By Les Simmonds

MPEG - The Standards and History

There has been a lot written about MPEG, some accurate, and some not accurate, so we thought it might be a good idea to present the history of MPEG using a Tektronix paper as the base of this article. Tektronix makes the test equipment necessary to test the various MPEG standards and we reckon they should be an accurate place to start. This article examines the history and structure of MPEG and the evolution of the various MPEG standards.

What is MPEG

MPEG is the Moving Pictures Experts Group, a committee that comes under the joint control of the International Standards Organisation (ISO) and the International Electro-technical Commission (IEC). IEC handles international standardisation for electrical and electronic technologies; ISO handles virtually everything else. At the start of the information technology age ISO and IEC formed a joint technical committee (JTC1) to address IT issues. JTC1 has a number of working groups including JPEG (Joint Photographic Experts Group) and WG11, which is MPEG. The committee was formed in 1988 under the leadership of MPEG convener Dr. Leonardo Chiariglione of Italy. Attendance at MPEG meetings, normally held four times each year, has grown from about 15 delegates in 1988 to some 300 in 2003. It established an enviable track record of generating standards that achieve widespread adoption, MPEG-1, MPEG-2, and the MP3 audio compression standard (MPEG-1 Audio, layer 3). This reputation was somewhat tarnished by MPEG-4, not because of deficiencies in the standard, but as a result of the long delay in publishing licensing terms and the strong adverse reaction to the first terms that were eventually published in early 2002.

It should be noted that MPEG itself has no role in licensing. As a committee under ISO and IEC, it requires that technology included in its standards be licensable under "reasonable and non-discriminatory terms," but there is no accepted definition of "reasonable." Licensing is the responsibility of the holders of the relevant patents and typically this means many organisations throughout the world that have contributed to research and development and wish to see some recompense.

For MPEG-2, the patent holders grouped together and formed MPEG-LA (MPEG licensing authority). All the essential patents are certified by this group and are licensed as a block to any organisation wishing to implement the standards. These worked well for MPEG-2 but, as noted above, greater difficulties are being experienced with MPEG-4 and many hold the delays in publishing licensing terms responsible for the current lack of commercial success of MPEG-4. (This, of course, may change. The MPEG-4 Industry Forum is working hard to find solutions acceptable to patent holders and potential users and revised proposals released in mid-2002 are likely to be more readily accepted.)

MPEG-1

The MPEG-1 system, ISO/IEC 11172, is the first international compression standard for motion imagery and was developed between 1988 and 1992. It uses DCT transforms, coefficient quantisation, and variable length coding in a similar manner to JPEG, but also includes motion compensation for temporal compression.

It is in three parts:
ISO/IEC 11172-1, the multiplex structure.
ISO/IEC 11172-2, video coding.
ISO/IEC 11172-3, audio coding.

In its day MPEG-1 represented a remarkable technical achievement. It was designed to compress image streams with SIF picture size, 352x288 (25fps PAL) or 352x240 (30fps NTSC), and associated audio, to approximately 1.5 Mbps total compressed data rate. This rate is suitable for transport over T1 data circuits and for replay from CD-ROM, and corresponds approximately to the resolution of a consumer video recorder. A measure of this achievement may be seen by comparing the numbers for an audio CD. A normal audio CD, carrying two-channel audio, at 16-bit resolution with a sampling rate of 44.1 kHz, has a data transfer rate of up to 1.5 Mbps. MPEG-1 succeeds in compressing video and audio so that both may be transmitted within the same data rate!

The CIF format is a compromise between European and American SIF (source input format) formats: spatial resolution for 625 SIF (352x288) and temporal resolution 525 SIF (29.97 fps). This is the basis for video conferencing. MPEG-1 was designed for CIF images and has no tools to handle interlaced images, so it had little obvious impact in the world of broadcast television.

