

SOME RESEARCH ISSUES IN A HETEROGENEOUS TERMINAL AND TRANSPORT ENVIRONMENT FOR MULTIMEDIA SERVICES

Paul Haskell

Compression Laboratories, Inc., USA

David G. Messerschmitt

Department of Electrical Engineering and Computer Sciences
University of California at Berkeley, USA

Abstract

Continuous-media (CM) services like voice, audio, video, and animation utilize three primary signal-processing operations -- compression, encryption, and error-correction coding -- that have a substantial impact on the network architecture. Future networks will be heterogeneous, consisting of combinations of different types of subnets, such as wireless access, the public telephone network, Internet, and broadband ATM. Even in the distant future, we expect wireless access to a broadband network to be common. We point out a number of other issues relating to the signal processing aspects of CM services, with particular emphasis on the traffic efficiency of wireless access links, low delay, high subjective quality, and privacy by end-to-end encryption. We point out that popular but simplistic approaches involving the use of transcoding (conversion from one compression standard to another) have a number of undesirable characteristics, among them the realization of a network infrastructure relatively closed to change and inconsistent with privacy. We define a networking framework based on a "medley gateway" between heterogeneous subnets that is open to new services, allows privacy (end-to-end encryption under user control), and good traffic efficiency on all links of the network (through joint source/channel coding as appropriate). The medley gateway also opens up new possibilities for exploiting network characteristics in CM services such as video. A key feature is a substream structure that makes certain critical properties of the source visible within the network, even with encryption. We mention a number of open issues relating to resource allocation in session establishment and the design of medley source and transport elements.

1. INTRODUCTION

The idea of an "open system architecture" has been popularized by the Internet, which allows applications and services to be defined by users or third-party vendors transparently across LANs, MANs, and WANs. However, the Internet and its internet protocol (IP) were defined for non-real-time services, and are not suitable for continuous-media (CM) services like voice and video, at least under high traffic conditions, because they do not offer quality of service (QOS) guarantees. Efforts to extend IP to offer real-time guarantees by appropriate resource reservations are underway [3][4]. The real-time aspect is but one major issue confronting high-speed networks of the future [1]. Simultaneously, there is underway an effort to define a broad "Global Information Infrastructure (GII)" that subsumes and extends many existing networks [2]. While some existing networks do not interoperate gracefully¹, within the context of the future GII one goal is to allow applications and services to span a diverse and heterogeneous set of subnetworks [1][2].

In the realization of CM services, as distinct from other services, there are three critical signal processing technologies: *compression*, *forward error-correction coding (FEC)*, and *encryption*. These signal processing technologies modify or hide basic syntactical and semantic components of a bit stream. Subjective quality is important for CM services, and is affected by both signal processing and transmission impairments. Signal-processing considerations should thus play a major role in decisions about network architecture. A goal of this

1. For example, how would one place a telephone call through the Internet and the public telephone network?

paper is to understand qualitatively the implications of signal processing considerations to network architectures, for networks capable of achieving high traffic efficiency and high subjective quality.

These issues are critical on wireless access subnetworks. Radio physical media resources (like bandwidth and transmitted power) are in limited supply, and the traffic capacity is thereby limited. Obtaining consistently high reliability on wireless links is expensive, but high bit error rates can jeopardize subjective quality. Wireless access links will therefore be the bottleneck that limit the subjective quality of CM services. It is important to consider the needs of wireless access links in the design of network architectures, and we make some observations about the architecture of backbone networks for the accommodation of these links.

2. BASIC CONSIDERATIONS

In this section we discuss briefly and qualitatively some of the interactions between signal processing functions and the transport networks within which they are embedded.

There are some fundamental syntactical constraints that we should keep in mind while designing a network architecture for CM services. In particular, operations must be performed in the following order: compression, binary FEC, encryption, binary or signal space error-correction coding and decoding, decompression, decoding of the binary FEC, decryption, and decompression. This is for the following reasons:

- Compression (and other signal processing performed as a part of the service semantics) must precede encryption (for privacy) and decryption must precede decompression. Encryption would hide basic statistical characteristics of an uncompressed audio or video signal, such as spatial and temporal correlations, that are heavily exploited by compression algorithms.
- Compression must precede FEC and decompression must follow FEC, since otherwise the FEC would attempt to correct for transformation in the signal bits by the compression/decompression algorithm.
- We divide error-correction coding into two classes: *binary* (such as algebraic and convolutional coding) that transform a bit stream into another bit stream, and *signal space* (such as trellis and lattice coding) that are integrated into a modulation system and involve Euclidean-space manipulations [14]. Since encryption, like binary coding,

transforms one bit stream into another, it can precede or follow binary error-correction coding. However, since a signal-space code generates an output in the real-number field, it cannot precede encryption and signal space decoding cannot follow decryption.

