

8. SYSTEM ARCHITECTURE CONSIDERATIONS

8.1 Classification of Applications

The classification of the applications is important to understand the requirements that different types of applications impose on the overall system architecture. The following approach is valid for both wireline and wireless networks.

The taxonomy of applications appearing in Fig. 24 is given by Larry Peterson and Bruce Davie [Pet 96]. It is based on the concept of integrated services in the Internet issued by Bob Braden, Scott Schenker and David Clark [Bra 94]. In this classification, Peterson and Davie divide applications into two main groups depending on their time criticality. The elastic applications include such as FTP, Telnet, e-mail and Web browsing. They have the property to accommodate gracefully delay variations, but on the other hand, they tend to become unusable if a single bit of the received data block is corrupted.

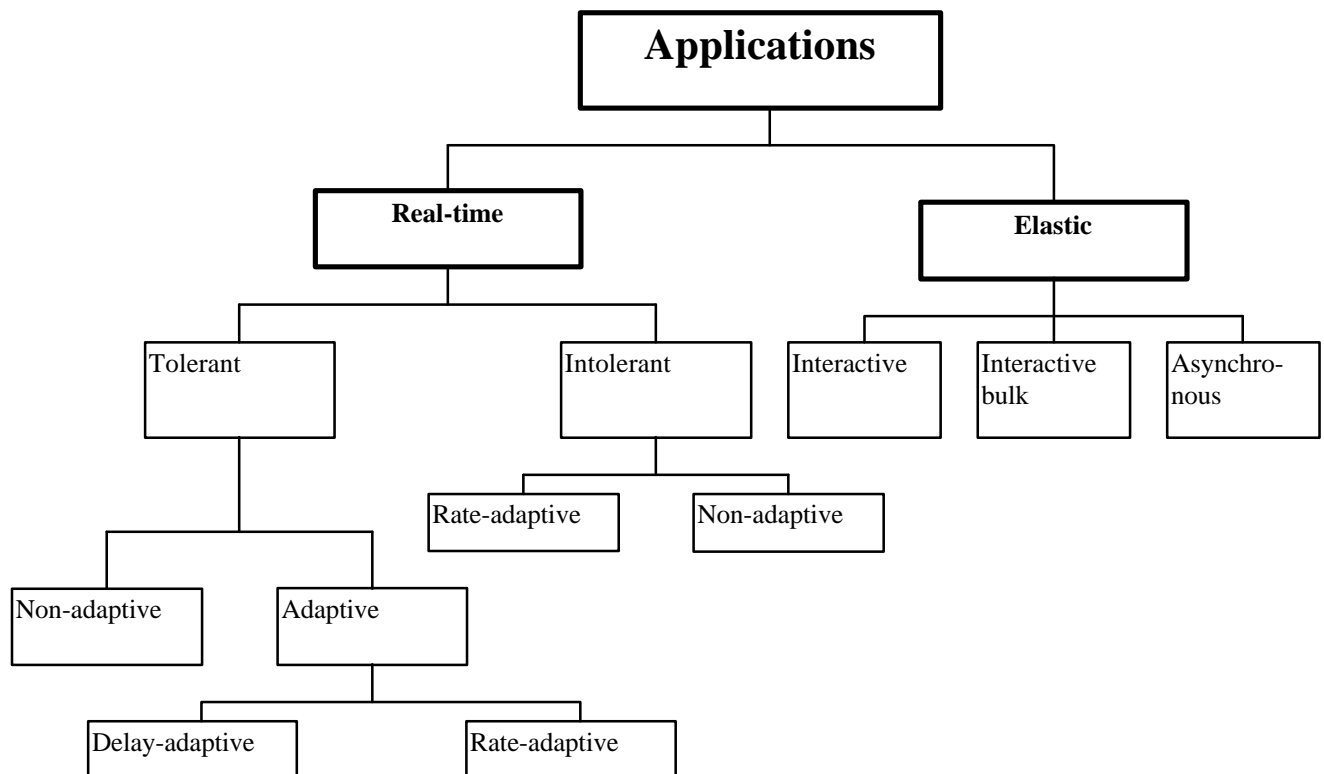


Figure 1 Taxonomy of Applications [Pet 96]

Real-time applications usually tolerate some loss of data, because of the redundancy included in speech and video signals. Loss of data degrades the QoS, but does not render the data useless. These real-time applications are classified as *tolerant* ones [Bra 94]. Some real-time applications do not tolerate the loss of data. Duly they are classified as *intolerant* real-time applications [Bra 94]. An example of such an application is remote

operation of an excavator. Short delays of the control data are of course necessary for such tele-operation.

All real-time applications share the inherent common feature that the data must be available within predetermined limits. Any data falling beyond those bounds is useless. Therefore, the data must be delivered in a continuous stream. In an IP network context, the term real-time streaming is used to characterize the continuous flow of data. To cope with the inevitable variations of delay encountered by the packets in the network, the incoming packets need to be buffered and played back after a constant or variable offset time. The price we have to pay for handling the delay variations is an increased overall delay i.e. higher latency. The shorter the buffering delay is, the higher is the probability of a packet loss. Apparently, we face a trade-off situation between impaired audiovisual quality and latency.

The taxonomic model of Fig. 24 characterizes real-time applications further on based on their *adaptability to variations in delay and rate*. Applications, which are capable of adjusting their playback point are categorized as *delay-adaptive* applications. On the other hand, many multimedia applications can tolerate lower transfer rates, but only at the expense of impaired QoS. Typical example is videotelephony. The reduction of the transfer rate results in impaired temporal performance and the received video becomes jerky. This kind of real-time application is characterized as a *rate adaptive* one.

Another way of categorizing applications is presented by [Kwo 97]. The author divides them in three main categories: real-time streaming applications, real-time block transfer and non-real time applications. The real-time streaming application delivers time-critical information in real-time. This means that the average transfer rate of the delivered data is similar to the rate at which the information is displayed at the receiving end. The receiver has to be able to cope with the delay jitter the network incurs in the packet stream. Discontinuities in the stream are handled in the receiver by buffering the incoming data as was done in the previous model. The buffer size i.e. its playback time determines the tolerated maximum delay variation.

The real-time block transfer is distinguished from the real-time streaming in the sense that there is no strict timing constraint between consecutive blocks of data. The concept of real-time transfer of individual blocks stems from the fact that individual blocks need to be delivered within a specified deadline. The application layer has the knowledge of handling the timing relationship of the blocks and the network treats them equally with any data packet. The author mentions Web browsing as an example of a real-time block transfer application. A data block handled by the application layer is split into packets of variable length at the network layer.

The fundamental difference between these two approaches is that the concept of a real-time system has a different meaning. This result in a different interpretation of Web browsing. The first approach treat it as an elastic application i.e. non-real-time one, whereas the latter classification treat Web browsing as a real-time application. However,

in this case, the inherent meaning of real-time is changed. Regardless of the timing relationship, we feel that the term real-time is inappropriate in this context. There are other timing constraints, stemming from system itself or the required user response times, but they differ significantly from the definition of real-time applications presented by [Bra 94] and [Pet 96].

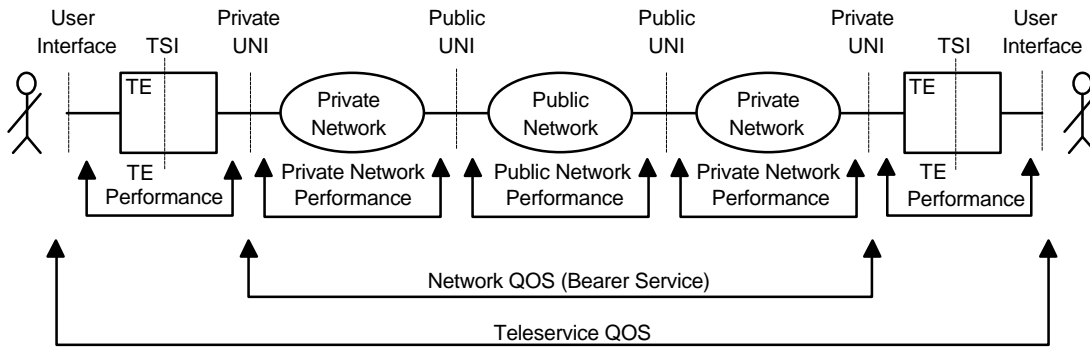
In wireless environments, we confront another problem, packetization of real-time data. It is faced also in the wireline IP networks. The packet length should not be too long to avoid drastic degradation of the quality in case a packet is lost. In particular speech packets are critical, because such interruptions, which impair significantly intelligibility, i.e. remove syllables or entire words, are very annoying. Albeit the data is permuted to spread error bursts in time, the probability of unrecoverable error patterns increases as well as the overall delay.

