H.320

The H.323 Recommendation is a component of Recommendation Family, which defines the specifications for the provision of multimedia services over different network topologies.

Other members of this family include the H.324 Recommendation, for the analog *General Switched Telephone Network*, the H.320 Recommendation, for digital ISDN networks, the H.321 Recommendation, for broadband B-ISDN networks, and the H.322 Recommendation, for LANs with guaranteed QoS.

H.323 does not define a new functionality and services topology; it is simply an extension of the previous H.320 Recommendation published by the *CCITT (Consultative Committee for International Telephony and Telegraphy)*, which in December 1990 became the ITU-T.

The extensions introduced by the H.323 Recommendation came both in the form of improvements in the signalling and compression techniques, which had been developed during the same period, and the exploitation of the different characteristics of PBN networks compared to SCN networks.

ISDN networks, as referred to in the H.320 Recommendation, provide a synchronous, guaranteed and bi-directional data flow: once the connection has been established, the bandwidth is guaranteed in both directions (*full-duplex*) for the entire duration of the connection. The result is a

temporally and sequentially predictable data stream.

Over IP networks, on the other hand, the information is divided into fixed length packets and transmitted asynchronously through pseudo-randomic paths.

Each packet is subject to different and unpredictable delays (and in some circumstances packets are lost) and the effective data rate can change on the basis of the instantaneous bandwidth availability (that essentially depends on the number and the type of the devices that are contemporarily active on the network).

IP networks transmissions are mono-directional (*half-duplex*) and to achieve the same performances as a 128 kbps ISDN line a 256 kbps IP network must be used.

Like the previous H.320 Recommendation, the H.323 Recommendation defines only the audio service (G.711 audio coding) as mandatory and leaves the video and data services as optional, but these, however, if they are implemented, they must support the H.261 (video) and T.120 (data) coding, respectively.

Therefore the signalling and control protocols and also the audio, video and data algorithms are common to both the H.323 and H.320 Recommendations.

To guarantee the interoperability between different networks, especially when connecting over a bandwidth of less than 64 Kbps, the H.323 entities can even implement other coding and decoding algorithms.

H.323 transmission protocols

PBN networks are intrinsically affected by a series of limitations that become critical when used in real time applications.

Essentially these limitations are:

- the superimposition of signals due to the (constant) transmission delay caused by the coding, decoding, packetisation and buffering algorithms;
- the losses in synchronisation due to the (non-constant and unpredictable) delays the packets experience, (while being transmitted on different paths), before arriving at their destination (*jittering*).

Additionally, when the protocol used for the data transmission is TCP (Transmission Control Protocol), the extent of the above-mentioned limitations is made worse by the relevant overhead introduced both by the packets transmission confirmation and lost packet retransmission algorithms, as implemented in this protocol.

Basically, the critical aspect of the data transmission over PBN networks is therefore the Quality of Service (QoS) for the audio and video streams (real-time traffic with restrictive requirements).

The QoS can be described as a kind of contract laying out a set of parameters, whose values are negotiated between two entities respectively named as service carrier and service customer.

Parameters typically included in this type of contract are: connection delay, bandwidth amplitude, maximum number of packets lost, transmission delay and packet priority. The H.323 Recommendation assigns the multimedia data transportation to the UDP (User Datagram Protocol).

Although less reliable than other protocols, the UDP allows an higher temporal

predictability. The implementation of the *end-to-end* services between the multimedia applications is, instead, entrusted to the RTP (*Real Time Protocol*) and RTCP (*Real Time Control Protocol*).

In detail, the UDP layer is responsible for the multiplexing and checksum services, while the RTP is deputed to the services of identifying the type of charge, numbering the packet sequences, temporal identification, and monitoring and managing the packet delivery.

The RTCP protocol, is responsible for media transport control, QoS monitoring and traffic congestion management.