Packet transport networks inevitably introduce three types of impairments. There is packet *loss* (failure to arrive), packet *corruption* (bit errors occurring within the payload), and packet *delay*. Packet loss can occur due to several mechanisms, such as bit errors in the header or buffer overflow during periods of network congestion. Data networks do not make a distinction between loss and corruption, since a packet that is corrupted is useless and hence is discarded. CM services can tolerate some level of loss and corruption without undue subjective impairment. Lost data must be masked, for example in video by repeating information from a previous frame, or in audio substituting a zero-level signal. Under some circumstances it is possible to make good use of corrupted information, for example by displaying it as if it were correct. The resulting subjective impairment may be less severe than if the corrupted data were discarded and masked.

Some CM compression standards, generally those presuming a reliable transport mechanism (such as MPEG video [6]) discard corrupted data and attempt to mask the discarded information. Other standards -- those designed for a very unreliable transport (such as the voice compression in digital cellular telephony [18] and video compression designed for multiple access wireless applications [17]) -- use corrupted data as if it were error-free, and minimize the subjective impact of the errors.

CM services are real-time, meaning that they require transport-delay bounds. However, there is a wide variation in delay tolerance depending on the application. For example, a video-on-demand application will be relatively tolerant of delay, whereas it is critical that transport delay be very small (on the order of 50 msec or so) for a multimedia editing or video conferencing application. Much recent attention is focused on achieving bounded delay through appropriate resource reservation protocols [3][4].

Joint source/channel coding is a way to increase the traffic capacity of a network subject to a subjective quality objective. It is viewed differently from the perspective of the "source" and the "channel", where channel is usually taken to mean a given physical-layer medium, but which we take here to mean the transport network. From the perspective of the network, joint source/channel coding requires the allocation of network resources (buffer space, bandwidth, power, etc.) to maximize the network traffic capacity subject to a subjective quality objective. From the perspective of the

source, joint source/channel coding means processing the signal in such a way that transport network impairments have minimal subjective effect, subject to maximizing the network's traffic capacity. This suggests that the source coding must take account of how the transport network allocates resources, and the effect that has on end-to-end impairments, and conversely the transport needs to know the source coding strategy and the subjective impact of its resource allocations.

A simple example of joint source/channel coding is compression [5]. The usual goal of compression is to minimize bit rate, which is intended to maximize the traffic capacity of the network without harming the subjective quality appreciably. However, minimizing the bit rate (say in the average sense) is simplistic, because traffic capacity typically depends on more than average bit rate. To cite several examples:

- The statistical multiplexing advantage in congestion-dominated subnets depends very much on the peakiness of the offered bit streams, at least for a constant loss and delay objective, and the manner in which the bit rate varies with time is an important factor in the traffic capacity.
- A side effect of compression, at least at a relatively constant subjective quality, is usually to generate a variable bit rate, and exploiting that variable bit rate through statistical multiplexing results inevitably in packet losses, which causes subjective impairment.
- Compression normally results in an increase in the susceptibility to bit errors. On interference-dominated subnets, such as cellular radio wireless access links, it is expensive (in terms of traffic capacity) to provision consistently low error probability, since that requires large transmitted power and hence increased interference to other users. Thus, the traffic capacity of such a subnet depends strongly on the reliability requirement, as well as the bit rate, and it is not automatically the case that minimizing the bit rate is equivalent to maximizing the traffic capacity.

Even standards such as MPEG targeted at widespread use commonly make specific limiting assumptions about the transport. Generally MPEG makes the assumption that errors are infrequent enough that corrupted blocks of data can simply be discarded, and that such errors can propagate to the next intraframe-coded image, without substantially degrading subjective quality. This results in very tight error rate requirements, depending on the application [12][13]. While this is feasible in storage, fiber, and broadcast wireless applications (such as terrestrial

HDTV [5]), this is likely not feasible in multiple access wireless applications¹. (Voice standards intended for multiple access channels and mobile receivers with fading generally assume a worst-case error rate in the range of $1.0e-2$ to $1.0e-3$, which is more representative on these types of channels during deep fades [18].) There are also compatibility issues in MPEG on transport systems with delay jitter.

MPEG illustrates the difficulty in designing compression standards with sufficient flexibility and scalability for a variety of transport scenarios. For a heterogeneous transport environment, we need more agility, and that is a major goal of the architectures proposed later.

Maximizing traffic capacity subject to a subjective quality criterion results in an intricate coupling of the design of the compression and the transmission. This is unfortunate from a network architecture and complexity management perspective. We would like a maximum *decoupling* between the design of the CM service and the network transport, so that they can evolve independently.

