Voice over IP Technology

2/4/2004

At World Telecom Labs, we have identified several limitations of the existing VOIP technology available on the market. We have addressed these one by one and finally produced a commercial product providing toll quality telephony service over an IP network.

Our technology is based on industry standard hardware and software. We came from the telephony service provider market whose characteristics are: high volume, digital switching, powerful least cost routing (LCR), fraud protection, great variety of interconnection protocols (SS7, ISDN, E1, T1,) and added value applications: prepaid, postpaid, automatic call distribution (ACD) and callback. When we started to work on VOIP technology three years ago, our approach was to implement VOIP as just another telephony protocol so that existing applications could use VOIP seamlessly. Our number one concern was high voice quality as that is what our customers are used to. Our number two concern was performance: our switch can carry high volumes (up to 20 million minutes a month and more using ATM technology) and it was a requirement that such high volumes could eventually go through VOIP although the market was still in its infancy. The third concern was interoperability with other equipment, so although World Telecom Labs' IPNx has a number of unique benefits it will operate perfectly in an existing H.323 environment.

The purpose of this paper is to explain in detail some of the solutions we found to address these challenges.

Voice quality improvements

The voice quality is determined by several parameters: algorithm, jitter on voice packets and end to end delay.

The algorithm

Several well known algorithms exist in the market: G.729A, G.723.1, SX7300 and SX9600 (both Lucent proprietary). Many companies also provide less common proprietary compression algorithms. At World Telecom Labs we support all the main standards but, for several reasons, we normally recommend G.723.1: it provides excellent voice quality at low bit rate (6.4 kb/s), it is used by the H.323 standard, it generates fewer packets (33 p/s). The G.723.1 standard is optimized for the human voice, other inband signals like call progress tones, DTMF, FAX are completely distorted by the compression algorithm and so must be specially handled (see below).

The compression technology we have selected includes many improvements to the G.723.1 algorithm such as a DTMF decoder/encoder, runtime voice/FAX switch, modem detection and bypass. Short silence frames are part of the G723.1

standard but are often ignored by other vendors. In contrast WTL makes extremely efficient use of them.

The short silence frames are important to the quality of the call: when the inband signal level drops below a certain level, the algorithm generates a 4 byte silence frame containing enough information to produce a comfort noise.
The alternative, used by H.323, consists in sending no packet as long as the signal level is below the noise level.
This causes a "dead line" impression to the listener as no signal is generated at the decoder until the talker speaks again. Failure to remove silence frames causes excessive bandwidth utilization, as the voice packets must be embedded into an IP frame which is bigger than the voice packet itself.
Note: because of the large IP overhead, the short silence frame is really beneficial when used in conjunction with NOP, another improvement invented at WORLD TELECOM LABS which is explained later.

The DTMF decoder/encoder is able to detect and transmit keypad activity and represent it accurately via IP messages. Both the content and length of the DTMF tone are handled. Another enhancement is that DTMF is encoded and sent as part of the packet header thus giving a further bandwidth saving.

The voice/FAX switch consists of automatically detecting a FAX signal (CED, HDLC flag) and switching to a FAX encoder/decoder for the rest of the call. The entire FAX transmission, including the T30 protocol will be demodulated and transmitted as data. This has numerous advantages over simply sending the FAX through a transparent 64K data channel: low bandwidth during T30 protocol (300 bit/s), half-duplex communication during image transmission, clean FAX signal generation even on an international call. In addition, the FAX encoder will automatically remove any non-standard facilities (NSF) which might be requested by the faxes. This improves the interoperability of the faxes, which might eventually not work when they are communicating directly to each other.

Modem traffic is also detected within the call, compression is switched off and currently the data is routed as a transparent 64K stream. A future development will allow this modem traffic to be demodulated and sent as data, thus taking less bandwidth.

