INTERNET telephony or Internet TELEPHONY?

Henning Schulzrinne

June 13, 2004

Abstract

Internet telephony (or VoIP) has been viewed in two different ways. First, some hold that it is largely similar to earlier architectural changes in the PSTN, continuing the transition from analog to circuit-switched digital and now to packet-switched transmission. Others believe that Internet telephony is primarily another Internet application that happens to interconnect with a "legacy" network. In this note, we argue that the latter is closer to technical reality, as Internet telephony is largely indistinguishable from other Internet applications at all layers of the protocol stack and that the integration of Internet telephony and other applications is likely to increase rather than decrease, while similarities to the existing voice network are likely to be transition phenomena.

- Voice traffic is essentially indistinguishable from other data traffic, particularly if the senders of such traffic have incentives to mask the true nature of the traffic.
- Signaling and media flows can traverse completely disjoint network paths, with signaling servers located essentially anywhere in the world relative to the caller or callee. The functionality of common channel signaling has been split into two completely independent functions, with the media path setup usually unnecessary due to the connectionless nature of IP networks.
- With the advent of presence and instant messaging, the limitations of ring-and-answer telephony will become apparent, leading to a transition to rich presence and negotiated communications.
- Calls are likely to contain a mixture of voice, video and data streams, added and removed dynamically within a session.
- The notion of a call as a symmetric, duplex voice session of relatively short duration may vanish as modes closer to push-to-talk, intercom or continuous voice monitoring take over.
- Calls are likely to be embedded into other applications, including interactive computer games, rather than being placed by manual dialing.
- The centrality of telephone numbers will diminish, replaced by email-like identifiers that avoid the problems of scarcity. Telephone numbers will no longer necessarily identify a PSTN-connected device.

- Even ignoring the benefits of broadband Internet service, switching from narrowband Internet and phone service to broadband Internet service and VoIP is becoming financially attractive and thus, VoIP may become a driver for broadband deployment.
- Regulation needs to take into account protocol layering and the related interfaces, as this is a fundamental property of all modern communication technologies.
- Many users of voice services will be able to provide their own voice services, similar to how they provide their own email or web services.
- Lower-layer providers can interfere with the provision of upper-layer services.
- Emergency services are not primarily voice services, and their efficient delivery requires cooperation from the providers of physical and possibly network layer services.
- Burdensome regulations, e.g., for emergency services or lawful intercept, may encourage regulatory evasion by locating services outside the geographic reach of the regulatory authorities.

1 Introduction

The International Engineering Consortium (IEC) describes Internet Telephony as follows:

> Internet telephony refers to communications services – voice, facsimile, and/or voice-messaging applications – that are transported via the Internet, rather than the public switched telephone network (PSTN) . The basic steps involved in originating an Internet telephone call are conversion of the analog voice signal to digital format and compression/translation of the signal into Internet protocol (IP) packets for transmission over the Internet; the process is reversed at the receiving end.

More technically, Internet telephony is the real-time delivery of of voice and possibly other multimedia data types between two or more parties, across networks using the Internet protocols, and the exchange of information required to control this delivery.

The terms Internet telephony, IP telephony and voice-over-IP (VoIP) are often used interchangeably. Some people consider IP telephony a superset of Internet telephony, as it refers to all telephony services over IP, rather than just those carried across the the Internet. Similarly, IP telephony is sometimes taken to be a more generic term than VoIP, as it de-emphasizes the voice component. While some consider telephony to be restricted to voice services, common usage today includes all services that have been using the telephone network in the recent past, such as modems, TTY, facsimile, application sharing, whiteboards and text messaging. This usage is particularly appropriate for IP telephony, since one of the strengths of Internet telephony is the ability to be *media-neutral*, that is, almost all of the infrastructure does not need to change if a conversation includes video, shared applications or text chat.

VoIP is not a protocol or single network service, but rather a description of an aggregated, user-visible system that incorporates a plethora of technologies and protocols, namely network transport, media transport, signaling and directory services.

Network transport: VoIP leverages the same basic Internet protocols that all other Internet-based services, such as email or web, use, including IP as the network layer, UDP and TCP as end-to-end transport layers, DNS for name-to-address resolution and DHCP for end system configuration, in addition to the core network protocols used for routing, network management and other tasks. VoIP, like other services, may make use of differentiated or guaranteed services, where certain packets are either given priority treatment or are set aside bandwidth resources to avoid statistical interference from other services. These services can be provided by multiple different entities. For example, it is quite common to have the authoritative name server for the DNS domain of a corporation be provided by a hosting service rather than the ISP.

Media transport: The media transport functionality (Section 2) carries media information between end systems, including media servers. Typically, RTP is used for this functionality for conveying audio and video. The protocol is generally only visible to end systems and possibly firewalls, not routers. RTP is media-neutral, i.e., it does not need to know about the details of each media encoding.

Signaling: Signaling (Section 3) establishes, manages and removes sessions between two or more end systems, where session can comprise any data exchange, not just audio and video. As discussed in more detail in Section 3, the traditional dial-and-ring interaction is being supplemented with context-aware session initiation, where presence information provides hints as to whether a call attempt is likely to be successful or appreciated by the callee.

