NTNU
Department of telematics

LARMemo-04N10
V 1.6. 15dec04
L A Rønningen

Distributed Multimedia Plays

Virtual Dinner

Contents:

1 Scope and background 3
1.2 Distributed Multimedia Plays (DMP) 3
1.3 Distributed Multimedia Plays and QoS requirements 7
1.4 Some established media systems 8
1.4.1 Conferencing 8
1.4.2 High definition TV 9
1.3.3 Streaming in Internet 9
1.5 Scene characteristics 9
1.5.1 MPEG-4 scene and stream description 10
1.5.2 Sub-objects and 2D interlace 12
1.5.3 Service description as part of scene description 13
1.6 3D multiview 13
1.6.1 Display techniques 13
1.6.2 Video segmentation and multiview shooting 17
1.7 Compression, coding 19
1.8 Resolution & Traffic Control 20
1.8.1 The RL Resolution & Traffic Control Algorithm 20
1.8.2 Other traffic control algorithms 21
1.9 Service setup and address allocation 22
2 The Virtual Dinner Scenario 24
2.1 The user's quality requirements ten years ahead 23
2.2 The Virtual Dinner 24
3 Traffic generated by advanced scenes 27
4 Core networks 28
4.1 A possible European fiber packet network Architecture 28
4.2 RTP-UDP-IPv6 applied for adaptable resolution & traffic control 29
4.3 Addressing 30
4.4 The RL-C Resolution & Traffic Control Algorithm 31
4.5 Example network 33
5 Access networks 35
5.1 Fiber to the home – FTTH 35
5.2 Extended IEEE1394 access to packet networks 35
6 Packet switch architectures 36
6.1 Short review of packet switch state-of-the-art 36
6.2 Guaranteed and-to-end time delay and acceptable resolution variation 37
6.3 A packet switch implementing the RL-C algorithm 37
7 Simulations 40
8 Testing time-varying resolution of scenes 41

References 42

1. Scope and background

Forecasting technological and market development 3-5 years ahead is feasible. But the forecasting approach aiming 10-20 years ahead probably has to be a mixture of classical technological research, feasibility studies, science fiction and arts.

The intention of this study is to present some futuristic thoughts on how DMP, Distributed Multimedia Plays, may be provided 10 -20 years from now. In the report, a large number of topics needed to design a total system are included. Some of the topics are merely touched, and in-depth descriptions can be found in the references. Some topics, such as the RL resolution & traffic control algorithm are treated in some detail.

This report is not meant to be a text-book. The intention is to present an outline background so the reader can get an understanding of the system components needed in a high-QoS DMP system. This understanding can then be used as bases for long-term and basic research on the elements necessary for DMP.

It is a fact that we have not mastered (or have not tried hard enough) time delays in Internet, and due to that, at lot of (unnecessary) synchronization and buffering mechanisms had to be introduced. Other networks, for example those transporting TV signals have so far been basically synchronous, and jitter problems in synchronous networks are easy to handle.

The capacity of Internet, at least of the last mile, has until recently been quite low. Hard compression of video and sound has been absolutely necessary, with large delays and reduced resolution quality as the result. The compression of video in TV is also extensive, in order to provide more and cheaper channels, given a certain capacity. This means long delays (seconds for MPEG-2 codecs), which is a mess for interactive multimedia.

However, end-to-end time delays in packet switched networks can be guaranteed, but since 'there is no free lunch', the resolution of the audio-visual content have to be adaptive and vary with the traffic. This combined approach is a main topic of this report.

Another requirement is that DMP services shall approach natural, real-life perception, and this can be accomplished by applying immersive, multi-view, auto-stereoscopic 3D shooting and presentation. These topics are focused in this report.

1.2 Distributed Multimedia Plays (DMP)

The concept of Distributed Multimedia Plays, DMP, was introduced as a proposal for an extension to the coming digital TV system, Multimedia Home Platform, MHP, in a Telenor project in 1996-1999 (RON99).

The main features of DMP can be summarized as follows (RON04a):

'The Distributed Multimedia Plays Systems Architecture (DMP) provides 3D multiview video and 3D sound collaboration between performers over packet networks. To guarantee the end-to-end time delay less than 10-20 milliseconds, and obtain high network resource utilization, the perceived quality of audio-visual content is allowed to vary with the traffic in the network. Parameters that are included in the quality concept, and that can be controlled adaptively, are the end-to-end delay, the number of 3D scene sub-objects and their temporal and spatial resolution, adaptive and scalable compression of sub-scenes, and the number of views. To approach the natural level of human perception, the quality has to be increased to levels that temporarily require data rates of Giga bits per second between users. An integrated resolution and traffic control algorithm denoted the RL algorithm, using traffic classes, measurement and forecasting of traffic, feedback control and traffic shaping (packet drop and source scaling), provides the necessary control. Actual DMP applications from arts are jazz sessions, song lessons, and distributed opera. Other applications are in coming generations TV (MHP extended with DMP), games, education, and near natural-seeing virtual meetings. In this paper, the focus is on the DMP application 'Virtual Dinner''.