Before leaving MPEG-1, it is important to note what is actually included in the standard and how interoperability is achieved. The standard defines a tool set, the syntax of the bit stream, and the operation of the decoder. It does not define the operation of the encoder – any device that produces a syntactically valid bit stream that can be decoded by a compliant decoder is a valid MPEG encoder. Also, it does not define the quality of the picture, or encoding quality. This allows for the evolution of encoding technology without change to the standard, and without rendering existing decoders obsolete. This model is used throughout the MPEG standards. The success of this strategy is obvious; although MPEG-2 is used for video, MPEG-1, layer 2 audio is still in use as the principal audio compression system in the DVB transmission systems today.

MPEG-2

MPEG-1 was frozen (i.e., subsequent changes were allowed to be editorial only) in 1991. In the same year the MPEG-2 process was started, and MPEG-2 eventu-

ally became a standard in 1994. The initial goals were simple; there was a need for a standard that would accommodate broadcast quality video width. This required the coding of "full size" standard definition images (704x480 at 29.97 fps, and 704x576 at 25 fps), and the ability to code interlaced video efficiently. In many ways, MPEG-2 represents the "coming of age" of MPEG. The greater flexibility of MPEG-2, combined with the increased availability of large-scale integrated circuits, meant that MPEG-2 could be used in a vast number of applications. The success of MPEG-2 is best highlighted by the demise of MPEG-3, intended for high-definition television. MPEG-3 was soon abandoned when it became clear that MPEG-2 could accommodate this application with ease. MPEG-2 is, of course, the basis for both the ATSC and DVB broadcast standards, and the compression system used by DVD. MPEG-2 was also permitted to be a moving target. By the use of profiles and levels, discussed below, it was possible to complete the standard for one application, but then to move on to accommodate more demanding applications in an evolutionary manner. Work on extending MPEG-2 continues.

MPEG-2 is documented as ISO/IEC 13818, currently in 10 parts. The most important parts of this standard are:

ISO/IEC 13818-1 Systems (transport and programs streams), PES, T-STD buffer model and the basic PSI tables: CAT, PAT, PMT and NIT.
ISO/IEC 13818-2 video coding.
ISO/IEC 13818-3 audio coding.
ISO/IEC 13818-4 MPEG test and conformance.
ISO/IEC 13818-6 data broadcast and DSMCC.

One of the major achievements of MPEG-2 defined in 13818-1, the transport stream, is described in Section 8. The flexibility and robustness of this design have permitted it to be used for many applications, including transport of MPEG-4 and MPEG-7 data.

Note: DVB and ATSC transport streams carry video and audio PES within "program" groupings, which are entirely different than "program streams" (these are used on DVD & CD). MPEG Transport Streams are normally constant bit rate but program streams are normally variable bit rate.

Profiles and Levels in MPEG-2

With certain minor exceptions, MPEG-1 was designed for one task; the coding of fixed size pictures and associated audio to a known bit rate of 1.5 Mbps. The MPEG-1 tools and syntax can and have been used for other purposes, but such use is outside the standard, and requires proprietary encoders and decoders. There is only one type of decoder compliant to the MPEG-1 standard.

At the outset, there was a similar goal for MPEG-2. It was intended for coding of broadcast pictures and sound, nominally the 525/60 and 625/50 interlaced television systems. However, as the design work progressed, it was apparent that the tools being developed were capable of handling many picture sizes and a wide range of bit rates. In addition, more complex tools were developed for scalable coding systems. This meant that in practice there could not be a single MPEG-2 decoder. If a compliant decoder had to be capable of handling high-speed bit streams encoded using all possible tools, it would no longer be an economical decoder for mainstream applications. As a simple example, a device capable of decoding high-definition signals at, say, 20 Mbps would be substantially more expensive than one limited to standard-definition signals at around 5 Mbps. It would be a poor standard that required the use of an expensive device for the simple application.

MPEG devised a two-dimensional structure of profiles and levels for classifying bit streams and decoders. Profiles define the tools that may be used. For example, bidirectional encoding (B-frames) may be used in the main profile, but not in simple profile. Levels relate just to scale. A high level decoder must be capable of receiving a faster bit stream, and must have more decoder buffer and larger frame stores than a main level decoder. However, main profile at high level (MP@HL) and main profile at main level (MP@ML) use exactly the same encoding/decoding tools and syntax elements.