The flip-side of a short packet length is the increase of relative portion of overhead in the form of header and error control data. In error-prone environments we have to sacrifice more of the available bandwidth for error protection, which makes this problem worse than in wireline IP networks. Since the overall bandwidth of a wireless system tends to be already narrower than that of the fixed network, the portion available for payload will be significantly less. Therefore, selection of the packet length is a critical issue as well as the packet protocol design. In this context, it is unclear, whether packet radio in general is feasible, both technically and economically for transmitting data of real-time multimedia applications. Moreover, the latency problem need to be addressed in packet radio environments. As we have shown in Section 8.3.4, the setup is even in this respect less favorable than in the wireline IP networks. Further research is needed to clarify this matter.

8.2 Quality of Service [Framework for Wireless Multimedia]

8.2.1 QoS Reference Model and Design Principles

Quality of service is also a concept with a lot of ambiguity. As a matter of fact, it has different meanings depending on from what and whose perspective we are looking to it. The foundation for all QoS architectures should be the subjective quality experienced by an end-user. This initial QoS, usually defined in subjective terms, translates into objective quality measures in different points of the communication infrastructure. The following quality of service (QoS) Reference Model issued by MMFC [Mmc 95], which is illustrated by Fig.24, is based on this approach.



- Notes
1. For some services only one user interface exists, in which case one of the TEs performs the function of the client and the other the function of a server. The server may be part of the public network.
 2. For some users or services, the private network will not be present. However, the possibility that it may exist should always be considered from a performance point of view.

Figure 2 Quality of Service and network performance reference model

The idea of the model is to define the teleservice QoS i.e. the overall QoS perceptible to the enduser. The multimedia system is decomposed into three classes of components, each of which having specific QoS requirements derived from the user perceived QoS. From a user's perspective, quality of service should be expressed by parameters which

- focus on user-perceivable effects, rather than their causes within the network
- are independent of the networks internal design
- take into account all aspects of the service from the user's point of view which can be objectively measured at the service access point
- can be assured to a user by the service.

The Internet has not been taken into account in the model. At present, the second and the last bullet point are not fulfilled by the best effort type delivery of information in the Internet. The guaranteed service class [Bra 94], which is being widely investigated, would fulfill the above parameter constraints. The QoS of predicted service class [Bra 94] cannot be *assured*, but service guarantees can be given with a certain level of probability. Therefore, the model is not fully applicable to Internet based services.

Bearer services or transport services, acting as transport vehicles of bit-streams, are constrained by given technical parameters such as BER, error free seconds, propagation delay, packet delay, packet drop-out rate, and they duly have a more technical content.

The question reads: on what grounds should we determine QoS parameters? Subjective user testing carried out on a testbed is a logical way of investigating user perceptible

quality parameters. The QoS is actually a subset of usability tests. Therefore, a usability testbed should support also QoS testing. The mapping of derived subjective QoS values into quantifiable, objective (technical) ones is a tricky task, because we need to know the influence of different types of impairments on subjective quality. The NCS Boulder labs has done extensive work in this area by developing an expert system, which uses a large base of subjective testing results as a reference. The system allows the assessment of heavily compressed audiovisual frames without any subjective testing.

On setting allowable *design* values for QoS parameters, the main principle is that the QoS should not impair significantly from that of respective wireline multimedia services and applications. The attribute “respective” means that essential transmission parameters, such as available bandwidth for payload, are similar in the two cases. For wireless Internet telephony this is a stringent requirement, because, by its nature, a wireless network tends to incur artifacts, errors, and additional delays in regard to its landline counterparts to an extent discernible to the user. As we have noted, the propagation delay is constrained by the system itself, which means that there is not much we can do about it. Reduction of other QoS impairments can be done only at the expense of the additional overhead i.e. reduced bandwidth for payload. If the delays are too long from the user’s point of view then the risk is high that the service will be rejected.

Subsequently our aim is to compare the subsequent MMCF parameter values with respective figures derived from assessments of different mobile systems to evaluate their feasibility for real-time multimedia communications over the Internet.

8.2.2 QoS Parameter Values

ITU-T has outlined a QoS framework in Recommendation E.800 [ITU 93a]. ITU-T defines QoS as follows: The collective effect of service performances which determine the degree of satisfaction of a user of a service. The E.800 QoS performance concept is depicted in Fig. 25. The dashed line is added by the author.

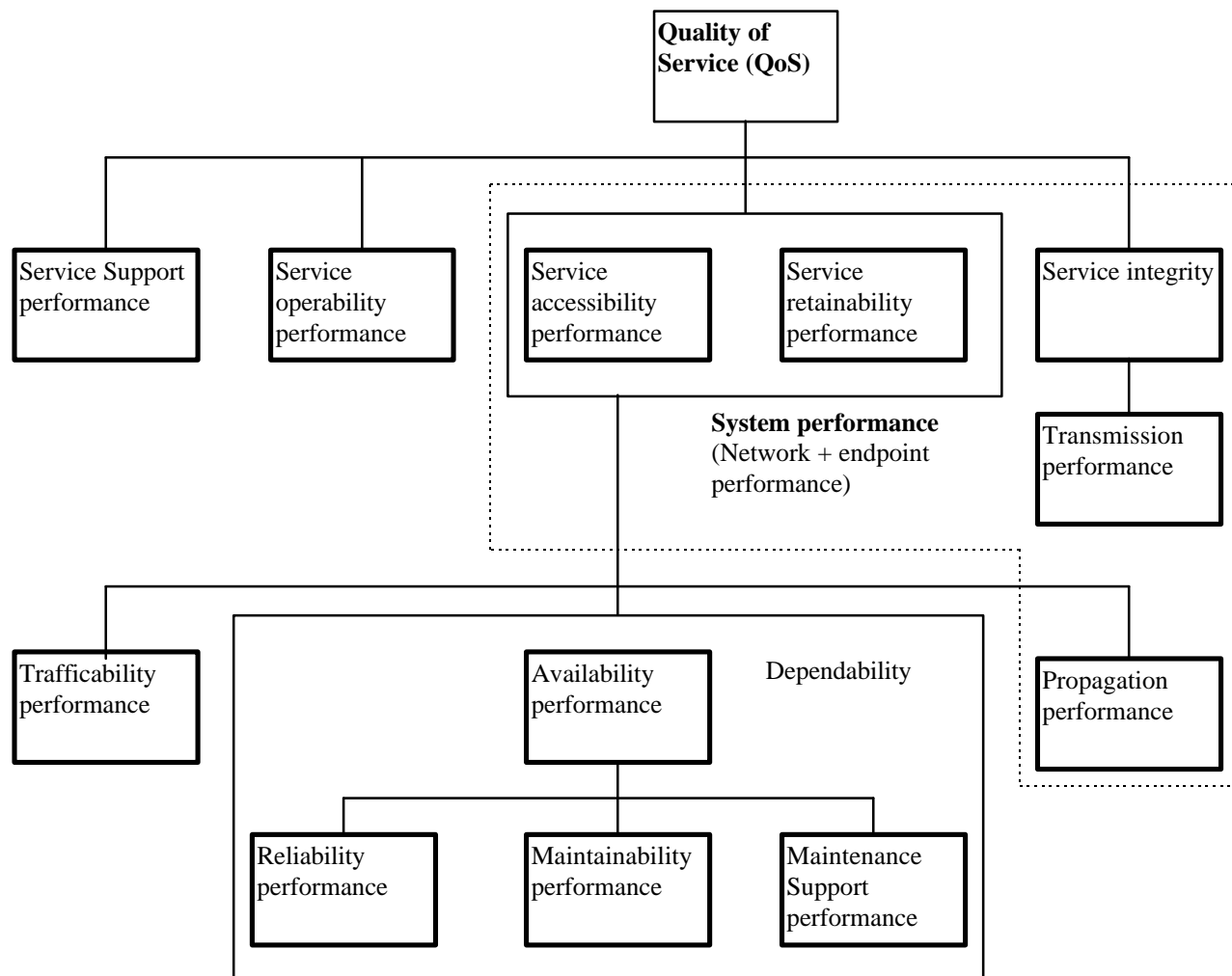


Figure 3 QoS performance concept of ITU-T Rec. E.800

The QoS parameters values appearing in the following Table 10 are issued by MMCF. They are given for a specific application category, namely multimedia desk-top collaboration (MDC). This is a typical real-time multimedia application, which bears some resemblance with Internet telephony, because MDC is exactly what H.323 systems are about. A wireless version of MDC is, of course, something else than the desk-top, but it could be a potential enhancement of multimedia Walkman. Rather than being an outcome of scientific research and testing, the MMCF parameter set reflects experiences and view points of MMCF members. The parameter values in the MMCF table are debatable as well as the MMCF approach itself, because rather than reflecting user oriented QoS parameter values, it presents measurable engineering values. Subjective quality has been translated into technical parameters such as bit rates and bit error rates etc. However, the bit-rates can only be lead-directing. Compression technologies for both speech and image are making constant progress albeit not so rapid. Before wideband wireless services are available, the service classes 1 and 2 are valid only. Therefore, in short to medium term, only those service classes are applicable for LBR wireless multimedia.