Directory services: Finally, directory services provide mapping between names and lower-layer protocol identifiers, such as SIP URLs. Since they are currently uncommon and are no different than other directories, e.g., those using LDAP for email address lookup, we will not consider them further. ITU Recommendation H.350 integrates SIP URLs and traditional email and phone directory information into a single LDAP record.

Some have treated VoIP has a straightforward extension of traditional voice telephony, continuing the evolution in voice transport from analog and digital circuit-switched facilities to a packet-switched network, i.e., essentially changes the multiplexing mechanism. Similarly, VoIP signaling can be seen as the next step in the evolution from channel-associated signaling to common channel signaling and now session signaling that splits media path and application-layer signaling.

In this paper, we make the point that VoIP is primarily another network service, similar to email, instant messaging and the web, and is indeed very likely to be integrated into systems that combine these core services into a single user-visible application.

2 Physical, Network and Transport Layer

While hardly a new notion, it bears repeating that it is all but impossible to reliably distinguish voice calls from other non-voice services on IP-based networks.

While most existing voice applications have distinctive fingerprints enumerated below, users can easily mask voice traffic to be almost indistinguishable from any other data traffic if they have an incentive to do so. Thus, if certain types of traffic bear a heavier burden of regulation or tariff, network entities will have an incentive to use applications that make their "high-burden" traffic appear to be some other traffic.

The three principal ways of distinguishing traffic are:

Protocols: Audio packets are typically carried over UDP to avoid the delays incurred by TCP retransmission and congestion control. However, some VoIP applications, such as Skype, will use TCP if forced to by firewalls or NATs.

Signaling uses UDP, TCP or SCTP, with a likely migration of TLS-over-TCP in the next few years.

Port numbers: SIP signaling usually uses port 5060; there is no assigned port number or port number range for media traffic (RTP), although it is common to set aside such ranges by local convention. However, SIP will work just fine on any other port number.

Traffic pattern: To keep delays to acceptable levels, voice packets will likely be sent at intervals between 20 and 50 ms.

With encryption, it is effectively impossible to tell what kind of content a protocol message is carrying. For example, it would not be hard to develop a VoIP system that uses TLS ("https") to carry voice data, with no reliable means to distinguishing the two.

There is clear precedent for such masking in other areas. As firewalls and ISP policies tried to block file sharing applications, these applications were converted into using port 80, the HTTP port. Thus, an ISP would have to inspect packets and guess whether a protocol exchange was part of a web transaction or file sharing. Such content-based distinctions are likely to be highly error-prone and invasive.

A step beyond masking is to *tunnel* protocols, i.e., carry the desired protocol information as payload of another packet. There are existing applications that tunnel IP packets over HTTP and there are rumors of using DNS to carry IP packets. In the former case, the HTTP connection is simply treated as another network interface.

3 Signaling

VoIP sessions are established using two kinds of signaling protocols, namely intra-domain device control protocols and inter-domain signaling protocols. Device control protocols are based on stimulus-reponse, invoking lower-layer actions without the device having knowledge of user or call state. Examples of such protocols include MGCP, Megaco/H.248 and

proprietary protocols such as Cisco's Skinny protocol. Since these device control protocols assume a single controlling entity, they can only be used within one administrative domain.

Signaling has two core functions, namely to maintain a shared notion of a session in end systems and to perform translations between identifiers (identifier binding). The notion of a session includes the type of media streams and their network and coding properties, as well as, for example, the security services that the session might use. Identifier binding maps a long-lived, user-visible identifier such as a URL or telephone number to a short-lived, network-visible identifier such as an IP address, possibly with additional intermediate identifiers, such as domain names.

Signaling protocols can be used across administrative domains since domains are peers. In the PSTN, Signaling System No. 7 (SS7) is the most common modern signaling protocol. SS7 combines two functions, namely session setup for voice calls and managing resources (circuits) in the switches between callers. VoIP separates the two functions into session signaling, described here, and, rarely, resource reservation along the router path. Since the media transport is connectionless, generally no media path setup is required.

In VoIP, H.323 and increasingly the Session Initiation Protocol (SIP) [1] are the protocols of choice. In addition to call setup, management and tear-down, SIP also offers instant messaging and presence within the same protocol.

The technical community has long distinguished in-band (call-associated) from out-of-band (non-call-associated) signaling protocols, but this nomenclature is somewhat misleading. The first telephony signaling protocols were in-band, in that they used tones or other voice path indications such as hook flash and path disruptions to signal actions and digits. MF signaling is an example of call-associated, in-band signaling.

Since the 1980s, non-call-associated signaling (NCAS) or common-channel signaling (CCS) became possible, where signaling for a number of voice trunks was combined into a physically separate signaling network between switches. However, CCS signaling still needs to visit each switching node along the voice path and is thus geographically relatively closely tied to the voice path.