The simple profile does not support bidirectional coding, and so only and P-pictures will be output. This reduces the coding and decoding delay and allows simpler hardware. The simple profile has only been defined at main level.

The Main Profile is designed for a large proportion of uses. The low level uses a low-resolution input having only 352 pixels per line. The majority of broadcast applications will require the MP@ML subset of MPEG, which supports SDTV (standard definition TV).
The high-1440 level is a high definition scheme that doubles the definition compared to the main level. The high level not only doubles the resolution but maintains that resolution with 16:9 format by increasing the number of horizontal samples from 1440 to 1920.
In compression systems using spatial transforms and requantising, it is possible to produce scalable signals. A scalable process is one in which the input results in a main signal and a "helper" signal. The main signal can be decoded alone to give a picture of a certain quality, but if the information from the helper signal is added, some aspect of the quality can be improved.

For example, a conventional MPEG coder, by heavily requantising coefficients, encodes a picture with moderate signal-to-noise ratio results. If, however, that picture is locally decoded and subtracted pixel-by-pixel from the original, a quantizing noise picture results. This picture can be compressed and transmitted as the helper signal. A simple decoder only decodes the main, noisy bit stream, but a more complex decoder can decode both bit streams and combine them to produce a low-noise picture. This is the principle of SNR (signal-to-noise ratio) scalability.

As an alternative, coding only the lower spatial frequencies in a HDTV picture can produce a main bit stream that an SDTV receiver can decode. If the lower definition picture is locally decoded and subtracted from the original picture, a definition-enhancing picture would result. This picture can be coded into a helper signal. A suitable decoder could combine the main and helper signals to recreate the HDTV picture. This is the principle of spatial scalability. The high profile supports SNR and

spatial scalability as well as allowing the option of 4:2:2 sampling. The 4:2:2 profile has been developed for improved compatibility with digital production equipment. This profile allows 4:2:2 operation without requiring the additional complexity of using the high profile. For example, a HP@ML decoder must support SNR scalability, which is not a requirement for production.

The 4:2:2 profile has the same freedom of GOP structure as other profiles, but in practice it is commonly used with short GOPs making editing easier. 4:2:2 operation requires a higher bit rate than 4:2:0, and the use of short GOPs require an even higher bit rate for a given quality. The concept of profiles and levels is another development of MPEG-2 that has proved to be robust and extensible; MPEG-4 uses a much more complex array of profiles and levels, to be discussed later with MPEG-7 and MPEG-21.

We continue with MPEG standards and history and will cover MPEG-4, MPEG-7 and MPEG-21. This history of MPEG uses a Tektronix paper as its base and examines the history and structure of MPEG and the evolution of the various MPEG standards.

MPEG-4

International standardisation is a slow process, and technological advances often occur which could be incorporated into a developing standard. Often this is desirable, but continual improvement can mean that the standard never becomes final and usable. To ensure that a standard is eventually achieved there are strict rules that prohibit substantive change after a certain point in the standardisation process. So, by the time a standard is officially adopted there is often a backlog of desired enhancements and extensions. So it was with MPEG-2. As discussed above, MPEG-3 had been started and abandoned, so the next project became MPEG-4.

Two versions of MPEG-4 are already complete and work is continuing on further extensions.
At first the main focus of MPEG-4 was the encoding of video and audio at very low rates. In fact, the standard was explicitly optimised for three bit rate ranges:
Below 64 kbits/s.
64 to 384 kbits/s.
384 kbits/s to 4 Mbits/s.

Performance at low bit rates remained a major objective and some very creative ideas contributed to this end. Great attention was also paid to error resilience, making MPEG-4 very suitable for use in the error-prone environments, such as transmission to personal handheld devices. However, other profiles and levels use bit rates up to 38.4 Mbits/s, and work is still proceeding on studio-quality profiles and levels using data rates up to 1.2 Gbits/s. More importantly, MPEG-4 became vastly more than just another compression system – it evolved into a totally new concept of multimedia encoding with powerful tools for interactivity and a vast range of applications. Even the official "overview" of this standard spans 67 pages, so only a brief introduction to the system is possible here.