Many of the values in Table 10 are derived from ITU-T Recommendations. For instance, we can calculate what bit error rates the requirements of error free intervals translates into. The results appear in subsequent Table 12.

Table 1 Quality of Service Classes for MDC Applications

QOS Parameter	QOS Class 1	QOS Class 2	QOS Class 3
Audio transfer delay with echo control (*)	< 400 ms	< 400 ms	< 150 ms
Audio frequency range	$\geq 0.3 - 3.4$ kHz	$\geq 0.3 - 3.4$ kHz	$\geq 0.05 - 6.8$ kHz
Audio level (typical)	- 20 dBm0	- 20 dBm0	- 20 dBm0
Audio error free interval	> 5 min.	> 15 min.	> 30 min.
Video transfer delay	(still image only) 10s	< 600 ms	< 250 ms
Video/audio differential delay	N/A	> -400 and < 200 ms	> -150 and < 100 ms
Video frame rate	N/A	≥ 5 frames/s	≥ 25 frames/s
Video resolution	N/A	$\geq 176 \times 144$	$\geq 352 \times 288$
Video error free interval	N/A	> 15 min.	> 30 min.
DSD/audio differential delay	< 1s	< 200 ms	< 100 ms
DSD, error free interval	> 5 min.	> 15 min.	> 30 min.
Data rate	≥ 5 kbit/s	≥ 50 kbit/s	≥ 500 kbit/s

(*) Note 1: An audio transfer delay limit of 25 ms is applicable when supporting connections to conventional telephones without supplementary echo control.

The MMCF table does not say anything about speech quality. A widely recognized methodology for measuring speech quality is defined by ITU-T SG 12 by using Mean Opinion Score (MOS) rating [ITU 96k], [ITU 96l]. The rating of speech quality in technical terms is extremely difficult, because the source information affects strongly the codec performance. Therefore, subjective MOS tests are carried out with different voices (men, women child, elderly persons etc.) speaking different languages, and with other tones such as music. The toll quality, corresponding to MOS values from 3 to 5, is widely considered as a minimum requirement. The need to improve the codec performance is clearly indicated due to impairments stemming from tandeming. There is also other reason to do that. Users are prone to require better quality, because of their expectations. The speech quality of current digital telephone network is far superior to the quality of the past analog one. Formerly users did not expect a better quality, because they were used to an inferior one. *Therefore, as a target for the speech quality of wireless multimedia as well as of Internet telephony should be a MOS value of at least 4.*

The overall end-to-end delay (called also transit delay or transfer delay) derived from ITU-T Rec. G.114 is too long. There is substantial evidence that 200 ms is the upper limit for smooth human interaction. Unfortunately, we face a major technical barrier to meet this requirement in conjunction with the video provided that lip synchronism is required. The processing delay of video codecs is a dynamically varying quantity, which depend on the amount of motion in consecutive video frames, implementation, and coding algorithm. The average value is 200-300 ms. Therefore, if a head-and-shoulders view is represented, and

we want to achieve lip synchronism the speech has to be delayed additionally. Under these circumstances, the overall delay tends to exceed 400 ms. In many applications, lip synchronism is irrelevant, and hence one has to be able to disable it to facilitate human interaction.

The SQCIF video resolution format has been included in H.263. Hence, we have used it instead of QCIF as the minimum value.

The MMCF values for differential delay between video and speech (audiovisual skew) need to be changed. For voice annotation, MMCF has defined a limit for out-of-synchronization, but not for video. Typical case is a remotely controllable pointer and the pointer data is delay sensitive, because it relates to a specific video frame (the synchronization error is called pointer skew). Synchronization errors have been investigated through subjective tests by Stainmetz and Engler [Ste 95] and they have derived the following results:

-80 ms <audiovisual skew < +80 ms

-1000 ms <pointer skew < +1250 ms

These values derived from extensive tests at IBM European Networking Center suggest that respective MMCF recommendations for synchronization of delay sensitive data are over-specified as well. On the other hand voice or textual annotation related to given video frames or data field may impose a more stringent synchronization requirement. In any case, the MMCF values should be taken as generic guidelines, which strongly depend on the source information content.

The IBM results are not directly applicable to low bit rate applications, in which repetition rate of video frames tends to be too low to insure fluid movements in the reproduced video. Given that the time between consecutive video frames can be as large as 200 ms, the above IBM figures are not reasonable. As lower limit for the video frame in Service Class 2, the MMCF rating is > 5Hz, which is rather low and results in a very jerky motion picture. The ITU-T Recommendation H.261 defines 7.5 Hz as a minimum frame rate. We do not see any reason to bargain on that.

The reason for applying the MDC parameter values also in a wireless environment is that presentation media of future wireless terminals is likely to be close to the same as those of desk-tops. This goes also for a multimedia Walkman and other portable multimedia appliances implying that in the future, inferior performance in the wireless realm will not pass unnoticed like today. The users do not have any other expectations now, because of the restricted terminal capabilities. The bandwidth restriction influences quality, but if the bandwidth is the same in the desk-top, similar QoS requirements apply. The results of our QoS study appear in Table 11.

Table 2 Suggested QoS classes for wireless low bit-rate multimedia applications

Source/QoS performance area	QOS Parameter	QOS Class 1	QOS Class 2
E.800,GSM/Service accessibility performance	Mean access delay = time to provide service (GSM)	< 4s ?	< 4s ? (to be tested in the testbed)
E.800/ Service accessibility performance	Terminal negotiation delay (start-up delay)	< 3s ?	< 3s ? (to be tested in the testbed)
E.800/Service accessibility performance	Post selection delay (the time lapsed until the ringing tone is sent)	< 3 s ?	< 3 s ? (to be tested in the testbed)
GSM/Service accessibility performance	Call success rate (mobile originated)	95% ?	95% ?
GSM/Service accessibility performance	Call reception success rate (mobile terminated)	99% ?	99% ?
ITU-T SG12 SQEG/ Propagation performance	Audio transfer delay with echo control (*)	< 200 ms	< 200 ms(to be tested in the testbed)
E.800/Service retainability performance	Service release delay	<u>≤ 2s?</u>	<u>≤2s?</u>

Table 11 (continuation)

GSM/Service retainability performance	Hand-over success rate	<u>98%</u>	<u>98%</u>
Source/QoS performance area	QOS Parameter	QOS Class 1	QOS Class 2
E.800/Service integrity performance	Billing integrity	<u>?</u>	<u>?</u>
ITU-T SG12 SQEG/ System performance	Mean Opinion Score (MOS) of speech quality	≥ 4	≥ 4
E.800/Service integrity performance	Interruption probability	$\leq 1\%$	$\leq 1\%$
GSM/Service integrity performance	Duration of interruption of user data	<u>?</u>	<u>?</u>
MMCF/ Transmission Performance	Audio frequency range	$\geq 0.3 - 3.4$ kHz	$\geq 0.3 - 3.4$ kHz
MMCF/ Transmission Performance	Audio level (typical)	- 20 dBm0	- 20 dBm0
E.800/System Performance	Noise level	? dBm0	? dBm0
E.800/System Performance	Distortion level	? dBm0	? dBm0
E.800/ Transmission Performance	BER for speech	10^{-4} [Rap 95]	10^{-4} [Rap 95]
E.800/ Transmission Performance	BER for video	10^{-6} (with error correction)	10^{-6} (to be tested in the testbed)
E.800/ Transmission Performance	BER for non-transparent data (e.g. TCP based)	10^{-8}	10^{-8}
E.800/System Performance	BER for transparent data	10^{-6}	10^{-6}
MMCF revised/ System performance	Audio error free interval	>0.5 min.	> 0.5 min.
MMCF revised/ System performance	Video transfer delay	(still image only)10s	< 400 ms (to be tested in the testbed)
MMCF revised/ System performance	Video/audio differential delay	N/A	> -150 and < 120 ms

	(skew)		
MMCF revised/ System performance	Video frame rate	N/A	≥ 7.5 frames/s [ITU-T H.261]
MMCF revised/ System performance	Video resolution	N/A	$\geq 128 \times 96$
MMCF revised/ System performance	Video error free interval	N/A	> 0.5 min. (to be tested in the testbed)
MMCF/ System performance	DSD/audio differential delay (skew)	< 1 s	< 200 ms
MMCF revised/ System performance	DSD/video differential delay (skew)	NA	-1000 ms<pointer skew<+1250 ms
MMCF revised/ Network performance	DSD, error free interval	> 0.5 min.	> 0.5 min.
MMCF revised/Network bearer capability	Data rate	≥ 9.6 kbit/s ≤ 28.8 kbps	≥ 28.8 kbit/s

It is clear that the MMCF parameter values are over-specified, at least in terms of existing or forthcoming wireless environments. Tests carried out for G.729 speech codec [Hay 96] indicate that as low as 1% BER is adequate for toll quality speech provided that no tandeming is present. For video, a BER value of 10^{-7} has been mentioned at bit-rates ranging from 64 kbps to 384 kbps [Rap 95]. On low bit-rate connections, 10^{-6} is a more appropriate figure, which translates into an error free interval of 1 minute for 28.8 kbps and 30 seconds for 57.6 kbps. The error robustness of the video is in particular important in mobile environments. This means that any remaining error is not allowed to cause severe impairment of video quality.