The network delay and quality
The jitter is the variation on the transmission delay of voice packet through the network. The jitter is dependent on the usage of the network: if the network is also used for bursty data traffic, the jitter on the voice packets is likely to be very high. Because of the continuous nature of speech, the decoder needs a voice packet every 30 milliseconds exactly; the only solution to compensate the jitter is to pile the voice packets in a jitter buffer up to the maximum transmission delay. When the network delays the voice packets, the decoder uses the voice packets stored in the jitter buffer to ensure a continuous speech generation. The World Telecom Labs software adapts the jitter buffer automatically to the actual jitter on the network. Of course, a larger jitter buffer means a longer end to end delay.

There are two other important factors to mention which affect voice quality. Firstly, echo cancellation is vital to the user perception of voice quality. World Telecom Labs has a very effective echo cancellation implementation based on G165. Secondly, World Telecom Labs has implemented bi-directional, configurable gain control. This means that levels can easily be adjusted between the local E1 (or analogue connection) and the compression engine. This is available for both record and playback (send and receive).

World Telecom Labs recommends that the voice packets are sent on a dedicated IP network with a low transmission delay and jitter. However, this is not always possible and special facilities are available for situations where the network link is more unreliable. The IPNx may be set up to regularly probe the link used to carry the VoIP traffic. If the link is discovered to be down or experiencing too much congestion and delay the IPNx can bring an emergency back up link into service (for example, a dial-up ISDN line) to ensure that some level of service is maintained. The sampling of the impaired line continues so that when it recovers traffic may be switched back to the original connection. The IPNx can fit well into whatever QoS (Quality of Service) or packet priority scheme a network operator chooses. VoIP traffic in the IPNx always uses a fixed IP port number which means that it can easily be identified and mapped in a MPLS, Packeteer or other router-based QoS environment. It will also possible to make use of the ToS bit where this is appropriate.

Performance improvement

A complete VOIP solution is made of many modules: the switching module, the compression module, the packaging module, and the network. The traditional VOIP solution separates the switching and compression module and the packaging module is therefore not under control of the system integrator. With this architecture, a telephony switch is connected to a compression device which converts an E1 trunk into data packets. This solution has many disadvantages: additional cost due to the duplication of E1 interfaces, loss of SS7 call information due to non-SS7 E1 interface, no network optimization because of the unsophisticated packaging module. Most compression devices will also have network size limitations due to the limited number of IP routes they can support. In addition, these external compression boxes are not designed for large network design. It is often difficult to build large nodes by stacking several boxes, as there is no automatic load distribution.

At World Telecom Labs we have chosen to integrate the switching module, the compression module and the packaging module into the same chassis. The compression is achieved by using a specialized DSP board, which gets the uncompressed channel from the internal TDM bus.

The compressed data is not sent to the network directly but to the host via the PCI bus. Although this architecture increases the load on the host CPU, it gives us a great deal of flexibility when implementing a powerful packaging module, which is the most important invention of the World Telecom Labs solution.
A single Pentium III processor is able to support 1024 compressed channels which is more traffic than most single chassis can handle (more than 32 E1s). However, even this limit is easily overcome by intelligent load distribution software embedded in the packaging module.

For the network interface, we simply use standard LAN cards, which provide the best performance/cost ratio, and leave the IP switching to the specialized routers. Using IP on Ethernet is the most universal form of transport and therefore gives the maximum freedom for the long distance link (for example, satellite, leased line, ATM or intranet). Using another network interface exclusively (frame relay, ATM) would restrict the number of possible network topologies, limit the network size and would not allow the use of standard routers. The packaging module that we have developed at World Telecom Labs takes the voice packets produced by the compression board and sends them to the LAN card via a TCP/IP stack. The unique features of this module are: SS7 compatible call control, frame bundling and payload switching.

Flexible Call Control

The importance of an SS7 compatible call control layer is obvious when the VOIP network is inserted between several SS7 or ISDN networks. The seamless integration of a VOIP solution into a telephony network is only possible if the SS7 information elements can be transported transparently through the network. External compression devices use, at best, ISDN signalling which passes a limited set of information. Even standards like H.323 only support basic call information.