Encryption is an important requirement for privacy and for preventing unauthorized interception in intellectual property protection schemes. Encryption techniques can be divided into two classes [16], the *binary additive stream cipher* (which does not have error multiplication and propagation effects but is susceptible to loss) and the *block cipher* (which does have error multiplication) There are two reasons to prefer performing FEC after encryption. One is the error-multiplier effect of some encryption schemes, increasing the correction burden on any FEC before encryption. Another is the redundancy introduced by an FEC before encryption, which weakens the security of the encryption.

3. SEPARATING TRANSPORT AND BEARER SERVICES

In order to allow different transmission media can work with the same source coding, and different source coders to work with different transmission media, it is necessary to separate the design of source coders from the transmission as much as possible. One group has made a proposal for an architecture for the future GII, and for consistency we draw upon their terminology [2],

1. While forward error-correcting coding may be able to achieve such error rates, countering the worst-case error rate environment during deep fades will require very high levels of redundancy, which, because it is present even during favorable channel conditions, will severely restrict the traffic capacity [17].

which is shown in Figure 1¹. Applications draw upon the *transport services layer*, which calls upon the *bearer services layer*, which carries the bits from one location to another. The transport services layer conditions the data for the bearer services (for example, the compression of audio or video) and accounts for impairments in the bearer services (for example, re-sequencing of packets, as in TCP, or retransmission of lost packets, as in TCP, or synchronization of packets to a global clock as in the MPEG-2 transport stream [7][8]).

While [2] does not attempt a detailed partitioning of functions between transport services and bearer services layers, we make a proposal here specifically with respect to signal processing functions, as shown in Figure 2. FEC has been placed in the bearer service layer, and compression and encryption in the transport services layer, where we have named the interface between these two layers the *medley gateway* for reasons delineated later. Compression is inherently a conditioning function, and hence belongs in the transport service layer. The reasons that we include encryption within the transport services layer are more subtle:

- Encryption must follow compression (and precede decompression) and hence cannot reside in the application layer.
- There may be two or more bearer service layers in a given connection in a heterogeneous environment

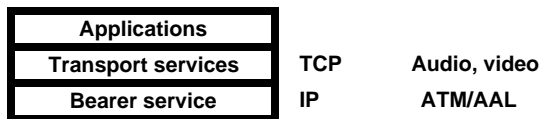


Figure 1: An architecture for the GII, including both CM and data services. Each layer may be subdivided into appropriate sub-layers.

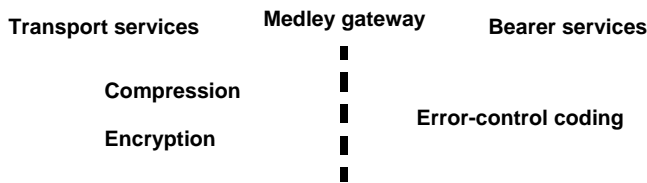


Figure 2: Partitioning of signal-processing functions between the transport services and bearer services layers.

(Section 2). Including encryption in the transport services layer opens the possibility of doing encryption on an end-to-end basis, with resulting simplified key management and higher level of security. Encryption in the bearer service layer could result in two or more encryption/decryption operations, with complications to key management and reduced security due to “in the clear” signals available at intermediate points. Also encryption would not be under the control of the user, but rather the bearer-service provider, dramatically reducing the security from the perspective of the user.

- The proposed architecture eliminates the increased burden of error multiplication due to block ciphers on the FEC algorithms.

The reasons we have placed FEC in the bearer services layer include:

- The most unreliable transmission media, wireless, are also the most critical with respect to spectral efficiency. On such media, signal-space coding techniques (for example trellis coding and multidimensional signal constellations [14]) are tightly integrated into the modulation system and hence are inherently localized to each transmission link in the connection.
- There are many error correction techniques available, such as retransmission, FEC, interleaving, etc. It is most efficient for these techniques to be tightly coupled to the transmission environment. For example, the temporal characteristics of wireless access links depend heavily on the level of mobility, and the level of interleaving (to counter error-correlation effects) and the coding techniques are best coordinated with that mobility.
- Achieving high traffic capacity on time varying media (such as wireless channels in the presence of terminal motion) requires techniques that take account of the state of the channel, so that parameters such as FEC redundancy, transmit power, etc., are varied with time. This important class of techniques is only practical to implement within the bearer service because of the close coupling to the physical layer and the need for low-latency feedback between modulation and coding and the physical layer.

There is no reason to dogmatically preclude the involvement of the transport service. For example, in “best effort” data services without delay guarantees, transport services retransmission protocols (as in TCP) may be acceptable.

1. Actually, [2] adds a fourth layer, middleware services, which we delete here because it is generally unrelated to signal processing functions.