Table 3 Relationship between MMCF error free intervals of QoS Class 2 and BER

Bit rate of the wireless connection/kbps	BER
28.8	$3.8 \cdot 10^{-8}$
38.4	$2.8 \cdot 10^{-8}$
48	$2.3 \cdot 10^{-8}$
57.6	$1.9 \cdot 10^{-8}$

The ITU-T and MMCF QoS models are based on circuit switched environments. In a packet switched one, we have also other important parameters influencing the over QoS. Packets will be lost due to buffer over-flow in the network or excessive delay jitter causing a playback deadline miss in the terminal. For real-time applications they lead to non-correctable errors. The probability of a packet loss and a deadline miss should be defined for different service classes of the Integrated Services in the Internet Architecture

[Bra 94]. Packet losses caused by network buffer overflow are beyond user control, but the latter cause of packet loss should be adjustable within certain bounds.

For real-time applications, packet loss of the wireline packet switched network tends to severely degrade the QoS, because a string of consecutive bits is inevitably lost. The size of the receive buffer determines the maximum playback time, which in turn depends on the required delay jitter range. The playback time affects the overall latency. Hence, we have a trade-off between data integrity and latency. Since, as we have seen, a number of other entities introduce delays between the endpoints, the need to keep the playback time as short as possible is urgent for real-time applications. On the other hand, under adverse network conditions, the user may wish to choose rather the impaired latency than put up with artifacts caused by missed playback deadlines. The playback time may also be readjusted automatically by the receiver as network conditions change into favorable ones. A playback time of no more than 100 ms seems appropriate from the user and applications point of view.

8.3 Real-Time Services over Packet Radio Links

8.3.1 Reference Model

The following reference model (see Fig. 26) is derived from ETSI GPRS standard [ETS96b]. The model is GSM oriented, but applicable in principle to other TDMA packet radio systems as well. The basic idea is similar to CDPD i.e. the common infrastructure used for mass scale phone services operates as a core for packet radio services. The highest priority is given to a phone call competing from same radio channel. The phone call will drive the packet radio data away from its channel to search for a free channel. This algorithm works well as long as the amount of packet radio traffic remains low. In the future, we may have to give also for packet radio some minimum service guarantees independently from concurrent mobile telephony traffic. The advantage of the overlay scheme described above is low additional cost compared to the cost of a separate network.

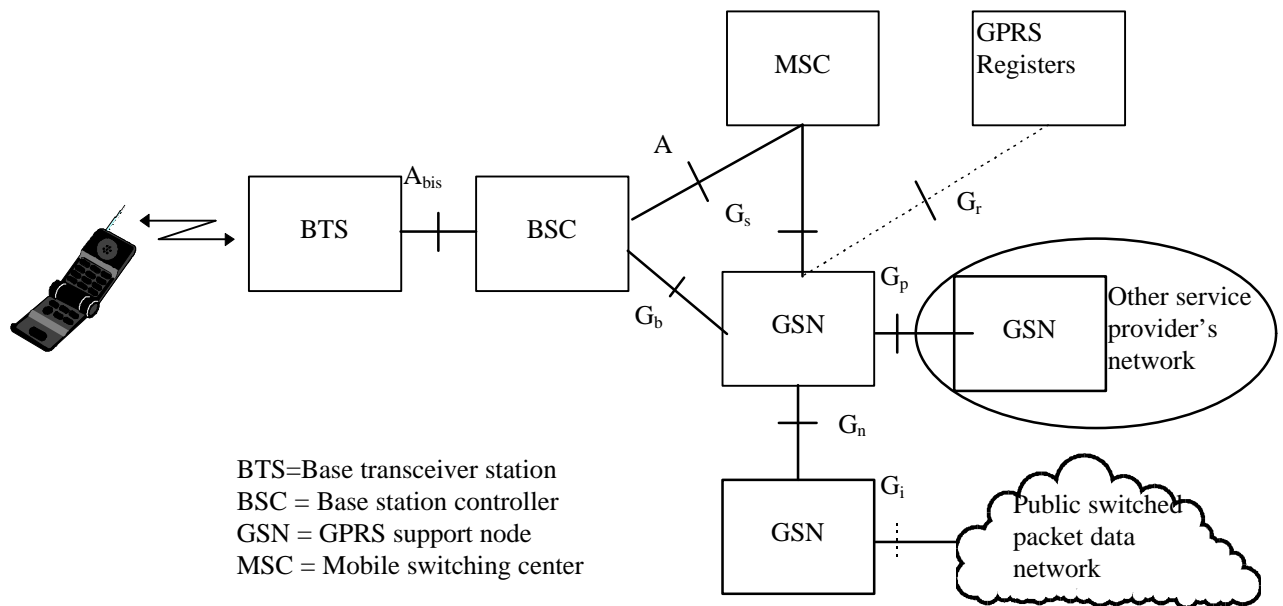


Figure 4 GPRS Reference Model

8.3.2 Routing and Mobility Management

The architectural issues related to mobile computing and communications over packet radio links have been widely studied [Per 96a], [Per 96b], [Joh 96], [Häm 96], [Che 97]. Jari Hämäläinen divides in his doctoral thesis the design problems into three main categories:

- packet routing
- mobility management
- packet radio protocols over the air interface, including access mechanism and dynamic resource allocation.

The subsequent handling of routing and mobility management are based on IP only i.e. a connectionless network layer protocol.

Packet routing and addressing scheme of the Internet has been designed by allocating a permanent Internet address for an stationary Internet host. This entails that every point of attachment to the network has an address, which consists of subnetwork identity and host identity. The first part of the address is used for routing the packets to the right destination network and latter part by the destination network only to route the packets to the selected host [Häm 96]. For mobile hosts, moving away from their home location, the point of attachment has to be changed via their home subnetwork.

Some years ago IETF launched a specific mobile IP activity to address the routing problems of a *mobile node*, which is a host or a router capable of changing its point of attachment. The results appear in [Per 96a]. Fig. 27 illustrates the basic idea of mobile IP routing. The packets are always routed to the destination subnetwork identity i.e. in the case of a mobile node. The basic idea of the scheme is that a temporary identity i.e. so called *care-of address* will be allocated by so called *foreign agent* of the foreign network to the mobile node, when it is absent from its home network. The foreign agent is a server

having an address pool for the visiting mobile nodes. After having received a care-of-address, the mobile node registers its mobility binding at that instant with its home agent by sending a registration request.

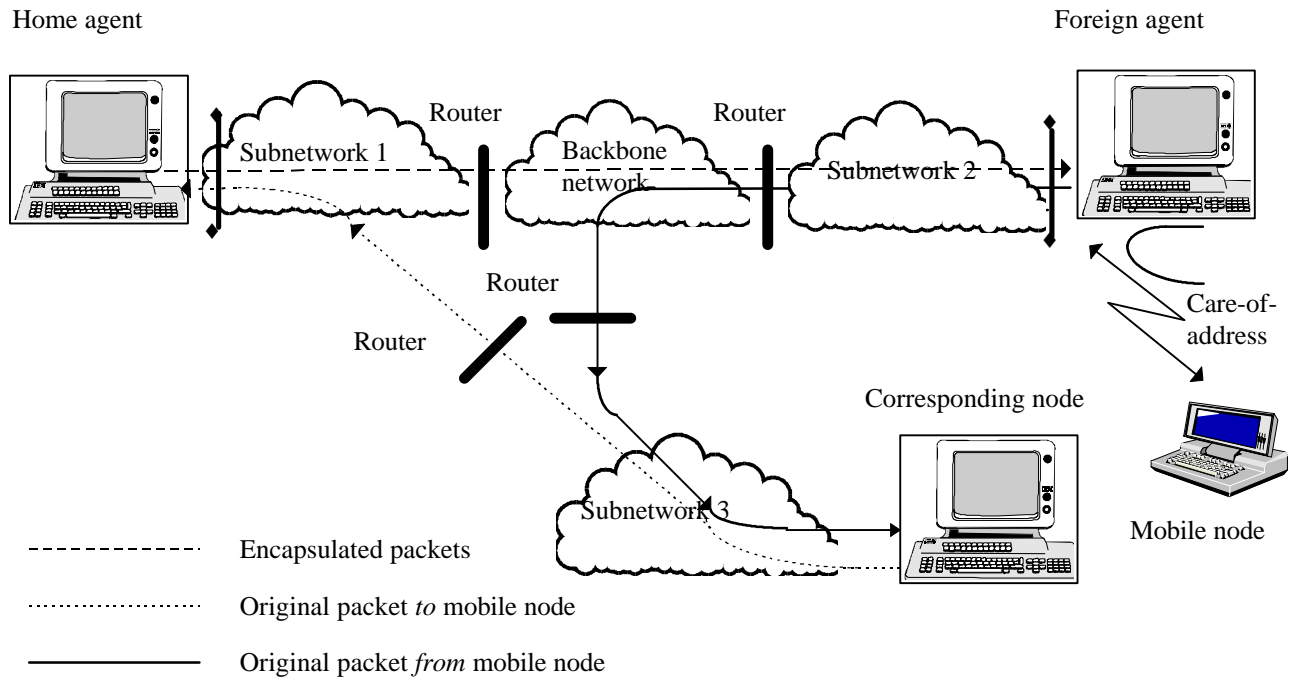


Figure 5 Routing of packets using the mobile IP protocol [Häm 96, Per 96a]