MPEG-4 Standards Documents

The principal parts of the MPEG-4 standards are:
ISO/IEC 14496-1 Systems.
ISO/IEC 14496-2 Visual.
ISO/IEC 14496-3 Audio.
ISO/IEC 14496-4 Conformance Testing.
ISO/IEC 14496-6 Delivery Multimedia Integration Framework (DMIF).

Object Coding

The most significant departure from conventional transmission systems is the concept of objects. Different parts of the final scene can be coded and transmitted separately as video objects and audio objects to be brought together, or composited, by the decoder. Different object types may each be coded with the tools most appropriate to the job. The objects may be generated independently, or a scene may be analysed to separate, for example, foreground and background objects. In one interesting demonstration, video coverage of a soccer game was processed to separate the ball from the rest of the scene. The background (the scene without the ball) was transmitted as a "teaser" to attract a pay-per-view audience.
Anyone could see the players and the field, but only those who paid could see the ball!

The object-oriented approach leads to three key characteristics of MPEG-4 streams:

☐ Multiple objects may be encoded using different techniques, and composited at the decoder.
☐ Objects may be of natural origin, such as scenes from a camera, or synthetic, such as text.
☐ Instructions in the bit stream, and/or user choice, may enable several different presentations from the same bit stream.

Video and Audio Coding

Many of the video coding tools in MPEG-4 are similar to those of MPEG-2, but enhanced by better use of predictive coding and more efficient entropy coding. However, the application of the tools may differ significantly from earlier standards. MPEG-4 codes video objects. In the simplest model a video is coded in much the same way as in MPEG-2, but it is described as a single video object with a rectangular shape. The representation of the image is known as texture coding. Where there is more than one video object, some may have irregular shapes, and generally all will be smaller than a full-screen background object. This means that only the active area of the object need be coded, but the shape and position must also be represented. The standard includes tools for shape coding of rectangular and irregular objects, in either binary or grey-scale representations (similar to an alpha channel).

Similarly, MPEG-4 uses tools similar to those of MPEG-1 and MPEG-2 for coding live audio, and AAC offers greater efficiency. Multiple audio "objects" may be encoded separately and composited at the decoder. As with video, audio objects may be natural or synthetic.

Scalability

In the context of media compression, scalability means the ability to distribute content at more than one quality level within the same bit stream. MPEG-2 and MPEG-4 both provide scalable profiles using a conventional model; the encoder generates a base-layer and one or more enhancement layers. The enhancement layer(s) may be discarded for transmission or decoding if insufficient resources are available. This ap-

proach works, but all decisions about quality levels have to be made at the time of encoding, and in practice the number of enhancement layers is severely limited (usually to one).

Later versions of MPEG-4 include the fine grain scalability (FGS) profile. This technique generates a single bit stream representing the highest quality level, but that allows for lower quality versions to be extracted downstream. FGS uses bit-plane encoding. The quantized coefficients are "sliced" one bit at a time, starting with the most significant bit. This provides a coarse representation of the largest (and most significant) coefficient(s). Subsequent slices provide more accurate representations of these most-significant coefficients, and coarse approximations of the next most significant – and so on.

Spatial scaling, including FGS, may be combined with temporal scaling that permits the transmission and/or decoding of lower frame rates when resources are limited. As mentioned above, objects may be scaled differently; it may be appropriate to retain full temporal resolution for an important foreground object, but to update to the background as a lower rate.

Other Aspects of MPEG-4

MPEG-4 is enormous, and the comments above just touch on a few of the many aspects of the standard. There are studio profiles for high-quality encoding which, in conjunction with object coding, will permit structured storage of all the separate elements of a video composite. Further extensions of MPEG-4 may even provide quality levels suitable for digital cinema.

Some of the object types defined within MPEG-4 are interesting. One example is a sprite. A sprite is a static background object, generally larger than the viewing port or display device. For example, the action of a video game may take place in front of a background scene. If a sprite is used, a large static background may be transmitted once, and as the game action proceeds the appropriate part of the background will be seen, according to the motion of the view port.