VoIP signaling combines both aspects, as it can be multiplexed onto the same "channel" as voice data, yet is logically completely separate from the voice data. Since VoIP signaling is end-to-end signaling and does not establish the actual route of voice packets, only the communicating end systems need to have voice and signaling packets coincide. Since signaling information consumes relatively few bytes compared to voice traffic[1] and signaling traffic has far less stringent delay constraints than voice traffic, there is no disincentive to having signaling services be provided by service providers that are far away from either the caller or the callee. For example, a caller in Toronto might use a SIP service provider in Bermuda or New York to route its calls, without significant costs in terms of performance, reliability or services. (The only drawback occurs if number portability remains local, as these providers may not be able to offer numbers in the user's home service area.)

Note that this differs somewhat from the wireless situation. In wireless calls, calls, includ-

[1] As a rough estimate, signaling probably consumes the equivalent of about one second of compressed voice traffic.

ing media streams, need to first be routed to the home network of the roaming user and are then routed within the network of the wireless provider, even if the caller and callee are within the same location. Due to the separation of signaling and media in VoIP, this is not needed in VoIP. The media stream would take the most direct route between two Internet-connected VoIP customers, regardless of where their home SIP proxies are located.

Classical user services, such as three-way calling, caller ID, conferences, or call waiting, are typically provided by a combination of user agent end systems and network servers [2]. The primary motivation for servers is that these are assumed to be available continuously, with constant and publically routable IP addresses. These servers are typically operated by entities other than the ISP, namely the domain owner for the subscriber to the voice service.

It should be noted that non-VoIP multimedia services, namely streaming multimedia services, use very similar signaling protocols, such as RTSP. They also use the same media transport protocol, so that the difference between streaming media and VoIP, from a protocol perspective, is marginal.

4 Application and User View

4.1 Transitioning from Voice-Dominated to IP Communications

It appears likely that the instead of thinking of email and telephone as the most basic electronic means of communications that we will make the transition to dividing communications into synchronous and asynchronous means, or message-based and interactive ones. Message-based communications, such as email, delivers messages reliably even if the recipient is not currently connected, with generally short, but unbounded delivery time and no human response guarantee. Interactive communications, comprising VoIP, instant messaging, text chat, and other collaboration modes, encourages instantaneous back-and-forth interaction and generally has the notion of a session (or call) between two or more participants. There seems to be a natural dividing line between those two modes, although sometimes rapid-fire email exchanges approximate IM interactions.

This blending of real-time interactions is already visible, despite somewhat limited tools and manual configuration. For example, it has become quite common to exchange instant messages as part of a kind of sidebar conversation in audio conferences.

4.2 Making Real-Time Communication Context Aware

The basic four-step user interaction to set up a telephone call has not changed since the 1880's: obtain dial tone, ring distant user, callee picks up and conversation ensues, with a well-defined termination point when one party hangs up. (Naturally, dialing is a more recent invention, but one can consider the switchboard as an early example of voice recognition dialing found on some mobile services today, except that the early 20th century version had better recognition accuracy.)

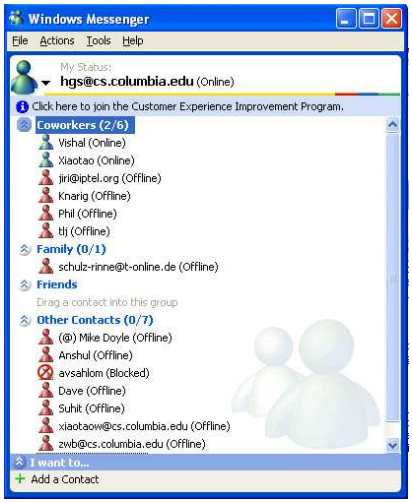

Figure 1: Windows messenger

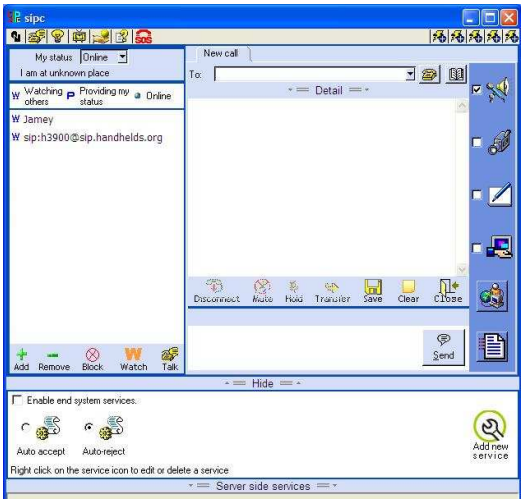

Figure 2: sipc SIP client

Figure 3: Nortel SIP client

This model of user interaction is simple enough to understand that four-year olds can place phone calls. Thus, while this model of user interaction for real-time communications is likely to continue to exist, its shortcomings have led to a number of alternate arrangements. The major shortcoming is that this process has become ever less likely to succeed, leading to repeated voice mail or no-answer. Thus, in many cases, the real-time service of voice communications has become an involuntary asynchronous messaging service. This is probably the major reason that email has become the most common means of corporate communications, as most people find it far more efficient to write a message than compose a speech in real-time. (Extracting information such as phone numbers, appointments and addresses from email is also far less error-prone. Email also makes it significantly less time-consuming to reach multiple individuals with the same message.)