Audio and video transmission does not strictly require the reception of any single packet, a trait typical to digital data transmission. What is, on the other hand really essential, is the reception of a

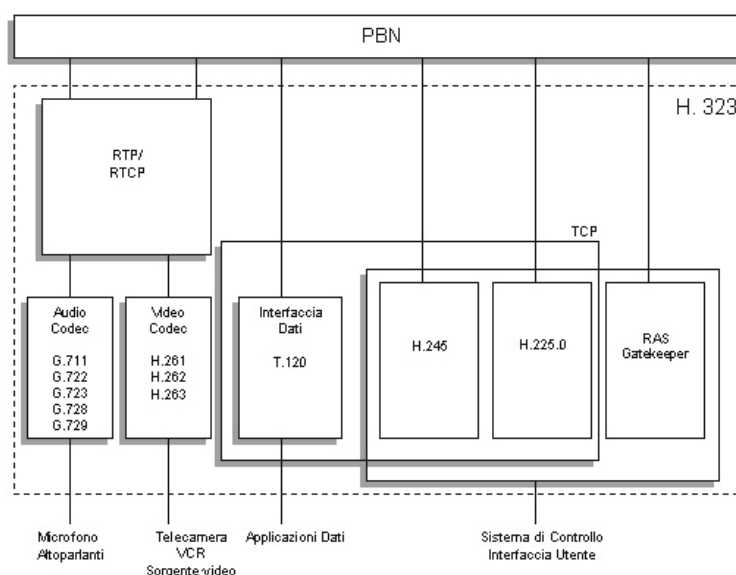
continuous stream of almost all the transmitted packets and in particular the packets transmission sequence.

The RTP protocol header, in particular, has a dedicated field used to carry the information on the precise moment the packet is sent (*timestamping*).

The timestamping is then used by the receiving H.323 device in order to know when the packet was expected and to verify the correct sequence.

supplies the addressing, authorisation and authentication services for both H.323 Endpoints and H.323 Gateways.

In addition, the Gatekeeper is responsible for the accounting and billing services and for call forwarding (e.g. it converts the E164 telephonic numbers into the corresponding IP addresses, if necessary).



H.261 and H.263 video coding

A frame is a sequence of pixels, grouped in vertical and horizontal lines, coded in their luminance (Y) and chrominance (CR, CB) components. A video stream is a sequence of frames transmitted at a certain rate over a physical medium.

Basically, video compression algorithms (video codings) use spatial redundancy (identical pixels sub-sets, or areas, in the same frame) and temporal redundancy (identical, or almost identical, consecutive frames) to reduce the total amount of information to be transmitted.

H.261 is the name of the video coding the ITU-T defined as mandatory for the H.323 Recommendation. H.261 applies to bit rates multiple of 64 kbps and it is the only video coding required in order to allow an H.323 video Endpoint to be compliant to H.323. Introducing the H.261 video coding, the ITU-T defined also the Common Intermediate Format (CIF, 352x288 pixels) as the basic frame format (resolution) and the QCIF (Quarter CIF, 176x144 pixel) as the mandatory resolution.

H.261 requires each frame, called Picture, to be hierarchically organised in Blocks (squares of 8x8 pixels), Macro-Blocks, MB (4 adjacent Blocks), and Group of Blocks, GOB (3x11 adjacent MBs). CIF and QCIF Pictures can be therefore divided in 12 and 3 GOB, respectively.

To remove temporal redundancy, H.261 demands the frames to be divided in I-frames, that are coded independently from any other frame, and P-frames, that are coded in correlation with the previous frame.

Frame classification is performed at MB level calculating the absolute value of the differences between each MB of the frame to be classified and the corresponding MB of the previous frame. To reduce to the minimum the information to be transmitted, the algorithm used for the calculation basically takes into account not only the single MB, but also all the adjacent MBs in the area given by the translation of 15 pixels in all the 4 directions (Integer Pixel Motion Estimation).

A frame is classified as an 'I-frame' if the absolute value of the differences is higher than a threshold, as a 'P-frame' otherwise. H.261 states an I-frame to be entirely transmitted and a P-frame to be transmitted in terms of the bidimensional vector (motion vector) of the differences with respect to the previous frame. By the way, to avoid an excessive deterioration of the image, a maximum number of 132 P-frames for each 2 I-frames is allowed by the H.261 coding.