3.1 Abstracted View of Transport and Bearer Service

To maintain flexibility and contain complexity, it is important that abstractions be defined at the medley gateway. These abstractions should retain information that is relevant and critical, while hiding unnecessary details. One of our major goals is to separate, insofar as is possible, the design of the transport service from the bearer service. Not only is this an important complexity management technique, but it also increases our ability to substitute different bearer service entities for a given transport service entity, and vice versa.

Since the bearer service performs the relatively simple function of transporting information elements (packets or cells) from one location to another, the abstracted view should focus on the fundamental impairments of corruption, loss, and delay. A basic model incorporating these three elements is shown schematically in Figure 3. Often, the transport service will be interested in the temporal properties of these impairments; that is, a characterization of whether impairments like losses, corruption, or excessive delays are likely to be bunched together, or they statistically spread out in time. In the presence of time-varying characteristics such as congestion or fading, the bearer service model will have to include a characterization of the temporal characteristics of the impairments, in order that the transport service be able to adequately distinguish between, for example, wireless access links with widely varying fading rates (related to terminal velocity). This issue is discussed further later.

Note what information is *not* included in the bearer service model. We deliberately exclude knowledge of the detailed transmission and switching structure within the bearer service. For example, we hide from the transport service any knowledge of whether loss and delay is caused by congestion, or by FEC and interleaving techniques, etc. Similarly, knowledge of whether corruption is caused by thermal noise, or interference, or is affected by time-varying mechanisms like Ricean or Rayleigh fading, is obscured. Imbedding such knowledge in the transport service layer not only creates greater complexity and dependency, it also largely rules out heterogeneous bearer-service scenarios.



Figure 3: Abstracted model of a bearer service entity.

The transport service passes on to the bearer service layer a *stream* of information elements. A description of the properties of the transport service stream is called a *flowspec*. The most relevant of these properties are:

- *Rate* parameters, such as average rate, peak rate, and a characterization of the temporal characteristics of the rate.
- *Quality of service (QOS)* parameters expected of the bearer service, including loss, corruption, and delay, and the temporal characteristics of these impairments.

A primary objective is to allow joint source/channel coding, in spite of our careful separation of the design of the two layers. To this end, we include in the transport service layer abstraction the substream structure shown in Figure 4. The stream of information elements is logically divided into *substreams*, which are visible to the bearer service entity. The QOS requirements of the substreams are different, and the substreams are also jointly specified in terms of a set of rate parameters. Joint source/channel coding becomes possible once the set of QOS requirements and rate parameters are established.

For example, the two-level priority schemes in video coding can be thought of as associating high-importance packets with one substream, and low-importance packets with another substream. The higher-importance substream would have a QOS requirement associated with a lower loss probability than the lower-importance substream. The bearer service can exploit the relaxed QOS requirement of the lower-importance substream to achieve a higher traffic capacity.

More generally, the transport service, knowing the QOS to be expected on the substreams, can associate information elements with substreams in a way that results in acceptable subjective quality. The bearer service, knowing the QOS expectations and rates, can allocate its internal resources, such as buffer capacity, power, etc., in a way that maximizes the traffic capacity. In the absence of the substream structure, the bearer service would have to provide the tightest or most expensive QOS requirements to the entire stream in order to achieve the same overall subjective quality. The distinction between a stream composed of a set of *substreams* and a *set of streams* with different QOS

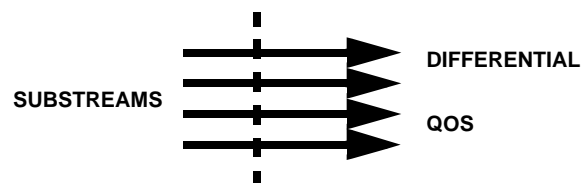


Figure 4: Abstracted view of the transport service.

requirements resides in the rate parameters. Distinct streams, which originate from distinct transport service entities, are assumed to be independent. Substreams, on the other hand, originate from the same transport service entity, and hence will typically have highly correlated rates.

The abstractions introduced in the bearer service model make opportunities in joint source/channel coding more transparent. The joint source/channel coding functionality is now divided between the transport services layer and the bearer services layer. The transport services layer, in an effort to maximize its traffic-carrying capacity, does the following:

- Affords no information element (such as a packet) a loss or corruption probability lower than required by the QOS specified for the substream with which it is associated.
- Takes maximum advantage of the delay flexibility afforded by the QOS on a per information element basis. This is a new opportunity in joint source/channel coding not anticipated in previous approaches.

Simultaneously, the transport service attempts to maximize the subjective quality afforded to the application or user within the constraints of the agreed flowspec. For example, information elements less sensitive to delay are associated with a substream with a relaxed delay specification.