Mobile phones have theoretically increased reachability, but many subscribers hesitate to publish or widely disseminate their mobile phone numbers due to the receiver-also-pays model and the common case that a mobile phone call is inappropriate at the time placed.

The underlying problem in both cases is similar: traditional telephony, both mobile and landline, is *context-unaware*. The caller has no efficient way of knowing whether, where and how he is going to reach the called destination. The normal "ring-no answer-forward to voicemail" cycle takes upwards of 30 seconds, making probing inefficient. The caller has no way of knowing where the mobile callee is located and whether a call might be appropriate and desired at the time or whether the call interrupts a theater performance or church service.

These deficiencies have led to a number of ad-hoc mechanisms that avoid the current technology limitations. For example, it has become quite common to arrange phone calls through an email or instant message (IM) exchange.

Better tools are starting to emerge. In particular, the combination of presence, instant messaging and real-time voice communication is now commonly available. Presence and IM

have always been linked, although there is no technical connection. In order to avoid typing a real-time message to somebody who is not at their PC, presence indication offers at least a first-order approximation to estimate whether the recipient is likely to see the message soon.

Most commercial IM systems rely on human users to update their status information, which is likely to be unreliable, or simply indicate "idle" status if a person has not touched the keyboard for a while. To improve the usefulness of context information, the Internet Engineering Task Force (IETF) is currently standardizing *rich presence* [3], which includes information about the activity, type of location (office, home, theater, etc.), communication privacy and mood in the presence updates. Many of these items can be derived from existing computerized data, such as appointment calendars. In addition, location information is likely to be an important indication of context, but needs strong privacy protections to make users feel comfortable integrating this into their presence information. Additional examples of such services are described in [4, 5, 6].

Examples of commercial and research applications are shown in Fig. 1 (Microsoft Windows Messenger), Fig. 2 (Columbia University sipc), Fig. 3 (Nortel Multimedia PC client). Almost all modern VoIP implementations offer variations on the same theme. These tools combine instant messaging, presence and voice (and, in some cases, video) communications into one software tool. Rather than ring-and-wait, the user interaction would likely be different:

- Subscribe to presence of colleagues and friends, possibly available only during selected time periods;
- When wanting to talk to somebody, check availability and estimate whether a call is appropriate, e.g., whether the person is likely to be in a meeting or about to leave the office;
- If the other party is available, send an IM and indicate desire to talk. The conversation may remain as text chat, or, at any point, may be converted into a voice or multimedia call, transparent to the user.

Thus, Internet real-time communication is increasingly unlikely to exclusively follow the ring-and-answer model of classical telephony. In addition, the very same tool can be used for pure "data" communication (text chat) or to initiate voice communications.

4.3 Breaking the Call Model

Much of the measurement and billing of telephony is based on a notion of call duration. However, this model is likely to be supplemented by models where the notion of a call or session is far more distended and the active voice periods are only a small fraction of the call duration, if there is a perceived notion of a call at all. In other words, calls or, better, application-layer sessions, may last for days or weeks, while only short bursts of audio are exchanged.

In addition, such sessions may well be largely simplex or half-duplex, rather than conversational, i.e., only one side may send audio or video, with turn taking spread over long intervals.

There are two existing service examples illustrating this, push-to-talk and hoot-and-holler. The former models trunked radio systems for dispatch on cell phones, while the latter provides multicast communications, primarily in banking environments. Both allow instant voice communications, without setting up a call as such.

Once the notion of per-minute charging becomes less common and the classical limitation of one concurrent voice call per "line" disappears, one may expect that many people will keep certain sessions open all the time, effectively like an intercom or walkie-talkie.

In summary, VoIP and closely related technologies are likely to add new call models to our communications repertoire, avoiding the overhead and annoyance of the classical call-ring user interaction. Many such "calls" will not even be perceived as phone calls, as they are simply continuations of IM chats, within the same tool set and without ever dialing a phone number.

4.4 Phone Numbers and Addressing

For decades, telephone numbers have been the single most visible communications identifier. Individuals and businesses could publish their phone number and be assured that just about anybody would know how to use this number to get in touch with them. Numbers were easy to convey orally, are reliably transmitted even in handwriting and work reasonably well across language boundaries. Phone numbers also provided an approximation of the location of the callee, at least to the country and often to the city level.

However, phone numbers appear unlikely to maintain this central role, as they are increasingly running into their inherent limitations:

- Phone numbers are relatively scarce, so that many individuals need to share a single number, e.g., within a family or small business.
- Numbers are tied to devices or phone jacks, which is rather unnatural. (One can obviously forward a number, but this incurs relatively high costs and does not work well when a family has one landline number, but several mobile phones, one for each member of the family.)
- Many metropolitan areas now have multiple area codes, thus making it more difficult to remember numbers, as one has to remember 10 instead of seven digits. As number portability becomes international particularly for mobile devices, phone numbers effectively become random digit strings with fifteen or more digits, which are difficult to remember and prone to mistyping.
- Since phone numbers are assigned densely, the likelihood of misdialing is relatively high, as any random digit combination is likely to be a valid, if unintended, phone number. Recently, phone number assignment has gotten more efficient, further increasing the probability of misdialed numbers disturbing somebody.