Some service scenarios for residences:

- Film or video that is shot, synthesized, edited and stored before broadcasted or streamed (groupcasted).
- Live TV, distributed camera shot, automatic object trace and video segmentation, and a producer selecting objects, cameras and effects, extended with interactive DMP. Examples are football matches, educational sessions, and entertainment sessions with 'actors' performing from their homes, and real-time games and net based games.
- Virtual extended family. Examples are 'virtual dinner', audio-visual chat and 'virtual birthday celebration'. This scenario is similar to present videoconference meetings, with enhanced quality requirements
- Advanced games

Some professional scenarios:

- Film or video that is shot, synthesized, edited and stored before broadcasted or streamed (groupcasted).
- Live video shooting and projection, distributed camera shot, automatic objects tracing and video segmentation, 'the system' selects cameras and effects, extended with two-way DMP (videoconferencing). Examples are educational sessions (MBA, engineering, music,,,), virtual collaborative research (bio-engineering), and meetings.
- Games, used for edu, info, marketing

As shown in the papers (RON02a) (RON02b) (LIE03), the basic RL resolution and traffic control approach guarantees end-to-end time delays of user interaction. The continental delays will eventually be limited by propagation delays, other delays can be handled by using pertinent technology. The RL approach simplifies the media

synchronization and timing substantially, but at the cost of time varying temporal and spatial quality, and high requirements of capacity.

A number of papers addresses the application of traditional control theory to IP networks, see (LIE04) for summary and references.

In the European continent, a full-coverage audio-visual fiber network can be built with maximum ten switching nodes in series between any two users (RON04a). The propagation delay is less than about 35 ms, using optical fibers each carrying 100 (500?) Gbps. Knowing the propagation delay, given the link capacities, and allowing controlled packet drop, the maximum permitted queue lengths can be specified to guarantee the wanted maximum packet delay.

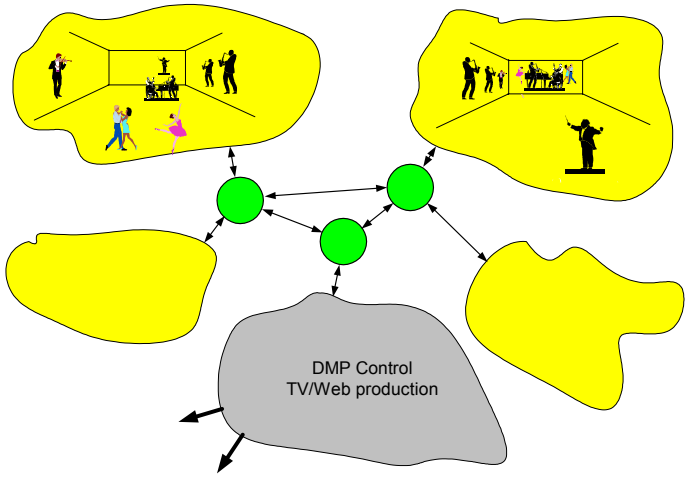

Figure 1. Distributed Multimedia Plays (RON99).

Assuming scalable and adaptive service quality, packets are dropped in a controlled way when queues are full in nodes (this is controlled service down scaling), and forecasted overloads are reported back to the contributing sources, that in turn scale down the quality and the packet rate. This, in short, is the RL Resolution and Traffic Control Algorithm (see more details and extensions later)

Eventually, the perceived audio-visual quality will determine acceptable quality of service requirements. The question is when and why will we need low delays and high resolution.

A DMP terminal system with adaptive integrated traffic and AV resolution control is shown in Figure 2. The two upper layers will always be present, while the two lower layers represent wireless and fiber alternatives.

Network nodes will in addition to normal routing functions also include traffic and resolution control.