The H.263 video coding shares with H.261 the same frame hierarchical organisation but it introduces some relevant differences in frame subsetting and classification techniques basically addressed to the optimisation of the video quality in case of bit rate lower than 64 kbps.

H.263 introduces a new frame format, called SubQCIF (128x96 pixels), and a new frame classification technique, called Advanced Prediction Mode (APM); this particular feature allows the use of four motion vectors (Half Pixel Search) for each MB, when classifying the frame as 'I' or 'P'.

H.263 differs significantly from H.261 also at GOB level. The H.263 GOB is composed of 11 MBs when the frame format is the CIF, and 8 MB when the frame format is the QCIF or the SubQCIF.

At the MB level, H.263 demands to transmit only a single bit (COD), instead of a VLC indicating the MB Address (MBA) in correspondence of a P-MB. As for Blocks the main difference between H.263 and H.261 resides in the quantisation algorithm.

H.323 Entities

The H.323 Recommendation defines four different categories of devices, or *entities*, that, once connected to each other, allow multimedia communication: *Endpoints*, *Gateways*, *Gatekeepers* and *MCUs (Multi Point Control Units)*.

The communication (*conference*) between H.323 entities is accomplished by the

transmission of audio, video and data streams, each with their own control signals, which are sent with the coded digital data. The H.323 Endpoint entity, defined by the recommendation as the only entity in the network that can start and receive calls, is the H.323 basic unit: it is the H.323 Endpoint that communicates in real time and bi-

directionally with other H.323 entities. The H.323 Recommendation states that a H.323 Endpoint must always supply an audio service, which can optionally be integrated with the supply of video and/or data services. The audio codec, and in some cases the video and data codecs, must therefore be implemented at the H.323 Endpoint level of the H.323 architecture, together with the

corresponding control algorithms. An H.323 Endpoint, for example an IP phone or Personal Computer, must guarantee compatibility with H.324 Endpoints connected to SCN or to 3G wireless networks, or with H.320 Endpoints, connected to ISDN networks, and with H.310 and H.321 Endpoints, connected to B-ISDN networks.

The H.323 Gateway is defined by the H.323 Recommendation as the entity that allows intercommunication between H.323 Endpoints when connected to PBN networks, or compatible endpoints connected to different networks, like for instance a SCN networks. An H.323 Gateway makes intercommunication possible by transcoding the *call-setup* and *call-release* protocols and properly converting the data format when the data packets go from one network to the other.

The H.323 Gatekeeper is defined by the H.323 Recommendation as the entity that supplies the addressing, authorisation and authentication services for both H.323 Endpoints and H.323 Gateways.

In addition, the Gatekeeper is responsible

for the accounting and billing services and for directing of the calls (it converts the E164 telephonic numbers into the corresponding IP addresses, if necessary). The H.323 Recommendation even states that the H.323 Gatekeeper is responsible for the monitoring and the management of the bandwidth state and instantaneous utilisation. Usually a H.323 Gatekeeper manages a H.323 *zone*, defined as a number of H.323 entities grouped according to a specific logical criterion that means they refer to a unique gatekeeper, independent of their position collocation in the effective network topology.

The H.323 MCU is defined by the H.323 Recommendation as the entity that implements communication between three or more H.323 Endpoints (*multipoint conference*).

The services assigned to the H.323 MCU, include monitoring the various resources of the multipoint conference, the negotiation of the audio-video-data coding to be used by conference participants, and management of the multimedia data streams.

The H.323 MCU receives the audio and

video streams from all the endpoints involved in the conference and sends back to each one a unique stream containing either the audio and video streams of each participant (*Continuous Presence, CP*), or the audio and video streams of the active speaker (*Voice Switching, VS*).

In CP mode, the audio stream is simply obtained by sending to each participant the sum of the audio signals received from all the other participants; the video stream is, instead, obtained by inserting (often after reducing the format) the video stream of each Endpoint into a unique video stream, which has the format agreed on during the multipoint conference set up process.

Inside the H.323 MCU, the audio-video-data stream is managed by the *Multipoint Controller (MC)*, while the management of the single control signals is assigned to the *Multipoint Processor (MP)*.