- Call routing for mobile numbers is very inefficient, as the call has to be routed to the area code for the device first.
- Subscribers either have to change their phone numbers when moving more than very short distances or phone numbers lose their geographic significance as global number portability becomes more common.
- Phone numbers provide no indication whether a number reaches a residence or a business, a landline phone, a mobile phone or a fax machine.

Given these limitations of phone numbers, it is increasingly likely that they will be supplemented by email-style addresses (or URLs), where a user can be reached under the same address for email, instant messaging and voice calls. About 90,000 staff and students at Columbia University, MIT, the University of Pennsylvania and Yale University are already reachable by phone via their email address, even though these campuses largely still use traditional telephony. (This effort is part of the sip.edu project in the Internet2 consortium.)

Email addresses have three principal advantages: their supply is effectively unlimited, their administration is decentralized and they can be made easier to memorize or recognize. The lack of scarcity makes it easy to assign multiple such identifiers to a person, to assign identifiers based on roles rather than devices or individuals and to create ephemeral identifiers for resources with a limited life time, such as audio and video conferences. Decentralized administration allows anybody to register a domain name for a few dollars each year and then delegate names from that namespace, without coordination with other entities. In many cases, personal names ("John.Doe") or functional designations ("info@example.com") can be embedded into email addresses, making them easier to remember and harder to misdirect communications since typos are most likely to generate an invalid address.

Clearly, due to their widespread use and their ability to be entered into a very limited user interface, phone numbers will continue to be important for years to come, but as end-to-end IP communications becomes more common, many of these entities will only be reachable by an email-like identifier. The ENUM directory service [7] will provide the bridge between E.164 phone numbers and these more expressive identifiers. With ENUM in particular, E.164 numbers as a means of distinguishing end-to-end Internet vs. Internet-to-PSTN calls will cease to be possible. Indeed, the same number may refer to an Internet-connected device, reached directly without ever touching the PSTN, on one day, and refer to a mobile phone the next day.

Even if we maintain numeric identifiers, increasing number portability makes it attractive to use non-NANPA numbers, either from another country or a special VoIP country code. Such usage could obviate the scarcity concerns.

If we were to assign numbers more like identifiers (license plate numbers or social security numbers), the existing 10-digit numbers could support a personal number for every person within the geographic reach of the North American Numbering Plan (NANP), with room to spare.

4.5 Breaking the Line Model

Many parts of the existing phone system depend on the notion of a "phone line". For example, numbers are assigned to lines and different regulatory charges are levied on the first line and subsequent lines. However, the notion of lines will cease to make sense in a VoIP environment for two reasons. First, as discussed in Section 4.3, calls may now last for extended periods of time, with only periodic voice or data activity. Thus, it is quite possible to maintain a large number of sessions even on limited bandwidths. Secondly, the peak bandwidth available to residential end users is already equivalent to 20 or 30 voice lines and could increase to hundreds of line-equivalents in the near future. As an aside, the disappearance of line scarcity also means that services such as call waiting become less relevant.

Thus, even if not all family members have their own number, they could easily all converse at the same time, in a manner similar to hunt groups in telephony today.

As noted in Section 4.4, identifiers are likely to be plentiful, either as non-NANP numbers or as email-style URLs. This further reduces the need for the notion of lines.

Existing hardware and software IP phones commonly support a large number of virtual lines, typically four to six. This number is primarily limited by the ability to handle concurrent audio streams and user interface issues, but if long-lived, intermittent sessions become popular, there would be little problem with extending the line count to just about any number. (Since user attention is limited, some virtual lines would presumably be put on hold, which causes the inbound and outbound voice data streams to cease.)

4.6 Multimedia

Current IP communications tools seamlessly integrate both classical audio and video communications as well as means of communications that are likely to be viewed as data. For example, most IM systems that support voice calls (e.g., Yahoo, Windows Messenger and AOL) also support file transfer, which is not substantially different from HTTP or ftp. They often also support shared web browsing or screen sharing. These services can be invoked seamlessly both as part of session setup or negotiated later. For example, the Session Initiation Protocol (SIP) in conjunction with SDP [8] can add and remove media, including "media" such as file transfer, screen sharing and shared web browsing, within the confines of a single session, without the user having to do more than check off another item on the media list.

4.7 Embedded Voice

Traditional voice and Internet communications were directly visible to users, i.e., a user would need to use an easily recognizable tool, such as a telephone, and perform a sequence of manual actions to place a call. Similarly, designated software tools, say, Eudora or Outlook for email, made it plain when an email was being sent.

More recently, all forms of IP communications are being *embedded* into other devices and software systems, ceasing to be visible as such.

Here are some examples of embedded IP communications: Burglar and fire alarm systems may start to send instant messages instead of placing voice calls. Audio and video monitoring systems, such as security cameras, may remain in a session ("call") for months, but only transmit audio and video when something of interest happens, e.g., when a motion detector is triggered.