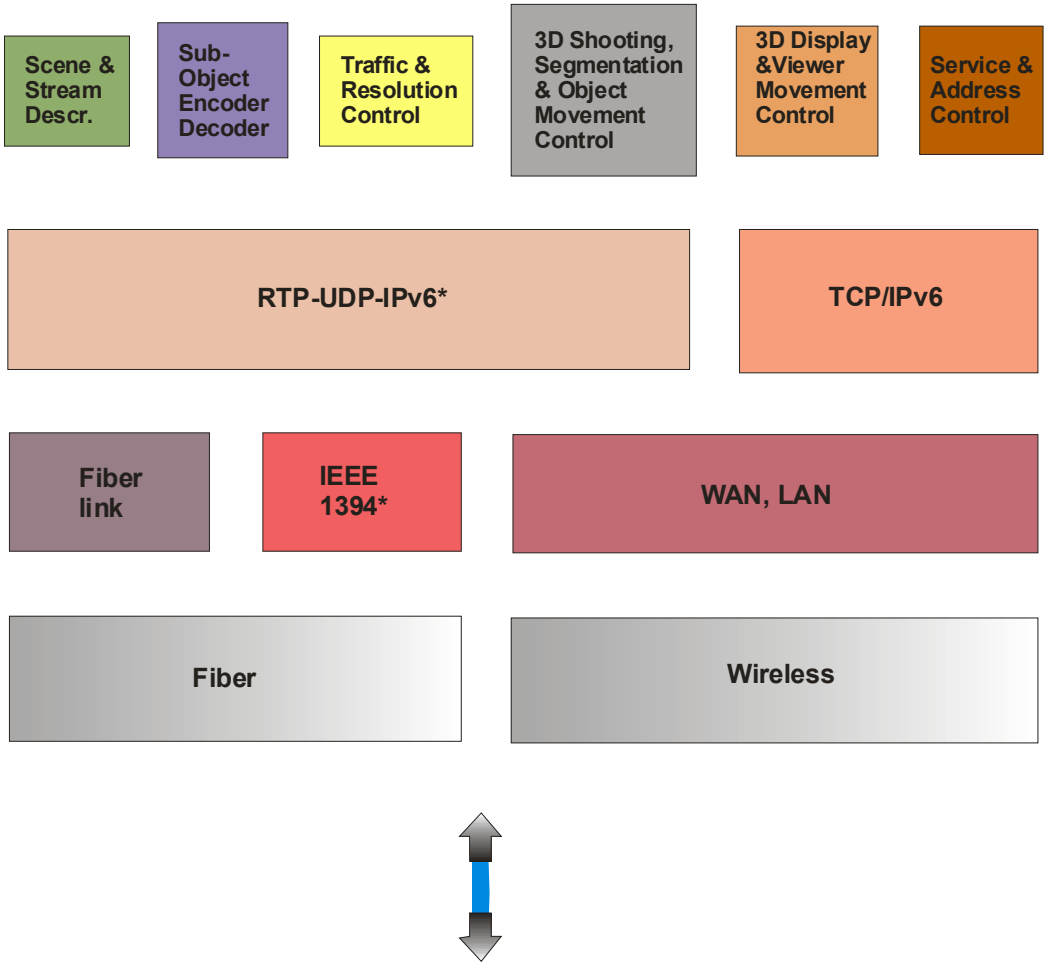

Figure 2.The DMP Terminal System (top picture, source (HHI04))

1.3 Distributed Multimedia Plays and QoS requirements

In this report, the Quality of Service concept is used in a broad sense, that is, the total quality of the service as experienced by a user (total QoS), and provided by several providers. This includes marketing, purchase, using/perceiving the service, and customer-provider relations. However, the focus is on traffic and visual resolution.

When the user buys services, the contract should also include the specification of the guaranteed QoS levels, such as time-varying audio-visual resolution and maximum end-to-end delays.

Given services as in 1.2, the QoS requirements have to be quantified in order to be measurable.

In the living room, in other residential rooms and the residential game room, family members will in the future be surrounded by 3D multi-view (multi-channel) auto-stereoscopic video shooting and displaying surfaces, and corresponding 3D audio. Various input 'devices' will be present. The 'system' will be capable of recognizing and tracking users, understanding the behavior (talk, movement and attitude) of users, and taking pertinent actions in order to fulfill the user's expectations regarding QoS.