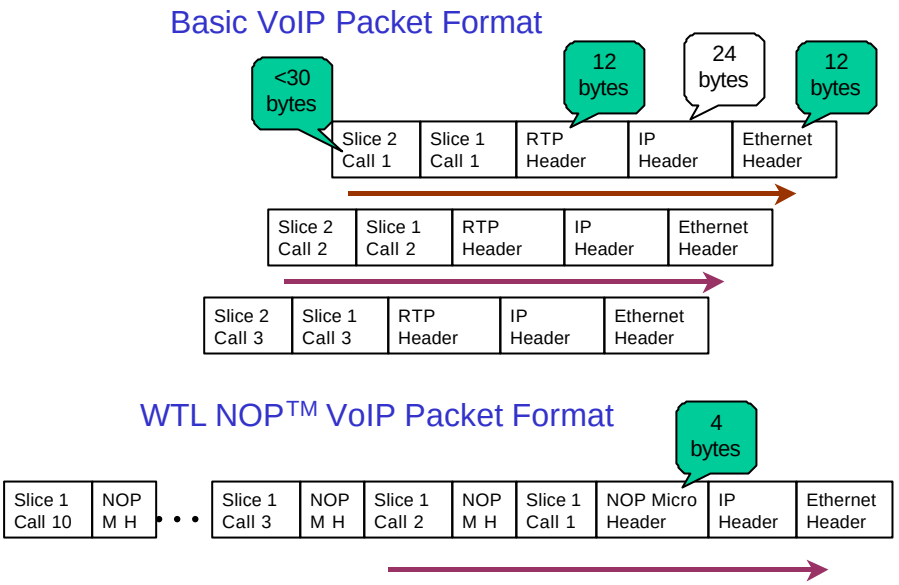
At World Telecom Labs we have developed a flexible, adaptable call control protocol based on UDP which is compatible with SS7 messages. Any information received from the switched network can be passed transparently through the IP network. Using UDP allows us to be independent of OS limitations like the number of TCP sockets that can be created at one time and to establish a call much faster than a TCP connection. These problems exist with H.323 which uses TCP as the basis for the call control protocol. Of course, using UDP requires detection and retransmission of lost messages. This quality of service layer is built into the packaging module and includes an automatic IP link test mechanism as in the SS7 protocol: every IP route being used is tested by sending regular test messages. When the test message fails, all the calls placed on that route are automatically closed and appropriate re-routing takes place in the switch. This technique allows a fast reaction on the network failure compared to TCP-based call control protocols.

To support large nodes made up of multiple chassis, the call control protocol includes an automatic call distribution mechanism: when a new call arrives at a multi-chassis node, the first call setup message is sent to a master chassis, which automatically forwards the message to one of the slave chassis. The remaining call control messages will be exchanged with the selected slave. This method is much simpler than the gatekeeper method used by H.323 and guarantees an automatic load distribution inside a large node.

NOP (Network Optimisation Protocol)

NOP consists of assembling several voice packets that are going along the same route into larger IP packets. This is the most important feature to improve performance as the CPU load and the network load depends mainly on the number of packets and not the packet size. Less packets means less interruptions from the LAN card, less IP headers to check, less allocated buffers, less context switching in the OS. It should also be noted that the same is true for the routers whose performance is influenced only by number of packets not by their length. Use of NOP can therefore postpone the need to upgrade to expensive, high packet-per-second throughput routers

But moreover, the bandwidth utilization is far better: the IP header of 40 bytes is replaced by a small 4 byte header which is inserted before the voice packets inside the large IP frame. This header identifies the type of packet (voice, silence, FAX, DTMF,), the size of the packet (a silence packet is smaller than a voice packet) and the destination channel. 4 bytes is enough because all the necessary voice packets combined in an IP frame are exchanged between the same IP addresses, therefore the number of channels that must be identified is finite (currently 4096 channels).