MPEG-4 defines both facial and body animation profiles. In each case a default face or body may be used, and commands sent to animate this object. Alternatively, the default object may be modified by the bit stream; for example, a specific face may be transmitted and then animated. Sophisticated animation commands related to language will permit a stored face to "read" text in many languages. Some describe MPEG-4 as the standard for video games, and certainly many of the constructs are ideally suited to that industry. However, even a cursory examination of the standard reveals such a wealth of capabilities, and such depth in every aspect, that the potential applications are endless.

The Future of MPEG-4

As discussed above, MPEG-4 is a wide-ranging set of standards with a rich offering of capabilities for many applications. This is the theory; in practice, MPEG-4 can show few successes. In particular, many observers expected that MPEG-4 would quickly become the dominant coding mechanism for audio-visual material transmitted over the Internet and replace the various proprietary codec's in use today. This has not happened, nor is there any likelihood that it will happen in the near future. There are two main reasons for this failure.

The first is technology, and the resulting performance. MPEG-4 uses video compression technology based on the ITU-developed H.26x, dating from the early 1990s. The distribution of audio and video over the Internet is a fiercely competitive business, and all the major players, Apple, Microsoft and Real Networks, have implemented proprietary video encoding schemes that outperform the current MPEG-4 codec.

The other reason for the failure (to date) of MPEG-4 is the patent licensing situation. Until early 2002, companies wishing to implement MPEG-4 did not know what royalties they would need to pay to the patent holders. The proposed licensing scheme for the basic levels of MPEG-4 has now been published, and met with strong adverse reaction from the industry. Licensing terms for the more sophisticated levels are still unknown. Certainly the initial offering of licensing terms has done nothing to increase mainstream implementation of the standard.

There is hope for the future. A joint effort of ITU and MPEG, known as the joint video team (JVT) is working on a codec known as H.26L. To quote the ITU, "The H.26L design is a block-based motion-compensated hybrid transform coder – similar in spirit but different in many specifics relative to prior designs. … H.26L significantly increases the number of available block sizes and the number of available reference pictures for performing motion estimation." The new codec also offers much greater precision in motion estimation (1/8 pixel in some implementations), and is based on a principal block size of 4x4, rather than the 8x8 used in most MPEG systems. H.26L is expected to show substantial improvements in coding efficiency, and it is the goal of the participants that the base level, suitable for Internet streaming, will be royalty-free. The first stage of the work of JVT is expected to be published as MPEG-4 Part 10.

MPEG-7

Because MPEG-3 was cancelled, the sequence of actual standards was MPEG-1, MPEG-2, and MPEG-4. Some committee participants wanted the next standard to be MPEG-5; others were attracted by the binary nature of the sequence and preferred MPEG-8. Finally, it was concluded that any simple sequence would fail to signal the fundamental difference from the work of MPEG-1 through MPEG-4, and MPEG-7 was chosen. MPEG-7 is not about compression; it is about metadata, also known as the "bits about the bits." Metadata is digital information that describes the content of other digital data. In modern parlance, the program material or content, the actual image, video, audio or data objects that convey the information are known as data essence. The metadata tells the world all it needs to know about what is in the essence.

Anyone who has been involved with the storage of information, be it videotapes, books, music, whatever, knows the importance and the difficulty of accurate cataloguing and indexing. Stored information is useful only if its existence is known, and if it can be retrieved in a timely manner when needed. This problem has always been with us, and is ad-

dressed in the analogue domain by a combination of labels, catalogues, card indexes, etc. More recently, the computer industry has given us efficient, cost-effective, relational databases that permit powerful search engines to access stored information in remarkable ways. Provided, that is, the information is present in a form that the search engine can use.

Here is the real problem. The world is generating new media content at an enormous and ever-increasing rate. With the increasing quantity and decreasing cost of digital storage media, more and more of this content can be stored. Local and wide-area networks can make the content accessible and deliverable if it can be found. The search engines can find what we want, and the databases can be linked to the material itself, but we need to get all the necessary indexing information into the database in a form suitable for the search engine.