The shooting, presentation and coding of video and audio will be object oriented, based on VRML (VRM97), MPEG-4 (KOE02) or similar. The importance of the concepts of image framing and aspect ratios as used today likely will be reduced. The natural spatial seeing of the eyes and corresponding perception will be dominating (RON04b). Of importance will be the absolute size and temporal and spatial resolution (given by pixel size and color and intensity accuracy) of displayed objects, and the seeing distance (RON04b). To avoid seeing individual pixels on an LCD color display with 0.25 x 0.25 mm pixel size, the seeing distance has to be about 40 cm. If the pixel size is 1 x 1 mm, the seeing distance should be more than 1.6 meter. To show moderate fast movements of objects over the screen, the temporal resolution of the object and the near surroundings should be at least 120 Hz (SIL04). Note that this pixel sizes should be satisfied for each view in a 3D multi-view situation.

Automatic focusing of shooting and display will partly be controlled by the 'system' that can follow and interpret the viewer's behavior and focus on various objects.

A number of interactive services (RON04a) require end-to-end delays below 20 ms. Experiments with distributed playing music have shown that delays more than a few milliseconds are disturbing (STR03). In order to guarantee this, the audio-visual content resolution has to be allowed to vary with time. The perceived quality and acceptance levels are not yet known, and have to be found by experimentation and testing. DVB addresses such problems, see (CRO04). The data rates needed to our houses will be in the Gbps range, instead of the Mbps range used today.

1.4 Some established media systems

1.4.1 Conferencing

ITU-T's recommendation H.261 (ITU93) was defined for fixed-rate media conferences. The newer H.323 (ITU03a) is intended for packet-switched networks (IP-networks). H.323 defines zones including Terminals, Gatekeeper (address translation, authentication, authorization, bandwidth management, billing, etc), Gateway (gateway between H.323 and non-H.323 networks, Router and Multi-point Control Units, MCU, supporting multi-user conferencing. A number of protocols are used, including RTP/RTCP (RTP96) for transport, H.225 (ITU03b) for registration, admission and status, and for call signaling, and H.245 (ITU02c) for control signaling. A number of audio codecs can be used, but the only mandatory is the H.261 video codec. However, the codecs are not object oriented, they are not scalable, they do not meet the requirements for high-resolution we need, and they do not support adaptive spatial and temporal resolution and individual streams for scene objects. The H.323 protocols do not support roaming. The concept of zone is usable, and maybe part of H.225 and H.245 can be used.

SIP (SIP02a) is essentially a protocol for setup, modification and release of media sessions between two or more participants over IP networks, described in RFC3261. SIP is open-ended and does not say anything about media formats. This is good because we want to define our own codecs. SIP uses RTP and RTCP for contents transport, and SDP (SDP02) for media content description. SIP defines main functional blocks corresponding to H.323. Several multiparty conference software packages are available. See for example (CIS04).

RTSP (RTS98) is a protocol for setup, control and release of media streams from a server to a client, and is described in RFC2326. It is extendable, it is possible to define own methods. It does not support varying displaying speed.

Shall we apply SIP and/or RTSP, H.323 or other? The pros and cons of SIP and H.323 have been analyzed. H.323 was characterized by 'complexity', 'structure', 'dependability', and SIP by 'simplicity', 'freedom' and 'chaos' (SIP02b). Since H.323 does not satisfy a large number of requirements for application in our future system (see above), and in many ways is a 'closed standard', we prefer to proceed with SIP's open characteristics (the 'chaos' has been reduces a little bit by establishing reference software (CIS04)). Lately, a 'SIP MCU' has been implemented (CIS04). It is a question if we need RTSP.

In a master thesis submitted at NTNU (HEL03), the DMP idea was further developed, and a specification for introducing videoconferencing into a MHP (Multimedia Home Platform) TV set-top box was presented. This service can be used as a normal videoconferencing system. But more important is the possibility of using the service to compose multi-party scenes with geographically distributed actors (at home or elsewhere), who participate with individual roles in a play or game. The composed scenes can be broadcast as a normal TV program, or be streamed over Internet, and the viewers will see the program as if it were produced in a TV studio (RON99).

Two vendors, Tandberg and Polycom, dominate the videoconferencing market today (DAG04).

1.4.2 High definition TV

HDTV has been in normal operation for many years in USA and Japan. For Europe and those countries that will use the MHP standard, an extension of MHP with an HD profile would be natural and quite straight-forward. At IBC 2004, HDTV via satellite using the DVB-S standard and MPEG-2 transport was demonstrated by DVB.

Recent research carried out by DVB members (CRO04) shows that the preferred viewing distance for watching TV is 2.7 m, and that 5-7 % of the population wants larger displays than 50". This is partly explained by the size of the living-rooms in Europe today. A comment to the DVB survey is that viewers have not experienced large, surround displays in home situations, and the questions about the wanted size of the display is not valid in our case. New audio-visual environments have to be built up/simulated, and test viewers have to be experienced users before they can give valid answers to such questions.