The home agent authenticates the request and sends an acknowledgment to the mobile node either directly or over the foreign agent. The node wishing to establish binding with the mobile node is called a *corresponding node*. The packets destined to a mobile node visiting a foreign network are encapsulated in another datagram [Per 96b]. The home agent of the mobile node validates security elements, keeps track of its mobility bindings and encapsulates packet streams for delivery to the mobile node. During the transfer of packets, the home agent intercepts the packets it has mobility binding with. Either the packets are delivered directly to the mobile node located within the home network or they are encapsulated for delivery to a foreign network. The encapsulated care-of-address is stripped off by the foreign agent and the original data is delivered to the desired application of the mobile node. The mobile node transmits packets to the correspondent node directly without any encapsulation.

There are some routing problems, which we are not going to describe in more detail in this context. In IPv6 they have been solved in such a manner that optimal routing can be supported. The main problem that remains to be solved is the protocol issue, to avoid additional latency and delay jitter. The BER of raw data is insufficient for acceptable speech and multimedia service quality. BERs of the order 10^{-5} - 10^{-6} should be aimed at. For the moment the GPRS protocols are designed for datagram delivery only i.e. for delay

tolerant applications. TCP/IP protocol need to be attuned to the mobile environment To keep additional overhead at its minimum, we need to compress the TCP header data [Häm 96]. In any case, the packets arriving from Internet have been delayed by the routing hops they have had to cross.

The mobile IP optimal routing capabilities will not be available before IPv6 is widely deployed. Routing the packets sent by the corresponding node always via the home agent results in excessive delays, which make mobile IP even less attractive to real-time communications. We can expect the current long routing delays of the Internet to reduce significantly in the future. Yet, it is not clear, whether mobile IP will be applicable to real-time delay intolerant applications.

8.3.3 Multiple Access over Packet Radio

A packet switched overlay technology can be applied in digital cellular network as well. This is the basis for GPRS. Furthermore, we can apply a multi-slot reservation mechanism for supporting also high speed packet switched data services [Hon 94]. The overlay technology is a very cost effective way of offering packet radio services, because the underlying physical infrastructure, primarily used for mobile phone services, remains basically intact. The packet switched mode requires different protocols. Due to the fact that packets are usually fairly short, bit-interleaving depth can be reduced and thus cut down overhead and propagation delay. The MAC protocol causes inevitably some delay, because of collisions with packets from other users. A multiple access reservation mechanism based on Aloha, slotted Aloha [Rap 96] or CSMA is commonly used [Rap 96]. For the latter one, detection and propagation delays are the most critical parameters. In this context, propagation delay stands for the relative time needed for a packet to travel from a base station to a mobile terminal. The detection delay is an implementation dependent variable. It denotes the time needed for a receiver to sense an idle channel. When the receiver has detected an idle channel, it starts sending packets. However, before the packets have traveled to the base stations and detected by potential other stations wishing to transmit, the other stations may start to access concurrently the same channel, leading to a collision. The propagation delay normalized into packet transmission units can be expressed by following formula:

$$t_d = \frac{t_p * R_b}{m}$$

t_p = propagation time in seconds
 R_b = channel bit rate
 m = number bits in a data packet

The overall delay i.e. latency is the most critical parameter for real-time applications. It is a trade-off between error robustness and data buffering. As a rule, a packet switched link is used by elastic, delay tolerant applications, which require error free transmission. The link access protocol has to be designed based on this precondition to insure correct operation at the IP level. In addition to underlying channel coding, error recovery is needed and ARQ is commonly used. However, ARQ is less suitable for real-time applications, because of large buffer sizes needed for delay jitter compensation and

increased propagation delay. Furthermore, for data, additional error control is needed at the transport layer to insure reliable transmission.

8.3.4 Voice and Multimedia over Wireless IP

Voice-over-IP is technically possible already in the 2nd generation digital networks using the packet switched mode. There are some obstacles however, which cannot be easily eliminated, at least not in current market situation. Like the Internet in its current state, a packet radio network does not give any service guarantees for delay intolerant applications such as voice and video. Ramifications of a best effort based packet transfer are probably more severe in a wireless environment.

Technically, using packet radio links for telephony may be possible, but it is unclear whether there is any economical justification for that. As we have noted previously, large buffers are needed for the real-time speech to accommodate delay variations introduced both by the Internet and the packet radio link. Besides burst errors caused by a fading radio channel, collisions of packets cause additional errors. The packet protocols such as TCP are designed for reliable transfer of non-transparent data generated by delay tolerant applications. For speech and LBR multimedia applications, we need a protocol on top of UDP, which enables a residual bit error rate of the order 10^{-6} , without clogging up the throughput. Note that the design value for plain speech is 10^{-4} in wireless networks [Rap 95]. However, the multimedia capability i.e. the video requires significantly better overall error performance. A combination of FEC and ARQ may be needed to meet the requirement. The FEC part of the error control handles the huge majority of errors. Those error patterns, which are likely to result in FEC coding errors, are handled by the ARQ. Provided that the number of retransmission can be kept low enough, the additional delay is only a fraction of the overall delay. However, a better error handling capability can be achieved only at the expense of additional overhead, which translates into lower link throughput or additional cost. The current wireless networks are optimized for speech. To be competitive, voice over packet radio needs other incentives rather than cost such as more versatile functionality or provision of a large plethora of IP based multimedia applications to the nomadic users. However, the error-prone low bit-rate wireless link permits this in a limited scale only

The propagation delay of the Internet depends on the number of router hops and traffic load. To get some idea about the delays encountered in the Internet in terms of best effort datagram delivery, we have measured Internet transit delays using the Netmedic software from Vitalsigns Software. The results appear in Table 13. As expected, the variations of Internet transit delays are huge. Therefore, the receive buffer should be able to automatically adjust its playback time for optimum performance as VoIP has proposed [VoIP 97]. The router delays are on the wane as more effective routers are launched in the market, but a transit period of several years is needed before a significant improvement can be expected in the overall network performance. Besides the Internet, site servers incur long delays. Therefore, also new servers are needed, which are designed to meet the latency requirements of real-time services.

Table 4 Transit delays of the Internet

Test	Average delay (ms)	Min [ms]	Max [ms]	Number of router hops
1	52	24	167	4
2	127	7	151	6
3	157	4	249	8
4	170	21	241	10
5	225	4	445	12
Mean/hop	18			

We have made some calculations of the overall delays for Internet telephony and multimedia applications based on a scenario in which a H.323/M encounters a H.324 terminal. (see Fig. 28 and Table 14). We assume that wireless packet radio links of 2nd generation digital wireless systems are used. The IP network is assumed to be the Internet operating in best-effort service mode. Packetization delay depends on the packet length. For real-time traffic in low bit-rate networks, the packet size should be preferably only a few bytes. The presented configuration supports full-fledged multimedia communication, but it can be used for speech only.