In the H.323 Recommendation, H.323 Gateway, H.323 Gatekeeper and H.323 MCU are distinct entities, but nothing prohibits them from being implemented in a single device.

H.264/AVC video coding

H.264 defines new and efficient coding algorithms and it is specifically designed for transmission over networks that may differ for both bandwidth and transportation layers. The H.264/AVC (Advanced Video Coding) architecture is composed of a Video Coding Layer (VCL), which efficiently represents video data, and a Network Abstraction Layer (NAL), which formats the VCL output and provides header information for conveyance on particular transport layers or storage media.

The NAL is responsible for video data formatting and header management. All data are contained in NAL Units (NALU), each of which containing an integer number of bytes. A NALU specifies a generic format matching both packet-oriented and bitstream transmission requirements. For both delivery modes, the format of the NALU is identical, with the only exception that each NALU can be preceded by a start code prefix in the bitstream-oriented transport layer. H.264 standard supports both progressive and interlaced frames, that, in addition, may be mixed together in the same sequence. The VCL uses a hybrid video-coding (in 4:2:0 chroma format) approach, based on inter-picture prediction, to exploit the temporal statistical dependencies, and a transform coding of the prediction residual to exploit the spatial statistical dependencies. There is no single coding element in the VCL that provides the majority of the dramatic improvement in compression efficiency, in relation to prior video coding standards.

Rather, it is the plurality of smaller improvements that add up to the significant gain. Substantially H.264 differs from previous video coding standards for the enhanced motion-prediction capability, the use of small block-size exact-match transform, the adaptive in-loop deblocking filter, the enhanced entropy coding methods. Each picture, either a frame or a field, is partitioned into fixed-size

macroblocks (MBs) that cover a rectangular area of 16x16 samples of the luma and 8x8 samples of each of the two chroma components. All luma and chroma samples of a MB are either spatially or temporally predicted, and the resulting prediction residual is subdivided into blocks. Each block is transformed using an integer transform, and the transform coefficients are quantised and transmitted using entropy-coding methods. The MB are organised in slices, which generally represent subsets of a given picture that can be decoded independently. The transmission order of MBs in the bitstream depends on the so-called Macroblock Allocation Map. H.264 supports five different slice-coding types. Intra-Slices are composed by MBs coded without referring to any other picture within the sequence (intra prediction). Predicted-Slices contain, on the other hand, MBs coded using inter-prediction with at most one motion-compensated prediction signal per prediction block. MBs coded using inter-prediction with two motion-compensated prediction signals per prediction block form the so called Bipredictive-Slices (substantially B-Slices are coded so that MBs or blocks are coded using a weighted average of two distinct motion-compensated prediction values). The above three coding types are very similar to those in previous standards with the exception of the use of reference pictures. The remaining other two slice-coding types are completely new. Switching P-Slices (SP-Slices) are coded so that efficient switching between different pre-coded pictures become possible. Switching I-Slices (SI-Slices) allow an exact match for a MB in a SP-Slice for random access and error recovery purposes. The H.264 standard introduces a new feature called Flexible Macroblocks Ordering (FMO) that allows to assign the MBs in a picture to several slice groups, so that each slice becomes an independently-decodable subset of a slice group. When used effectively, FMO can significantly enhance robustness to data losses by managing the spatial relationship between the regions that are coded in each slice. Every slice group is transmitted separately and if a slice group is lost, the samples in spatially neighbouring MBs, that belong to other correctly-received slice groups, can be used for efficient error concealment. In contrast with all major prior video coding standards, H.264 is based on a 4x4 transform (this allows the encoder to represent signals in a more locally-adaptive fashion, which reduces artifacts) and requires the prediction being conducted in the spatial domain by referring to neighbouring samples of already coded blocks. As for intra-frame prediction, two classes of intra coding types are supported, which are denoted as Intra-4x4 and Intra-16x16. When using Intra-4x4 mode, each 4x4 block of the luma component utilises one of nine prediction modes. When using Intra-16x16 mode, a uniform prediction is performed for the whole luma component of a MB. In addition to the Intra MB coding types, various predictive or motion-compensated coding types are specified for P-Slice MB. Each P-type MB correspond to a specific partitioning of the MB into fixed-size blocks used for motion description. The accuracy of motion-compensation is a quarter of a sample distance. The prediction values for chroma components are always obtained by bi-linear interpolation. B-Slices utilise an MB partitioning similar to that of P-Slices and support four different types of inter-picture prediction. The motion vector coding is also similar to that of P-Slices with the appropriate modifications because neighbouring blocks may be coded using different prediction modes. To be suitable to applications that differ for bit-rate, quality, services and resolution, the H.264 is provided of a simplified Profiles and Levels scheme. A Profile defines a set of coding tools or algorithms that can be used in generating a compliant bitstream, whereas a Level places constraints on certain key parameters of the bitstream. All decoders conforming to a specific Profile have to support all the features in that profile. Encoders are not required to make use of any particular set of features supported in a Profile but have to provide conforming bitstreams. In H.264 three Profiles are defined: Baseline Profile (for low-delay end-to-end applications), eXtendend Profile (for mobile applications and e-streaming) and Main Profile (for broadcasting application at SD level). The Baseline Profile does not support B-Slices, SP-Slices, SI-Slices and some other minor features. The Main Profile does not support the FMO feature. The X profile does not support some adaptive features. The same set of level definitions is used with all Profiles, but individual implementation may support a different level for each supported Profile. Eleven Levels are defined, specifying upper limits for the picture size (in MB), the decoder processing rate (in MB/s), the size of the multipicture buffers, the video bit-rate and video buffer size. To address high-end consumer and other applications that require high-resolution video without a need for extended chroma formats or extended sample accuracy some new Profiles (originally known as professional extensions) has been introduced. They are collectively called High Profiles: HP (based on 8 bits/sample and 4:2:0 chroma sampling), Hi10P (supporting 4:2:0 video up to 10 bits per sample), H422P (supporting 4:2:2 chroma sampling and up to 10 bits per sample), H444P (supporting up to 4:4:4 chroma sampling, up to 12 bits per sample and, additionally, supporting efficient lossless region coding and an integer residual colour transform for coding RGB video while avoiding colour-space transformation errors).