We might guess from knowledge of earlier standards that the MPEG committee would not concern itself unduly with mechanisms for generating data. MPEG rightly takes the view that if it creates a standardized structure, and if there is a market need, the technological gaps will be filled. In previous MPEG standards, the syntax and the decoder were specified by the standard.

In MPEG-7, only the syntax is standardized. The generation of the metadata is unspecified, as are the applications that may use it. MPEG-7 specifies how metadata should be expressed. This means that the fields that should go into a database are specified, and anyone designing a search engine knows what descriptive elements may be present, and how they will be encoded.

MPEG-7 defines a structure of descriptors and description schemes that can characterise almost anything. In theory at least, primitive elements such as colour histograms and shapes can be combined to represent complex entities such as individual faces. It may be possible to index material automatically such that the database can be searched for scenes that show, for example, President Bush and U.S. Federal Reserve Chairman Greenspan together. The constructs are not confined to images. It should be possible to use a voice sample to search for recordings by, or images of, Pavarotti, or to play a few notes on a keyboard to find matching or similar melodies.

The rapid advance of storage and networking systems will enable access to vast quantities of digital content. As technology advances to satisfy the needs of MPEG-7, we will be able to index and retrieve items in ways unimaginable a few years ago. We will then need a system to control access, privacy, and commercial transactions associated with this content. This is where MPEG-21 is targeted.

MPEG-21

MPEG-21 again differs in kind from the earlier work of the committee. The basic concept is fairly simple – though wide reaching. MPEG-21 seeks to create a complete structure for the management and use of digital assets, including all the infrastructure support for the commercial transactions and rights management that must accompany this structure. The vision statement is "to enable transparent and augmented use of multimedia resources across a wide range of networks and devices."

The scope of the MPEG-21 work is indicated by the seven architectural elements defined in the draft technical report.

1. The digital item declaration is expected to "establish a uniform and flexible abstraction and interoperable schema for defining digital items." The scheme must be open and extensible to any and all media resource types and description schemes, and must support a hierarchical structure that is easy to search and navigate.

2. The digital item representation of MPEG-21 is the technology that will be used to code the content and to provide all the mechanisms needed to synchronise all the elements of the content. It is expected that this layer will reference at least MPEG-4.

3. Digital item identification & description will provide the framework for the identification and description of digital items (linking all content elements). This will likely include the description schemes of MPEG-7, but must also include "[a] new generation of identification systems to support effective, accurate and automated event management and reporting (license transactions, usage rules, monitoring and tracking, etc)." It must satisfy the needs of all classes of MPEG-21 users.

4. Content management and usage must define interfaces and protocols for storage and management of MPEG-21 digital items and descriptions. It must support archiving and cataloguing of content while preserving usage rights, and the ability to track changes to items and descriptions. This element of MPEG-21 will likely also support a form of "trading" where consumers can exchange personal information for the right to access content, and formalisation of mechanisms for "personal channels" and similar constructs.

5. Intellectual property management and protection is an essential component. The current controversies surrounding MP3 audio files demonstrate the need for new copyright mechanisms cognisant of the digital world. It can be argued that content has no value unless it is protected. MPEG-21 will build on the ongoing work in MPEG-4 and MPEG-7, but will need extensions to accommodate new types of digital items, and new delivery mechanisms.

6. MPEG-21 terminals and networks will address delivery of items over a wide range of networks, and the ability to render the content on a wide range of terminals. Conceptually a movie should be deliverable in full digital-cinema quality to a movie theatre, or at a lower quality over a slower network to a consumer device (at a different price). In either case there will be some restriction on the type and or number of uses. The user should not need to be aware of any issues or complexities associated with delivery or rendering.

7. Finally, there is a need for event reporting to "standardise metrics and interfaces for performance of all reportable events." The most obvious example here is that if the system allows a user access to a protected item, it must also ensure that the appropriate payment is made!

Acknowledgment: Tektronix Inc.

Les Simmonds is an independent CCTV consultant.

Email:
les@cctvconsultants.com.au

Web:
www.cctvconsultants.com.au

This article was originally published in Security Electronics and Networks Magazine Australia.