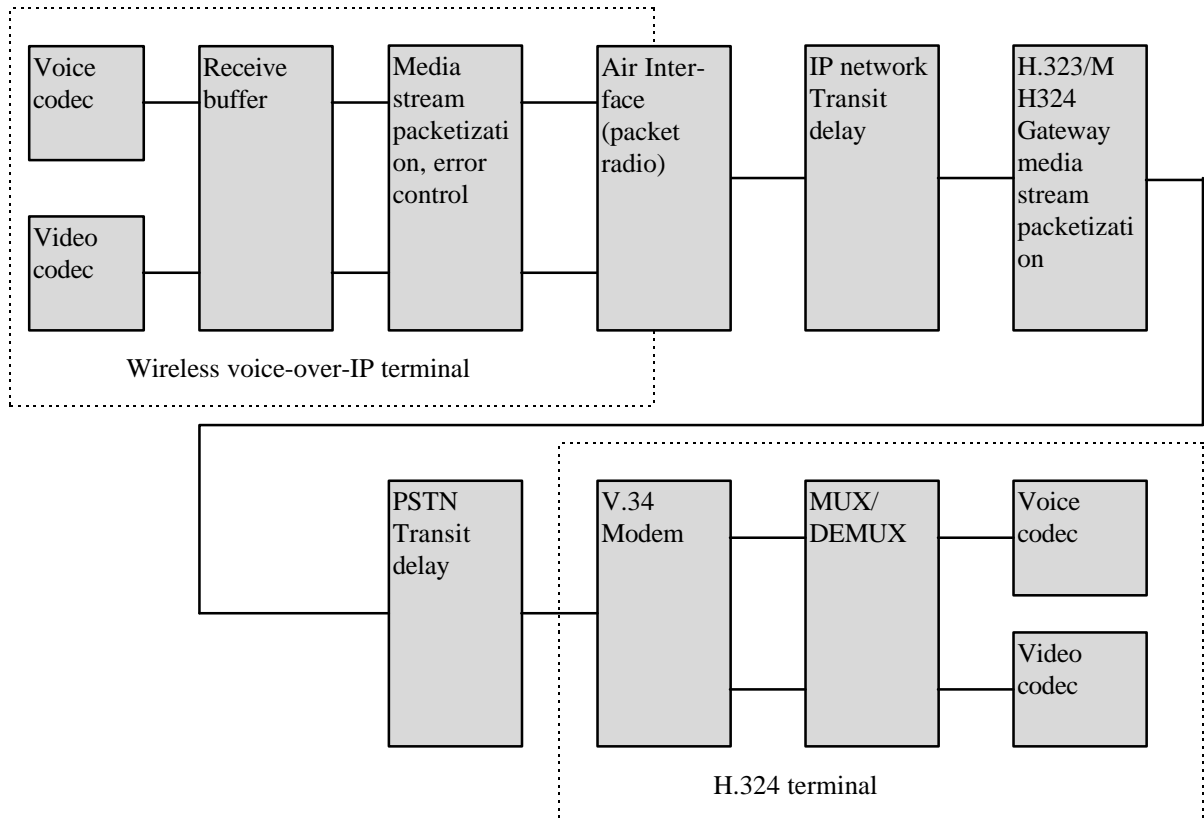


Figure 6 Delay introducing entities in a call between wireless IP and H.324 terminals

Table 5 Delay calculation of the configuration depicted in Fig. 28 [mobile IP to PSTN]

	low-tier packet radio (PACS)	high-tier cellular, GPRS
Delay introducing entity	Delay (ms), end-to-end	Delay (ms)
G.723.1 speech codec	80	80
Receiver buffer [delay jitter compens.]	10-100	10-100
Error control	30	30
Media stream packetization (wireless IP terminal and gateway)	15	40
Channel coding, equalization and bit interleaving	3	70
Packet protocol delay (Eg. CSMA/CD)	20	20
Avg. fixed IP network transit delay [depending on the number of router hops]	30-250	30-250
PSTN transit delay	10	10
H.223 MUX/DEMUX + modem	20	20
Total	218 - 528	310 - 620

Note The figures in Table 14 are rough estimates and should be treated as such. However, we believe that potential inaccuracies are low enough not to change subsequent conclusions

The results suggest that current wireless networks do not fit well for supporting voice over packet radio links as long as the transit delays of fixed IP networks cannot be kept reasonably low. Even if the delay would be 0, the 200 ms limit given by SQEG of ITU-T SG12 cannot be met. As we have pointed out already in section 4, the best candidate is a low-tier packet radio environment. This does not yet mean that such voice-over-IP would be feasible for other reasons. In case G.723.1 is used for speech coding, the Internet portion of the delay and delay variations need to be reduced substantially to keep the overall delay close to the recommended maximum delay 200 ms. In the short and medium term, this is not likely. Moreover results of our studies lead to the following conclusions:

- It is highly questionable, whether QoS of applications with conflicting requirements for delay and error tolerances can be controlled in a packet radio environment given that this is already very difficult in wireline TCP/IP networks
- Unlike the fixed IP network, wireless packet radio lacks the cost incentive. Therefore, wireless voice-over-IP for *speech-only* applications seems unrealistic.
- The current technology seems inadequate for real-time multimedia applications over low bit-rate wireless packet radio. Even if applications of low degree of interaction and without lip-synchronism would turn out technically feasible, HSCSD may offer a more competitive solution.
- In the future, market demand of the previously mentioned applications may justify dedicated packet radio networks, which are capable of meeting better the needs of wireless real-time multimedia applications. For the time being, we cannot envision such a demand. Nor are we convinced that packet radio is the right vehicle for real-time information transfer.

GPRS and other digital cellular overlay technologies enable moderately high data rates. The packet switched mode will be used mainly by non-real-time, delay tolerant and error intolerant applications. Possibly Web browsing with on-line speech will be offered to the users. In this case, users may accept longer round-trip delays, because less interaction is probably needed than in a normal POTS conversation.

8.4 Real-Time Services on HCSCD Links

8.4.1 Reference Model

The system architecture for GSM HSCSD has been depicted in Fig 4. The scheme is based on combining and splitting multiple subchannels each transmitted in its own time-slot. Both transparent and non-transparent data transfer is supported. The transparent one employs ISDN related V.110 protocol. The non-transparent mode uses an ARQ based RLP (Radio link protocol) between a MS and an IWF. Due to the low BER of raw data, retransmissions occur frequently, which translates into long overall delays due to protracted delay variations and playback times. Therefore, the non-transparent mode is unsuitable for real-time applications. In the call establishment phase, a user needs to define maximum and minimum limits for the desired capacity. In practice, the application itself sets the parameter values without any user intervention. The difference between the maximum and minimum represent the window for acceptable QoS.

8.4.2 Voice and Multimedia over Wireless Circuit Switched links to IP

Subsequently we have made some basic calculations on delays in a potential scenario in which a wireless LBR multimedia terminal (H.324/M) encounters a wireline multimedia IP terminal (H.323). The subsequent Figure 29 depicts the prime delay introducing entities of the call path. Like the previous configuration, the terminals have a full multimedia capability, which can be used in a limited scale i.e. for speech only.

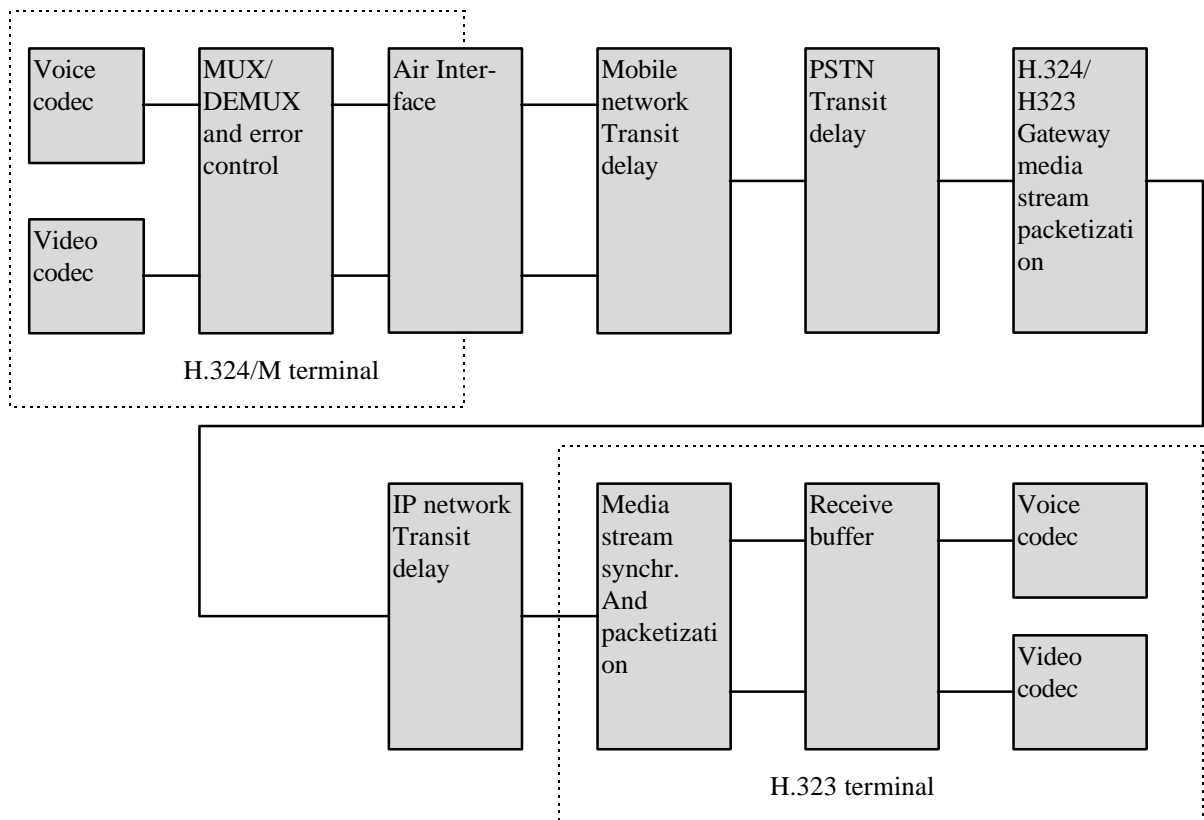


Figure 7 Delay introducing entities in a call between H.324/M and H.323 terminals