The medley gateway model does impose one limitation on joint source/channel coding. It does not include a feedback mechanism by which information on the current conditions in the bearer service layer can be fed back to affect the transport services layer. We do not allow it because we question its practicality in the general situation outlined in Section 2, where the transport services layer implementation may be geographically separated from the bearer services entity in question, implying an unacceptably high delay in the feedback path.

3.2 Scalability and Configurability Issues

Requiring transport services and bearer services to be interchangeable puts a much greater burden on both. A transport service entity that can utilize any bearer service entity must exhibit scalability to deal, for example, with both a broadband backbone bearer service and a wireless access bearer service. Similarly, the bearer service must be prepared to allocate its resources differently for different rate parameters and QOS requirements, for example to provision both an audio and a video transport service.

To deal with this problem, a *negotiation* must occur during the call setup phase. In particular, we envision a scenario such as the following:

- The transport service entity, based on subjective quality criteria requested by the application, requests a flowspec of the bearer service. However, since the bearer service can conceivably be anything between a broadband backbone and a wireless access link, this request may be wildly unrealistic or too expensive.
- The bearer service entity determines the feasibility of the flowspec, and if feasible passes back to the transport service a cost¹ associated with that flowspec.
- The transport service and the bearer service negotiate as appropriate, arriving at an acceptable trade-off between subjective quality and cost. This results in a final agreed-to flowspec.
- Both the transport service and the bearer service configure themselves. This implies appropriate resource allocation by the bearer service to guarantee that the agreed flowspec will be achieved. This also implies that the transport service configures itself to conform to the rate parameters in the flowspec and configures itself to maximize subjective quality subject to the agreed flowspec.

There are a number of challenges inherent in this process that will not be discussed further here.

4. EDGE VS. LINK ARCHITECTURE FOR TRANSPORT SERVICES

In Section 3 we addressed the problem of separating the designs of the transport service from the bearer service, while leaving open most possibilities for joint source/channel coding. Our motivation was to allow the flexibility to substitute freely the transport service or bearer service realizations. In this section, we consider a related set of issues in the provision of CM services through two or more heterogeneous subnets. Many of the issues addressed in Section 3 become more critical.

Consider two basic architectures illustrated in Figure 5 for a concatenation of subnets (two subnets in this example). We partition the connection into *links*, where each link corresponds to one homogeneous bearer service subnet. For example, in wireless access to a broadband network, the wireless subnet would constitute one bearer service link, and the broadband subnet would constitute

1. In a commercial context, cost is likely in monetary terms, or in other contexts it may be expressed in other terms. In any case, an important component of the cost will be the traffic capacity implications of the requested flowspec.

the second link. The distinction between the *link architecture* and the *edge architecture* is whether or not a transport services layer is included within each subnet. The back-to-back transport services layers in the link architecture include, for CM services like audio and video, a decompression signal processing function followed by a compression function. These functions together constitute a *transcoder*, as mentioned in Section 1. They may also include a decryption function followed by encryption.

The question is, which architecture is better? We believe that the edge architecture is superior. In favor of this architecture, we mention five factors:

- *Privacy and security.* The link architecture is incapable of providing privacy by end-to-end encryption under user control, since an encrypted signal cannot be transcoded. The best that can be done is encryption on a link basis by the service provider(s), with no ability for the user to verify that encryption has been performed. As encryption is an important requirement for some users, for example as one element of intellectual property protection, this problem in our opinion should preclude serious consideration of the link architecture.
- *Open to change.* The edge architecture is open to substitution of different transport service layers at the network edge (user terminal or access point). This leads to an economically viable method to upgrade transport services over time, as well as introduces new ones.
- *Performance.* The link architecture suffers from the accumulation of delay and subjective impairment through tandem compressions and

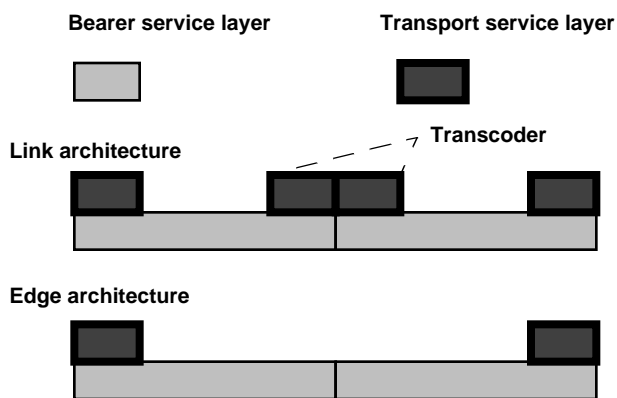


Figure 5: Contrast of link and edge architectures for concatenated heterogeneous subnets, where the former includes a transcoder function at the gateway between the two subnets.

decompressions of the CM signal. This problem has already become serious in digital cellular telephony. In more complicated heterogeneous scenarios, delay could become unacceptable for delay-sensitive interactive applications.