Audio, video and data codings

The H.323 Recommendation requires that the coding and decoding algorithms for audio, video and data streams are implemented by the H.323 Endpoints.

The audio coding defined as *mandatory* is the ITU-T G.711 coding (64 kbps *Pulse Code Modulation*, PCM, coding used with *Public Switched Telephonic Networks*, PSTN). H.323 Endpoints can *optionally* also implement other audio codings: the ITU-T G.722 (7 kHz *Adaptative Differential PCM* coding with a bit rate of 48, 56 or 64 kbps), the ITU-T G.723.1 (*MultiPulse-MultiLevel Quantisation* compression algorithm with a bit rate of 5.3 or 6.3 kbps), the ITU-T G.728 and G.729 (*Code Excited Linear Prediction* coding characterised by reduced delay and bit rate of 16 and 8 Kbps), the MPEG 1 and the GSM (13 Kbps coding used in mobile telephony).

The H.323 Recommendation defines video as optional, however, if a H.323 Endpoint supplies a video service, it is mandatory to implement the ITU-T H.261 coding.

This particular algorithm is especially suitable for the reduction of the constant and predictable delays introduced by the video stream coding and decoding procedures.

Another video coding allowed by the H.323 Recommendation is the ITU-T H.263, which is derived from the H.261 and introduces further techniques of frames prediction, particularly suitable for low bit rate transmissions.

Recently the H.264 coding has also been included in the family of video coding algorithms supported by the H.323 Recommendation.

The H.264 standard introduces innovative coding instruments more effective than previous ones and it is considered most suitable for video stream transmissions over networks with different bandwidth

amplitude, transportation layers, and error recovery procedures.

The H.323 Recommendation states that the data stream (data is intended as all information that is not audio or video) is totally independent from the audio and video streams: a data connection can be set up independently or when a conference is already active and its termination does not cause the conference to shut down.

If present, the data transfer has to be compliant with the T.120 Recommendation, that defines the entire gamma of protocols and services that must be used in real time multipoint transmissions.