Table 6 Delay calculation of the configuration depicted in Fig. 29 [CS digital mobile to IP]

	low-tier wireless (PACS)	high-tier cellular, (GSM HSCSD)
Delay introducing entity	Delay (ms), end-to-end	Delay (ms)
Voice codec [G.723.1]	80	80
Error control	0	40
Channel coding & equalization, bit interleaving	3	90
PSTN and Mobile transit delay	10	10
Receiver buffer [delay jitter compens.]	20-100	20-100
IP networks [transit delay]	30-300	30-300
Media stream packetization (IP terminal + gateway)	40	40
Total	183-533	310 - 660

Note The figures in Table 15 are rough estimates and should be treated as such. However, we believe that potential inaccuracies would not change the subsequent conclusion

What we have noted above about wireless Internet telephony using G.723.1 speech codecs over high-tier packet radio links goes also for H.324/M systems i.e. real-time multimedia systems over circuit switched mobile links (see Table 15). The error control is optional and needed for multimedia content only, in particular under hostile network conditions, typical of high-tier networks. The use of multiple wireless links incurs additional delay as does potential transcoding. However, in our scenario, we have assumed that no transcoding takes place in the network. This is not needed from the service point of view, since both terminals support G.723.1. Furthermore, for plain voice, neither error control, nor multiple links are needed.

The use of H.324/M systems for interactive multimedia communication between human beings is constrained by long delays per force on multiple HSCSD links of GSM. Even for voice only the delay is significantly longer than what is considered acceptable from fluid human interaction point of view. When the video is additionally present and an application requires lip synchronism, the coding delay is even longer. Video coding delay is usually a dynamically varying quantity, but under normal conditions we can assess that the average delay is of the order 200-250 ms. Therefore, we derive from the calculations an overall one-way delay that is close around 800 - 900 ms under adverse network conditions. This exceeds by a substantial margin the maximum overall delay 400 ms determined by G.114,

which is already too large for human interaction. The delay conditions of low-tier wireless networks are comparable to those of the PSTN. It is obvious that without delay control in the IP network the overall delay becomes too long even for voice-only communication.

8.4.3 H.323/M Scenario: Using CS Air Interface for Packet Transfer

As we have noted in Section 7, ITU-T SG16 the possibility of using PSTN copper lines or ISDN for Internet access of H.323 terminals is under investigation. We discuss briefly a respective idea of using a circuit switched multi link air interface for packet transfer. Such a concept has some attraction in the sense that IP would offer more versatile functionality.

FEC based error control mechanism is needed to protect IP headers. This option includes a number of difficult questions that need to be addressed, before any further actions are taken. Albeit being technically feasible, this might lead into an economically non-viable concept. The expensive multi-link air interface is reserved during the whole duration of the call. Therefore, the service would address mainly business users. It is likely that within a few years there will be quite a large installed base of H.323 terminals. At that point of time, the need to provide a IP based wireless access to those corporate LAN based terminals would increase substantially. Furthermore, a potential PSTN H.323/P standard and emerging applications might stimulate the demand for H.323/M. The forthcoming H.324/M system can do part of the job, but it is incapable of measuring up the functionality that a wireless IP based system could offer. A configuration of such a H.323/M scenario, in which the other terminal would be a PC interconnected to a corporate LAN, is depicted in Fig. 30.

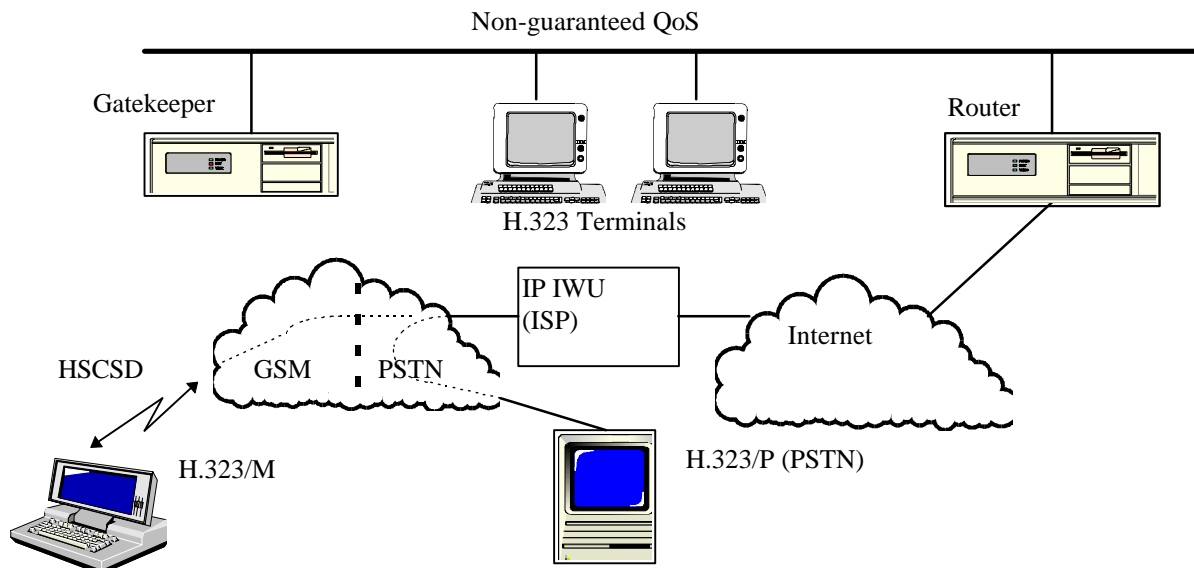


Figure 8 A scenario of H.323 potential developments

The mobile link is used in similar manner as the PSTN line i.e. only as access to the Internet. The routing can be done directly to the IP IWU or across PSTN as in the picture. In the latter case, the data transmission scheme on PSTN sets the limit for the maximum desired data rate. The GSM MSC may be interconnected via a 64 kbps links to the IP IWU.

A wireless H.323 system for supporting real-time multimedia services is not realistic based on technology currently available or envisioned. Solving the problems of keeping delays within reasonable limits and controlling QoS seem currently too tough a challenge.

Voice Web browsing may become the killer application. A protocol for such a speech + data mode over mobile IP would be needed as well. Like in H.323 system, voice packets can be sent over UDP without any retransmissions. However, further studies are needed to proof the feasibility of such a Web application at enhanced data rates for example those provided by GSM HSCSD or GPRS.

8.4.4 Addressing and Numbering

Addressing in telecommunications networks is determined by ITU-T Recommendation E.164 [ITU 91]. The same principles that are valid for PSTN also apply for wireless networks [Alt 97]. As a matter fact, the numbering scheme for public land-mobile networks defined in ITU-T Rec. E.212 follows similar numbering structure as E.164.

The VoIP Forum of IMTC has defined three basic cases that need to be covered in the numbering and addressing of Internet telephony:

- E.164-like number ==> IP + e-mail-like address (call from POTS to an IP based terminal)
- E-mail-like address ==> IP + E.164-like number (eg. Call from IP based terminal to POTS)
- E.164-like number ==> IP + E.164-like number (A POTS to POTS call over H.323 gateways).

The most relevant case is a call from a mobile phone to an H.323 phone or a call from a H.324/M to a H.323 phone across appropriate gateways. The second bullet point i.e. a call from a mobile IP based terminal is not relevant in the short or medium term. If and when such wireless IP based phone services emerge, the IETF mobile IP routing scheme [Per 96a] may be applied. From addressing point of view, the scheme is transparent.

The above approach is based on the assumption that Internet telephony employs H.323 standard, which is based on a LAN access. In wireless environments, the access is not likely to be a cordless LAN, but a packet radio air interface of a public network. The sub-addressing and address translations executed by the gatekeeper will be replaced by a separate functional module in the mobile Internet telephony terminal. E.164-like numbers can be allocated as Aliases for IP addresses in similar fashion as in the fixed IP networks.

Support for mobility and roaming are required in mobile environments. Furthermore, nomadic users need respective directory services as those provided for POTS users.

8.5 User Data Management

In this Section our intention is to deal with the management architectures in particular from the end-user point of view. Therefore, we are not going to handle the general IN architecture and resource management problems encountered by wireless operators in a mixed operational environment. Currently, service management of IN architecture is service provider centric. Networks are composed of elements provided by multiple vendors, which from the network management point of view makes things difficult. The network management system must be able to access and share data related to billing, security, user profile etc. in a mixed operational environment, consisting of multiple service providers, mixed air interfaces complying to different standards [Gar 96].