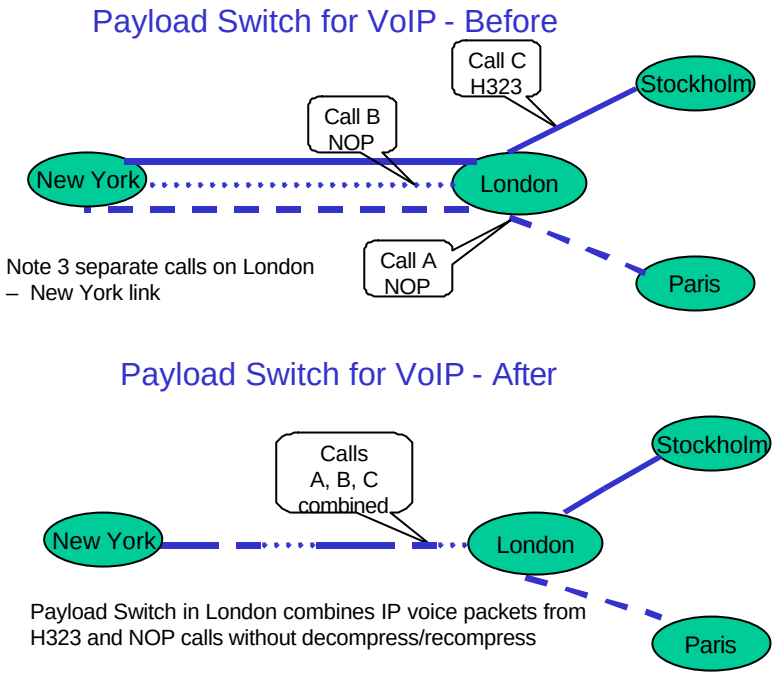


Note: Combining several voice packets does not increase the transmission delay as all these packets are generated at the same time by different channels. When 10 channels or more are combined, the IP overhead plus the 4 bytes header on every voice packet, together with the short silence frame, brings the overall bandwidth to 6.4 kb/s per channel, the actual compression algorithm rate. This outstanding performance allows us to put 10 toll quality calls on one 64 kb/s data channel while other products cannot support more than 5 or 6 calls. Whilst doing this the other products suffer longer delays (because they wait to pack two voice samples from the same channel) and higher packet per second load.

Payload Switching

The payload-switching module is closely integrated with the frame bundling module. The payload is the usable content of the IP frame; the voice packets with their small 4-byte header. Payload switching means 'the action of de-assembling one or more IP frames and re-assembling them into one or more other IP frames'. The way that WTL have done this means that many of the benefits of traditional TDM switching are available to packet-based traffic. Packet based voice traffic from different sources can be switched and multiplexed like TDM traffic but without the delay and quality loss associated with a decompress/recompress stage.

Payload switching for example allows the bridging of two or more IP routes: In the diagram below, to send a call to New York, a chassis in Paris could send the call to London. Inside the London node, the voice packets coming from Paris bundled in IP frames will be switched and combined with other IP frames going to New York (Calls B and C). This is equivalent to TDM switching, but in the packet space.
This technique also works for an H.323 device which can send a call to an **IPNx** chassis, which will convert it into the **IPNx** data format and insert it into the payload-switching module with other H323 streams or with NOP calls.



Interworking with H323

An H323 trunk is normally a public IP connection to other H323 entities: terminal equipment (NetMeeting station, IP phones) and other gateways. **IPNx** supports H323 Version 3 to allow these types of devices to be connected to an **IPNx** network and for them to gain the benefits of NOP and Payload Switching. You would normally not use the H323 protocol to interconnect **IPNxs** as the H323 protocol requires much more bandwidth than the NOP protocol and gives a lower voice quality. The authentication of H323 calls is done using their IP address or through a PIN code appended to the telephone number. The **IPNx** does not yet implement the sophisticated authentication methods that are specifically developed for H323 networks. To allow quick and easy set up, World Telecom Labs' VoIP solutions are designed to work without the cost and complexity of a separate H323 Gatekeeper. All the required Gatekeeper functions are built into an IPNx network: it resolves IP routing, authenticates calls, maintains CDRs, handles packet priority and deals with address translation between IP addresses and E164 standard international phone numbers.
The **IPNx** payload switching feature can be used to cross connect H323 calls with NOP calls without decompression. This configuration does not require any specialized hardware (low cost) and reduces the bandwidth usage by 40% without loss of quality. The H.323 call could equally be decompressed and converted into a TDM or ATM call.

'**IPNx**' is a registered brand name and trademark for World Telecom Labs Intelligent Switching Platform