The T.120 data transfer guarantees the reception of all the packets (packets are sent using the TCP protocol) and it is both platform and network independent.

The T.120 architecture is multi-layer, and each layer offers a specific service:

- the T.123 layer (*Transport Protocol*) offers a reliable and network independent transportation service for the *Protocol Data Units* (PDU);
- the T.122 and T.125 layers (*Multipoint Communication Services*) define respectively the available multipoint service and the data transmission protocol;
- the T.124 layer (*Generic Conference Control*) provides a collection of procedures suitable for the multipoint conference management;
- the T.121 layer (*Generic Application Template*) defines the modelling for the T.120 resources management procedures to be implemented at the application layer;
- the T.126 layer (*Still Image Exchange*) provides the still image exchange services between two or more applications (feature known also as *shared whiteboard*);
- the T.127 layer (*Multipoint Binary File*

Transfer) defines how to implement the file transfer.

The connection protocol

The connection protocol is managed by the H.323 Gatekeeper, which provides the connectivity services for a H.323 zone.

The first task of an H.323 Gatekeeper does is to open a RAS (*Registration, Admission, Status*) channel towards all the endpoints (H.323 Endpoints and H.323 Gateways) involved in the conference. All the registration, admission control, state change and disconnection messages for the conference travel over the RAS channel. Once all the endpoints have been registered and admitted, the H.323 Gatekeeper opens a second channel, located at the transport layer, called the *call-signalling* channel, for the transportation of H.225.0 control signals.

According to the H.323 Recommendation, the H.225.0 manages the set up of a communications between H.323 entities connected to the same packet switching network. The H.225.0 messages define the communication control architecture.

H.225.0 requires a reliable connection that guarantees the endpoints will receive the correct (not corrupted) information transported by each message.

The H.225.0 connection is therefore a TCP connection using a particular port (or *Transport Service Access Point*, TSAP), which is communicated to the endpoints by the H.323 Gatekeeper in the *Admission Confirm* RAS message. The port involved in this H.225.0 procedure is usually the port 1720.

An H.323 conference is considered 'set up' only once the sequence of H.225.0 ASN.1 (*Abstract Syntax Notation*) *Setup* (sent by the caller) and *Call Proceeding*, *Alerting and Connect* (sent by the called) messages have been exchanged.

Once the connection has been established, the exchange of H.245 messages sent over the *control channel* begins.

The H.245 protocol is responsible for managing transportation of the control signals between the conference participants, the definition of capabilities (i.e. the audio-video-data coding to be used during the conference), exchange procedures and managing the logic communication channels.

The logical communication channels are used for multimedia data stream transportation (the protocols used are the RTP and the UDP), to manage the bit rate of both the single channels and the overall channel (the channel comprised of all the logic channels), to measure the round-trip delay between two

endpoints and to define the master and the slave of the conference. The conference can be shut down by any of the participants. The disconnection protocol requires the interruption of the audio stream, the closure of all the logic channels, the shut down of the H.245 and H.225.0 sessions and the notification of the call release (via RAS message), in this exact order.

Conclusions

The H.323 Recommendation applies to only-audio-only (IP telephony), audio-video (videoconferencing), audio-video-data and video-data communications and requires the involvement of H.323 entities in point-to-point, multipoint (the streams are sent to a group of entities) or broadcast (the streams are sent to all

the hosts of an entire sub-net) connections.

Since the H.323 Endpoints can also have receiving functionalities only, the H.323 Recommendation allows the use of technologies such as video-on-demand and the presence of entities like the content providers that were not considered by the previous H.320 Recommendation.



Marco Di Donfrancesco
Aethra Product
Marketing Manager

REFERENCES

- [1] Kumar V, Sengodan S, Korpi M, "IP Telephony with H.323: Architectures for Unified Networks and Integrated Service", Wiley John & Sons Inc., ISBN 0471393436.
- [2] ITU-T Recommendation H.323 <http://www.itu.int/itu_t/>
- [3] Wiegand, Sullivan, Biontegaard, Luthra, "Overview of the H.264/AVC Video Coding Standard", IEEE Transaction on Circuits and Systems for Video Techonolgy, Vol 13, No 7, July 2003