A huge effort bearing the notation TINA-C [TIN 97] is being taken by large group of telecom operators and manufacturers to develop a common service architecture for multimedia and information services. TINA-C aims at managing services in a complex multi-vendor environment comprising traditional telecommunication networks and the Internet. The TINA-C separates the overall architecture in three entities. The first one is the DPE distributed processing environment, which comprise applications and the computing environment where the application run. The second architectural entity is the service architecture, which define different processes and interfaces for supporting coherent provision of services across different networks. The third architectural entity consists of a model for setting up connections and managing transport network operations. TINA-C is an important development for future wireless communications as well, which is particularly challenging for all levels of management.

The new TINA-C architecture enable dismantling of the binding between user and communication endpoint, which means that both user mobility and terminal mobility can be supported. In our model, the MMW is treated as a personal device, but this should not preclude the possibility of another person transferring his/hers identity to the terminal. Furthermore, a MMW user may wish to access a payterminal across an infrared link. In this case, the user identity will be transferred to the payterminal, the user using his/hers own terminal as an access device. [Eck 96] defines similar architecture titled PCSS (Personal Communications Support System) with special emphasis on personal mobility, service personalization and advanced service interoperability. We believe that the personalization of communication service mix i.e. a *personal service profile* is a crucial success factor for personal multimedia communication space, which would enable combining economies of scale and service customization in accordance with specific user needs.

8.5.1 Self-Configuration of Service Profile

In the future, it is apparent that the role and responsibility of the user to manage her or his own personal service profile and user data needs to be enhanced significantly. This is

particularly important in such an environment, where the service mix and content can vary within a broad range and where reconfigurations occur frequently due to constant release of new Web applications and content. As we have seen, Internet telephony is not only a telephone application, but a collection of different multimedia capabilities, which can be customized to meet a wide variety of user needs and preferences. In our model, we have drawn a scenario of the MMW, a personal multimedia information appliance. The ability to reconfigure the user profile whenever desired by the user is clearly indicated in this kind of highly versatile and rapidly changing personal communication environment. It is far too cumbersome and costly to let the service providers to do this on behalf of the user.

The mobility management capability is an inherent feature of wireless networks, corresponding to *terminal mobility*. User data is permanently stored in the home location register HLR. This data is temporarily copied at cell setup in the visiting locator register VLR of the mobile switching center MSC where the user has roamed to. After user has registered to VLR, all database operations are executed locally without any undue delays [Gar 97]. The briefly described HLR/VLR architecture is the basis for GSM and IS-41 for North America. *Personal mobility* is the user's ability to access subscribed services (i.e. in accordance with the user's service profile) from any terminal. In MMW case, we may wish to transfer our personal identifier to a wireline residential phone and use our MMW as its cordless extension. This is, however, a much more complex operation, because almost all parts of the network infrastructure, including the fixed network, are involved in this management process. Therefore uniform and consistent architecture is needed. The third form of mobility is *service mobility*, which means the user's ability to use also the customized service features in when roaming beyond his/hers home network. All these forms of mobility management need to be supported.

[Eck 96] defines a specific set of capabilities, UPMS (User profile Management Service) to manage the user profile. The UPMS include the following basic management functions:

- *authentication* for secure access to user's own data
- *user profile configuration* for modifying the data
- *user registration* for manually updating user's location data
- personal schedule configuration for scheduling the location data
- communication configuration for customization of personal call logic.

Editing the personal call logic enable the user to restrict incoming calls according to given call logic rules. For a personal terminal, the two first bullet points are the most important ones. It is nice to have all the versatility of the UPMS architecture, but the management of user profile tends to become too complicated from average consumers point of view and we doubt that the amount of usage would justify the cost of such increased level of sophistication. The location information update is automatic as far as the user has a personal terminal. The wireless network is aware of the base station user's location through location update signaling. A GPS transceiver opens new opportunities for making use of location data independently from the network management. For instance closed user groups could be established, where locations of the other members would be transmitted to each member.

A Network centric data base architecture has many disadvantages such as:

- Increased complexity, size and cost
- increased real-time internetwork communications burdening heavily the network.
- Increased latency of the database operations impairing the QoS
- inefficient use of network resources

For these reasons, a smart card approach suggested by [Gar 97] is particularly attractive. The user data would be stored in a smart card that can be loaded in a smart card reader of the mobile terminal itself. The changes in user profile are made by the user whenever desired. Of course a smart card oriented approach does not replace completely the network centric management of user data. It is always involved in some form or another such as for location and billing data, but all such information that needs not necessarily be managed in the network should be kept on the smart card.

8.5.2 Customization of Application Content

It apparent that the way we perceive information retrieval is changing. Traditionally it has been "pull" oriented, but gradually users are willing to receive the information, which is important to them with as little effort as possible. Since the amount of garbage, i.e. irrelevant information from an individual user point of view, is increasing drastically, there is a clear-cut market opportunity for various Web "push" technologies. There is always a snag when such information is "pushed" across a precious air interface. The control of the "pushed" data is less effective than for "pulled" one. Filters to single out wheat from the chaff need to be very effective. Users do not want to pay anything for meaningless stuff. Many of the available Web "push"-technologies support customization, but they are

developed from the wireline Internet access point of view. For mobile environments we need additional tools such as instructible user agents with advance knowledge of the user's behavior and preferences.

8.6 Charging and Billing Principles

Charging and billing of the Internet is a big issue, which constitutes an inherent part of the overall architecture. In this context, we deal with the topic from the wireless link and mobile user point of view only.

8.6.1 Wireless Link

The air interface is expensive and it remains to be such for some time. Currently all calls are metered in cellular networks i.e. the pricing is usage sensitive. Flat rate charging is unrealistic in short and medium term. In circuit switched multi-link operation, users are likely to be charged in proportion to the amount of links reserved. This is not entirely clear-cut, because the network resources needed for multi-link operation is not the same than if same amount of single links are allocated to separate users. In the latter case, call setup procedures are multiplied, which means that the network resources serving the subscriber level and call signaling need to be multiplied as well. On the other hand, the multi-link operation incurs additional cost on the MSC, which probably weighs up the cost savings of call signaling.

In packet radio networks, the charging is usually proportional to the amount of packets transferred over the link. For the time being, the charging rates are fairly high, probably due to the low market volume of mobile data. Since it is very likely that multimedia Walkman users need to down-load applications from the network, the expensive and low bit-rate air interface is bound to limit the applications to light ones.

8.6.2 Advice of Charge

Advice of charge is an important matter, which has not yet gained the attention it would have deserved. Today the charging rules are too complicated even for POTS. Therefore, we can imagine that for complex audiovisual services the situation would get completely out of control unless appropriate support services are made available for end-users. We can envision that in the future, when portion of the billing is allocated for content, a portion for the services, a portion of used network capabilities affecting the QoS with a large number of different combinations, the users definitely need itemized advice of charge during different phases of the call i.e. prior the call, during the call and after the call has been terminated. Introduction of standard set of profiles does not only serve the user in a better way, it is also technically reasonable, because the negotiation time of the terminal capabilities at the beginning of the call can be reduced to a fraction.

In mobile environments, charging may vary depending on the operator, roaming conditions, and used air interface, which confuses additionally the user. The transfer of the billing information across network boundaries is another essential requirement that need to be considered on designing the overall management architecture.

As far as multimedia services are concerned, the network should send a sample of the audiovisual quality corresponding to the selected profile together with charging rates.

8.7 Electronic Commerce and Payment

Electronic commerce and shopping across the Internet directly from a portable personal terminal sounds a bit eccentric, but within a few years a substantial number of users will place orders and pay for goods and services across a wireless link. Electronic commerce is rapidly gaining more popularity, for the time being, across a wireline Internet access only. The key success factors, security and reliability of electronic transactions are even more crucial issues in wireless environments due to their vulnerability to eavesdropping and authentication frauds. This goes also for the Internet, but a very high level of security is required at least for direct transactions from bank accounts. Encryption is a must in most cases.

The presented multimedia Walkman scenario is consumer oriented and therefore, payment methods such as electronic currency (cash), credit card, and electronic checks are important. Since the price of the content (e.g. a Java applet) residing in the Internet may vary from a few cents to a few dollars, a system such as Millicent from Digital Equipment Corporation supporting micro payment is needed. Because of the high performance required, public key cryptography is not feasible.

Air interfaces of wireless networks are metered. Therefore, content may be billed as a kiosk service as well i.e. the wireless operator bills the subscriber for the call surcharging the portion of the content provider. This is the traditional way of handling micro payments in the telecommunication networks. However, It is likely that a universal system for micro payments, independent from network billing system, will gradually gain a dominant position.

Other important issues for a consumer are anonymity, which is protected fairly well by most electronic currency methods [Mck 97]. For check and credit card payment the identity of both the payer and payee must be known. The public key encryption method can be used to encrypt the credit card number and other relevant data. An electronic signature of the payer is needed for both checks and credit cards. In the former case, also an endorsement is required in electronic form from the payee.

The technology to support electronic payments reliably and securely is at hand. However, it is likely that consumers are not prone to change their habits in money transactions very quickly regardless of convenience and ease of use, in particular in a wireless environment.