- *Complexity.* The edge architecture has a number of challenges as discussed later, but overall we believe it substantially reduces the complexity of establishment and configuration.
- *Mobility.* The link architecture embeds considerably more state within the network associated with the realization of a CM service, creating additional requirements for migration of state when terminals are mobile (requiring the movement or the dis-establishment/establishment of multipoint connection spanning trees).

Joint source/channel coding is important for achieving adequate traffic capacity, especially on critical wireless access links. A basic problem in achieving joint source/channel coding in the edge architecture is illustrated in Figure 6, where substreams are *not* utilized. In this case, in order to be able to perform joint source/channel coding, the bearer service link must know the full details of the syntax of the transport service. For example, it can then give different bits appropriately different error rates. However, this introduces two problems:

- Knowledge of the transport service syntax within all bearer service links has many of the problems of the link architecture, including complexity and inflexibility.
- Encryption hides the transport service syntax, destroys the relationship between bearer service QOS and the QOS of the decrypted bit stream, and thus effectively precludes joint source/channel coding.

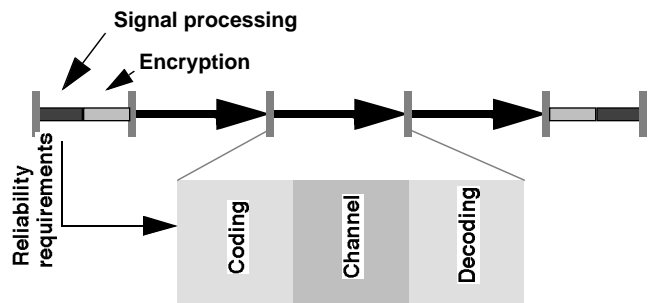


Figure 6: The reliability requirements for the overall bearer service must be somehow accommodated at an intermediate link.

The substream solves these problems. As shown in Figure 7, each bearer service link is obligated to maintain the structure of the medley gateway at its output. That is, the medley gateway is the interface between transport service and bearer service layers, and *also* the interface between distinct bearer service entities. This is why we call it a *gateway*, since it serves as a common protocol interface between heterogeneous bearer service subnets. The substream structure is visible to each bearer service link, which is able to allocate resources efficiently in accordance with joint source/channel coding.

Encryption need not interfere with joint source/channel coding, as illustrated in Figure 8. If encryption is performed independently on each substream (there is no dependency among the states of the distinct encryption units) there is a one-to-one correspondence

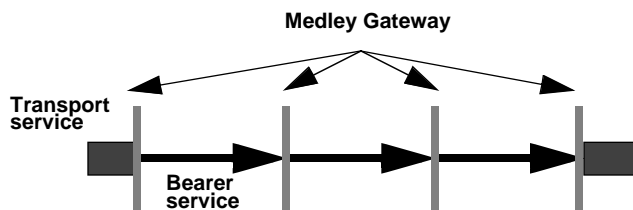


Figure 7: Each bearer service link maintains the structural integrity of the medley gateway, making the structure available to downstream bearer service links.

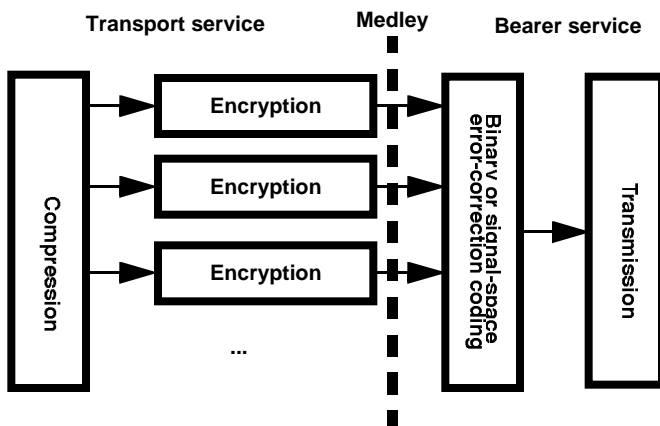


Figure 8: A proposed architecture including compression, encryption, and error-correction encoding. Compression and encryption are performed on an end-to-end basis, error correction coding is associated with each transport link. The medley gateway is reproduced at the input and output of each transport link.

between QOS delivered to each substream in the bearer service and the QOS experienced by the decrypted substream¹.

Both the link and the edge architectures raise important issues in resource allocation in session establishment. In both cases, for CM services the overriding objective is to obtain acceptable and controllable subjective quality in the audio or video service. Subjective quality is measured objectively by parameters such as frame rate and resolution (for video), bandwidth (for audio), and delay (for both video and audio). It is also measured by other factors more difficult to characterize, such as the perceptual impact of artifacts introduced by information discarded in the transport service (in the compression) and in the bearer service (packet losses), and also artifacts introduced by corruption in the bearer service.