Interactive, multi-player computer games may set up voice communication channels between participants, either via end-to-end IP communications or even bridging into the PSTN. Such sessions may be triggered if two avatars approach each other (Fig. 4), without any explicit user interaction. Before users are in proximity to each other, they may simply exchange coordinates and other activity coordination. As illustrated before, there is thus a seamless and user-invisible transition and integration between real-time voice and data exchange.

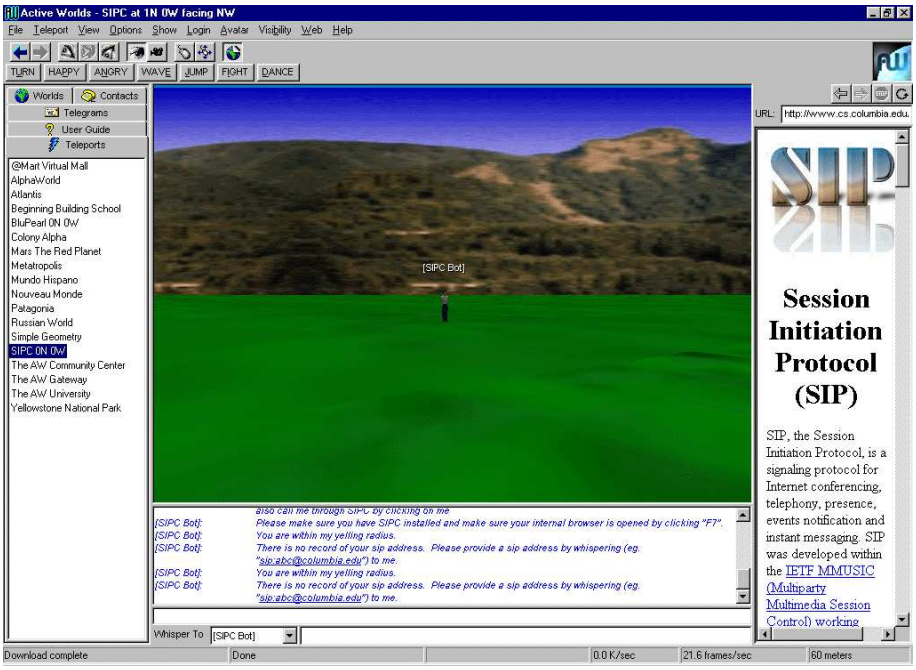

Figure 4: Voice communication embedded into a virtual reality game

5 VoIP as a Driver for Broadband Deployment

According to the Federal Communication Commission (FCC)[2] report, average household expenditures in 2002 for local and long-distance services are $36 and $12/month, respectively.

There is a clear motivation for consumers who already have purchased broadband Internet access to use VoIP. Depending on the assumptions on user behavior, one can distinguish three

[2]Trends in Telephone Service

categories of users, depending on whether they are willing to drop their existing telephone service:

Long distance only ("second line"): These users, the most conservative group among the three, want to maintain their regular phone line, e.g., for local and incoming calls, and are only interested in using VoIP if it reduces their expenditures for long distance services. Since typical unlimited VoIP plans cost \$20-\$30/month and average household expenditure for long distance has dropped to \$12/month, only the heaviest users would find this attractive.

Local and long distance: Some users have installed broadband Internet access, without considering VoIP and would continue to pay for it even without VoIP. These users should be willing to replace local and long-distance service with VoIP, since the total cost of VoIP service, \$30, is significantly below the \$48 spent on local and long-distance service. However, the \$12 for long-distance service includes international calls, which are generally not included in VoIP packages.

Broadband because of VoIP: A final class of users wants to minimize overall expenditures, but is not particularly interested in broadband for Internet access only. For this class of users, VoIP is attractive if the following cost relationship holds:

dial-up Internet access + telephone service $<$ broadband + VoIP

If we assume a cost for dial-up Internet access of around \$20 and telephony cost of \$48, and broadband costs of \$40 and VoIP of \$30, the total expenditures are about the same for both the narrowband and the broadband option (\$68 vs. \$70). Thus, even a minimal perceived advantage offered by either VoIP or broadband would make it advantageous to switch from traditional to broadband services. If broadband drops even marginally in cost, this arrangement is likely to motivate a transition. Again, the caveat about international calls should be noted.

Since the expenditures for local and long distance service differ significantly by location, by the number of enhanced services (caller ID, voice mail, ...) selected and the volume of long-distance calls placed, the numbers above are only estimates. Naturally, heavy users of long-distance services and those that already have broadband are the most likely users of VoIP.

Costs for local services have not decreased significantly in the last few years, so that there is an increasing proportion of users that will find VoIP as a strong additional motivation to obtain broadband service, even from a purely financial perspective. This will likely require, however, the ability to transfer numbers to a VoIP provider, i.e., local number portability (LNP), which does not appear to apply to many DSL customers, and the perception that VoIP is as reliable as regular phone service.

6 Regulation

Telecommunication regulation addresses multiple objectives, including ensuring competition in the face of natural or legacy monopolies, ensuring quality of service and safe-guarding social goals such as universal service and provision of emergency services. Below, we briefly investigate some of these issues.

6.1 Competition

From the very beginnings of data networking in the 1960s, the notion of protocol layering and well-defined interfaces between layers has been a cornerstone of network engineering. While the seven-layer OSI model was less than successful at the upper layers, the Internet has operated on a five-layer model for more than twenty years, namely physical, link, network, transport and application. These layers offer increasing specialization as one moves up the layers and natural end-to-end vs. hop-by-hop semantics. While "shim" layers have been introduced, e.g., between layer 2 and 3, to create virtual private networks or to perform traffic engineering, the layering structure seems to have stood the test of time, with very little motivation to change it. Very similar models have emerged in different technical communities, so that it is likely that it is more than an accident. These interfaces have also become commercial interfaces, i.e., service specifications where there is a large body of suppliers for several core interfaces, both of hardware and software as well as services. For example, there are service providers at the physical (dark fiber, DSL, SONET), network layer (ISPs) and application layer (web hosting providers, ASPs, SIP service providers).

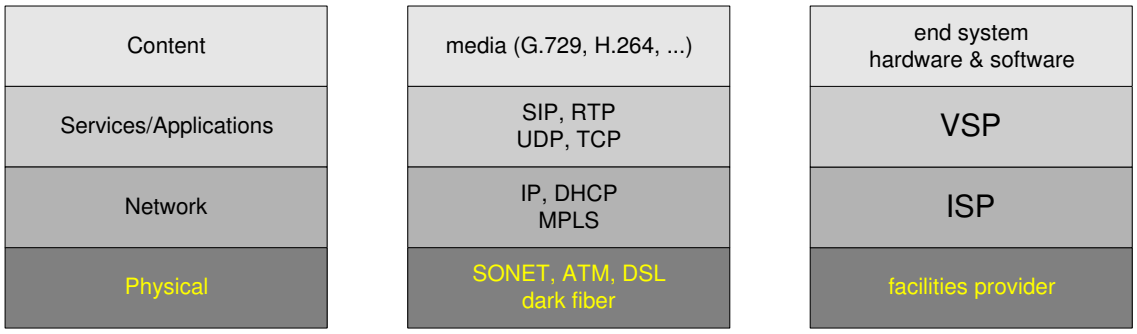

Figure 5: Logical network layers

The basic premise of the layer model is information hiding, i.e., upper layers only have a very narrow interface to express their requirements for service to the lower layer and lower layers transport upper layers as opaque data. (Layer violations occur, but experience with them has been uniformly bad, even if they are sometimes necessary due to non-technical constraints.)

Given the longevity of the model and its clear definition, it makes sense to extend a layering model to regulatory issues as well. Compared to other division, such as basic and

enhanced services, it maps well to well-defined commercial interfaces.

The precise number of layers is less important than the notion of layering. For example, some have proposed a four-layer model that separates physical, network or logical, services or applications and content layers[3] Figure 5 illustrates the model as applied to VoIP. Note that services shown within one rectangle do not necessarily have to be provided by one service provider. For example, one can split up the physical layer into two, e.g., the "naked" (dry) copper to the premises and the DSL modulation and coding. Similarly, services such as DNS can be offered by third parties unaffiliated with one's primary ISP.

Given the mapping to well-defined interface specifications, the layering model appears to be significantly easier to apply than the traditional telecommunication vs. information service distinction, although one can argue that the former meant the physical layer and information service is anything above that.

While physical access networks are often natural monopolies, at least for the same communication technology and particularly for residential users, facilities can readily be shared by multiple network-layer providers (ISPs) and any number of service and content providers. Such sharing is increasingly easy the higher one moves into the protocol stack. For example, it is technically challenging to have two ISPs share the same DSL or cable modem, while there is obviously no reason that a web sites cannot be visited by subscribers to two different access providers.

Since VoIP is just another network service, there is no inherent natural monopoly and the barrier to entry is very low. (Indeed, providing IP-to-IP service requires only some server capacity, or in the case of peer-to-peer services like Skype, a web server.) However, this assumes that the natural delay, delay jitter and packet loss offered to subscribers is sufficient along the path from the caller to the destination or the caller to the PSTN gateway. In particular, the VoIP provider has no control over any quality-of-service impairments incurring between the last-mile provider and the peering point, even if the provider chooses an appropriately engineered tier-one Internet carrier rather than just connecting as an end customer to various points of presence. Broadband providers could, if so inclined and not legally prevented, add impairments to the VoIP services offered by third parties, assuming they can identify VoIP packets. (Naturally, they could also attempt to block all SIP signaling. Even the usage of network address translators (NATs) already makes it more difficult to provide high-quality, reliable VoIP service as it may require that the customer maintain a TCP connection to the VoIP service provider.)

If differentiated services are required, where some packets receive better treatment than others, the more flexible solution is to have the end application invoke a generic signaling service that sets aside resources or assigns higher delivery priorities to certain packets, presumably for extra financial consideration or as part of a volume-limited part of the network service. Such differentiated treatment is also helpful for non-VoIP services, including distributed computing and multiplayer games.