In the link architecture, overall subjective quality objectives must be referenced back to the individual links, since each link will contribute subjective impairments and those impairments accumulate across links. It is relatively straightforward to partition objective impairments like delay among the links. Other objective parameters like frame rate, bandwidth, and resolution will be dictated by the worst-case link, and are thus also straightforward to characterize. Subjective impairments due to loss and corruption mechanisms will, however, be very difficult if not impossible to characterize in a heterogeneous bearer service environment. Simple objective measures like mean square error are fairly meaningless in the face of complex impairments like the masking of bearer service losses. Thus, as a practical matter it will be very difficult to predict and control end-to-end subjective quality.

The situation in the edge architecture is quite different. The first step is to generate an aggregated bearer service model for all the concatenated bearer service links. That is, the loss models for the individual links must be referenced to a loss model for the overall connection, and similarly for corruption and delay. There are no doubt serious complications in this aggregation, like for example correlations of loss mechanisms in successive links due to common traffic. Nevertheless, this is a relatively straightforward task susceptible to analytical modeling. Once this is done, the aggregate bearer service model must be related back to transport service subjective quality measure, much in the fashion of a *single* link in the link architecture. There is no need

1. Encryption may affect the QOS (through error multiplication effects) and must be taken into account in establishing the bearer service QOS.

for characterizing the overall subjective impairment in concatenated transport service entities.

Overall the prediction and control of subjective quality in the edge architecture is much simpler than in the link architecture, and this is an additional advantage. In practice it should be possible to maintain better control and prediction of end-to-end subjective quality.

The link architecture suffers from the accumulation of delay due to multiple compression/decompression steps, which as illustrated by digital cellular is often quite significant. In addition, the link architecture suffers from an accumulation of impairment due to information discarded in each compression step, such as by quantizers. For an overall subjective quality objective, each link in a multiple-link connection will have to meet a more stringent requirement to take into account this accumulation of impairments. Each link will thus have a lower traffic capacity than in the edge architecture.

4.1 Multipoint Connections

The problem of multipoint connections is illustrated in Figure 9. With heterogeneous receiving terminals, or heterogeneous subnets, we may need different representations (say with different bandwidth or resolution) of the CM service after a splitting point, but to conserve bearer service resources we want to share a common stream before the splitting point. An obstacle to this is encryption, which will hide the syntax of the originating stream. One solution is to locate transcoding at the splitting point, preceded by decryption and followed by encryption, but this introduces all the disadvantages of the link architecture. The medley gateway provides a framework for the solution to this problem as shown in Figure 10. At the point where two representations are split, a (not necessarily proper) subset of the medley substreams is extracted for each downstream branch, with the great simplification that the splitting function can be accomplished entirely within the bearer service layer.

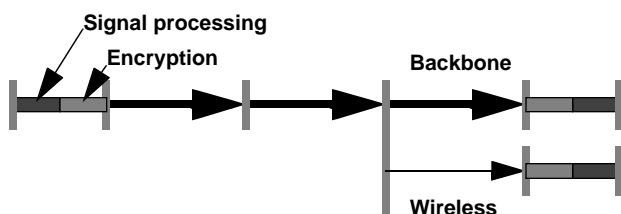


Figure 9: Illustration of a multipoint connection with heterogeneous receiving terminals.

If each substream is independently encrypted, encryption does not interfere with this splitting function.

In the edge architecture, the substream structure is used for three distinct but complementary purposes:

- *Joint source/channel coding.* It allows the transport service to present to each bearer service entity, in a generic fashion separated from particular transport service standards, the differing QOS requirements of different information elements, thus allowing the bearer service to efficiently allocate its resources.
- *Layered coding.* It allows the transport service to decompose its layered encoding in a way that is also generic and visible to the bearer service layer, so that the splitting function required in multipoint connections with heterogeneous terminals can be performed entirely within the bearer service layer.
- *Privacy and security.* Independent encryption of the substreams allows the privacy and security of end-to-end encryption without interfering with either joint source/channel coding or multipoint splitting.

5. CONCLUSIONS

There is much research to be performed for medley transport services and medley bearer services, and this paper has presented a general framework. Examples of the numerous issues that are raised by the medley gateway include:

- The design of medley transport services that take advantage of the enhanced bearer services at the medley gateway (such as delay/loss trade-offs) and which have the needed level of scalability and configurability needed for future open networks.
- The design of medley bearer service subnets that maintain the structural integrity of the medley gateway, which have the ability to configure to different impairment profiles for different substreams, and which exploit the substream structure for higher traffic capacity.

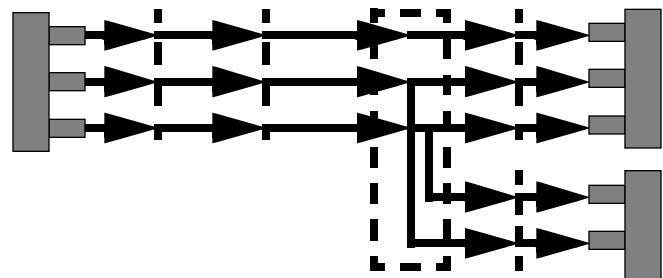


Figure 10: Illustration of a multipoint splitting function entirely within the bearer service layer.

- An understanding of joint source/channel coding, as constrained by the structure of the medley gateway. Similarly, an understanding of the design of hierarchical compression algorithms for multipoint heterogeneous terminals, as constrained by the structure of the medley gateway. In both cases, possible enhancement of the gateway based on understanding of the transport service design.
- Understanding of issues inherent in the aggregation of concatenated bearer service links for CM services.
- Development of negotiation strategies for resolving transport service subjective quality vs. bearer service QOS and cost.
- The upgrade of signalling systems to provide the needed capabilities in support of the edge architecture, including aggregation of bearer service links and negotiation between the endpoint terminals and the aggregated links.

6. References

1. G.M. Parulkar and J.S. Turner, "Towards a framework for high-speed communication in a heterogeneous networking environment", *IEEE Network*, March 1990, p. 19.
2. National Research Council, Computer Science and Telecommunications Board, *Realizing the Information Future; The Internet and Beyond*. Washington D.C.: National Academies Press, 1994.
3. L. Zhang, S. Deering, D. Estrin, S. Shenker, D. Zappala, "'RSVP: a new resource reservation protocol", *IEEE Network*, Sept. 1993, p. 8.
4. Ferrari, D., "Real-time communication in an internetwork" *Journal of High Speed Networks*, 1992, vol.1, (no.1):79-103.
5. Anastassiou, D., "Digital television", *Proceedings of the IEEE*, April 1994, vol.82, (no.4):510-19.
6. Le Gall, D., "MPEG: a video compression standard for multimedia applications", *Communications of the ACM*, April 1991, vol.34, (no.4):46-58.
7. MacInnis, A.G., "The MPEG systems coding specification", *Signal Processing: Image Communication*, April 1992, vol.4, (no.2):153-9.
8. Holborrow, C., "MPEG-2 Systems: a standard packet multiplex format for cable digital services", *Proc. 1994 Conference on Emerging Technologies, Society of Cable Television Engineers*, Phoenix, Ariz., Jan. 1994.
9. Kawashima, M.; Cheng-Tie Chen; Fure-Ching Jeng; Singhal, S., "Adaptation of the MPEG video-coding algorithm to network applications" *IEEE Transactions on Circuits and Systems for Video Technology*, Aug. 1993, vol.3, (no.4):261-9.
10. Pancha, P.; El Zarki, M., "MPEG coding for variable bit rate video transmission", *IEEE Communications Magazine*, May 1994, vol.32, (no.5):54-66.
11. Zhu, Q.-F.; Wang, Y.; Shaw, L., "Coding and cell-loss recovery in DCT-based packet video", *IEEE Transactions on Circuits and Systems for Video Technology*, June 1993, vol.3, (no.3):248-58.
12. S-M Lei, "Forward Error Correction Codes for MPEG2 over ATM", *IEEE Transactions on Circuits and Systems for Video Technology*, April 1994.
13. Montreuil, L., "Performance of coded QPSK modulation for the delivery of MPEG-2 stream compared to analog FM modulation", *Proc. National Telesystems Conference, 1993*.
14. E.A. Lee and D.G. Messerschmitt, *Digital Communication*, Second Edition, Boston: Kluwer Academic Press, 1993.
15. W. Ford and B. O'Higgins, "Public-key cryptography and open systems interconnection", *IEEE Communications Magazine*, July 1992, p. 30.
16. Massey, J., "An introduction to contemporary cryptology", *Proceedings of the IEEE*, Special Section on Cryptology, May 1988, p. 533.
17. T. Meng, "Portable video-on-demand in wireless communications", *Proceedings of IEEE*, Dec. 1994.
18. Natvig, J.E.; Hansen, S.; de Brito, J., "Speech processing in the pan-European digital mobile radio system", *Proc. GLOBECOM*, Dallas, TX, USA, 27-30 Nov. 1989.
19. Lao, A., Reason, J., and Messerschmitt, D.G., "Layered asynchronous video for wireless services", *Proc. IEEE Mobile Computing Conference*, 1994, to appear.
20. Yun, L., Messerschmitt, D.G., "Power Control and Coding for Variable QOS on a CDMA Channel", *Proc. IEEE Military Communications Conference*, Oct. 1994.