Thus, to ensure competition, it appears appropriate to treat the physical and possibly

[3]See also *A Horizontal Leap Forwards*, MCI public policy white paper, March 2004, for a much more detailed treatment of this issue.

network layer as deserving special attention, while only ensuring that lower-layer providers do not discriminate in their transport amongst service and content providers.

Note also that particularly at the upper layers the notion of a provider may become fuzzy. For example, while we currently have a number of voice service providers (VSPs) like Vonage and Primus that provide call routing, number translation and gateway services, such services can also be provided by individuals and enterprises for their own need. For example, a university could offer VoIP services to its students and staff, without the assistance of a service provider, just as it offers email and web services without relying on a separate service provider. (The institution might rely on PSTN gateway services provided by third parties as this is likely to be more efficient than running its own set of gateways, but this is an economic optimization, not a technical requirement.) Thus, any consumer or enterprise, large or small, that can purchase a domain name and contract for hosting services can effectively be its own micro-VSP, indistinguishable technically from a VSP offering its services to the public. Naturally, just as with email services, a single broadband customer can also easily use multiple different VSPs. In a residential environment, each family member may use a different VSP, for example.

6.2 Quality of Service

Currently, PSTN carriers in the United States need to report major outages to the FCC, allowing some indication as to whether a carrier is operating a reliable network. There are no such metrics for Internet Service Providers, either residential or tier-one, making an informed choice difficult, even assuming that a customer could pick the equivalent of a "long distance IP provider". Unfortunately, the choice of long-distance ISP is generally given by the choice of the destination and, to a lesser extent, the source of the data.

6.3 Emergency Services ("911")

Even landline VoIP services are effectively nomadic, since users can pick up their IP phone and install it anywhere in the world, while still being reachable and appearing to call from the same E.164 number.

Emergency services for VoIP require three components: a common emergency number, a way to route calls to the location-appropriate emergency call center (PSAP) and a mechanism to deliver caller location to that PSAP so that emergency services can be dispatched efficiently. Both the second and third requirement involve the provision of caller location.

Due to layering, the responsibility of providing emergency calling (911) cannot rest only on the provider of call routing services, i.e., the VSP. Only the provider of physical access, such as the DSL or cable modem provider for residential users (residential ISP), is likely to have reliable information about the current location of the caller and thus needs to provide this information to the caller. The residential ISP does not know which VSP, if any, the customer is using, so it needs to convey the location information to the customer, as only it will know the current VSP used to place the emergency call.

6.4 Regulatory Evasion

Regulators need to be aware that national regulation invites evasive action if too burdensome. Such evasion has become increasingly easy.

Since costs for transnational calls have plummeted[4], it is quite feasible to provide service across national borders. For example, a Canadian service provider could offer service to United States customers, using signaling services and gateways located only in Canada. The US customers would reach the PSTN gateways via IP.

7 Conclusion

This memo has attempted to put VoIP services into a broader architectural perspective, illustrating that what appears to be a single monolithic service offering is actually a set of protocols, each of which is also used for services other than VoIP. For example, signaling protocols are used for instant messaging, event notification and presence, media transport protocols are used for streaming media and directory services are used for email and other contact information. This protocol reuse is reflected also in end systems and applications that increasingly go beyond simple voice call to multimedia collaboration, involving both streaming media such as audio, video and real-time text as well as shared data and applications.

VoIP offers the opportunity to break out of a number of artificial scarcities: the scarcity of identifiers (telephone numbers), lines, services and media that have limited the advances in legacy PSTN systems.

This blending of services is one of the great strengths and promises of Internet telephony, more so than temporary financial advantages. Trying to break apart this symbiotic and natural relationship by regulatory intervention is likely to impede this progress, going back to the "silo" of standalone applications with far smaller utility.

References

[1] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. R. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: session initiation protocol," RFC 3261, Internet Engineering Task Force, June 2002.

[2] X. Wu and H. Schulzrinne, "Where should services reside in Internet telephony systems?," in *IP Telecom Services Workshop*, (Atlanta, Georgia), Sept. 2000.

[3] H. Schulzrinne, "RPID – rich presence information data format," Internet Draft draft-ietf-simple-rpid-01, Internet Engineering Task Force, Feb. 2004. Work in progress.

[4] X. Wu and H. Schulzrinne, "Programmable end system services using SIP," in *Conference Record of the International Conference on Communications (ICC)*, May 2003.

[4] For the author, it is now significantly cheaper to call London than to call another city within 20 miles of his home town. Fairly soon, it will be cheaper to route such intrastate calls first to London and then back to the United States, as the price difference is approaching a factor of two.

[5] S. Berger, H. Schulzrinne, S. Sidiroglou, and X. Wu, "Ubiquitous computing using SIP," in *ACM NOSSDAV 2003*, June 2003.

[6] H. Schulzrinne, X. Wu, S. Sidiroglou, and S. Berger, "Ubiquitous computing in home networks," *IEEE Communications Magazine*, pp. 128–135, Nov. 2003.

[7] P. Faltstrom, "E.164 number and DNS," RFC 2916, Internet Engineering Task Force, Sept. 2000.

[8] M. Handley and V. Jacobson, "SDP: session description protocol," RFC 2327, Internet Engineering Task Force, Apr. 1998.