Japan uses the 1920x1080, 60 Hz interlaced scan HDTV format, while Europe (EBU) proposes 1280 x 720, 50 Hz progressive scan for transmission (720P50 for short). USA uses both these formats. The goal worldwide seems to be 1920x1080, 50 Hz progressive scan, but the technology is not fully developed yet. The most popular codec is the H.264 (BJO03) and Microsofts VC1 (MIC04). MPEG-2's HD profile requires a data rate of 15 (12-18) Mbps, while the same quality is obtained by H.264 and VC1 at 8 Mbps. The uncompressed data rate is about 1.5 Gbps for the 720P50 case.

The spatial resolution of HDTV is excellent at a viewing distance of 2.7 m, but at 50 cm each pixel is observable. Panning and zooming also introduce some visible artifacts (RON04c)

1.4.3 Streaming in Internet

Streaming in Internet is normally performed using the RTP protocol for content transfer, and the RTCP protocol for buffer control. Implementations of streaming systems are the Envivio (ENV04) and the Windows Media 9 for MPEG-4 streaming, and QuickTime, RealVideo and JMF (Java Media Framework) that support a number of codecs (QUI04), (JMF04).

1.5 Scene characteristics

The scenes considered here are object oriented scenes, which can be described by VRML,Virtual Realty Modeling Language (VRM97), or MPEG-4 (KOE02) which is based on VRML and further development.

The DMP systems considered here are used at home or office with a small number of viewers. Note that when the audience is large, for example in public cinema, the principles described below may not be applicable.

The concept 'importance of object' is introduced here to help quantifying the temporal and spatial resolution wanted for each object, and to help tracking them. The temporal and spatial resolution of objects may vary with time, and from object to object. Important objects are the objects viewers actually focus on, and in many cases viewers look at the same objects, but not always. The importance of an object normally varies with time from several hundred milliseconds to seconds (minutes).

To approach natural-seeing, the size of objects shall be natural (that is, the maximum size of humans and other objects should correspond to their natural size (not as in present movies where close-ups often cover most of the screen)), the displayed picture shall be flicker-free, and individual pixels shall not be visible at a 50 cm viewing distance. The normal distance between cooperating human objects may vary from 0.5 to several meters in close-up situations, but may be several hundred meters.

The natural viewing area of human eyes should become the 'frame' of the scene. Frame of the scene, as used in present TV or movie, likely will be less important.

It is quite normal that a background or another object deliberately is shot un-sharp. Today, the data rate for these object are not reduced. With object oriented, individual resolution the data rate can is such cases be dramatically reduced.

1.5.1 MPEG-4 scene and stream description

MPEG-4 defines objects, the technical term used for parts of a scene (In the art of film the term subject is mostly used (KAT91)). Figure 3 shows an example of a hierarchic scene description with three objects shown in the leaf nodes. BIFS (Binary Frames for Scenes) is the description language used to reduce the amount of data needed for each seen description (DUF02). A textual description language called XMT-A (KIM02) is used for giving textual scene descriptions.

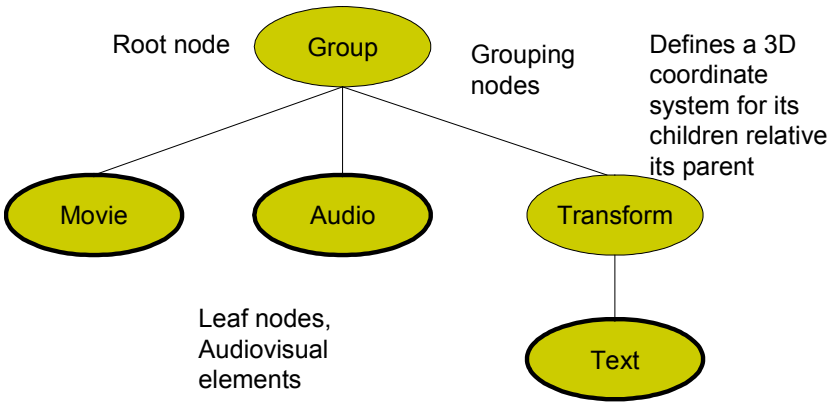

Figure 3. The hierarchic scene description of MPEG-4 (aggregated).

The audiovisual and textual data associated with each object are sent in separate streams as the example in Figure 5 shows. The relations between scene descriptions and streams are described by Object descriptor as suggested. The example shows that the movie content resolution can be made scalable, by dividing into a base layer and an enhancement layer.

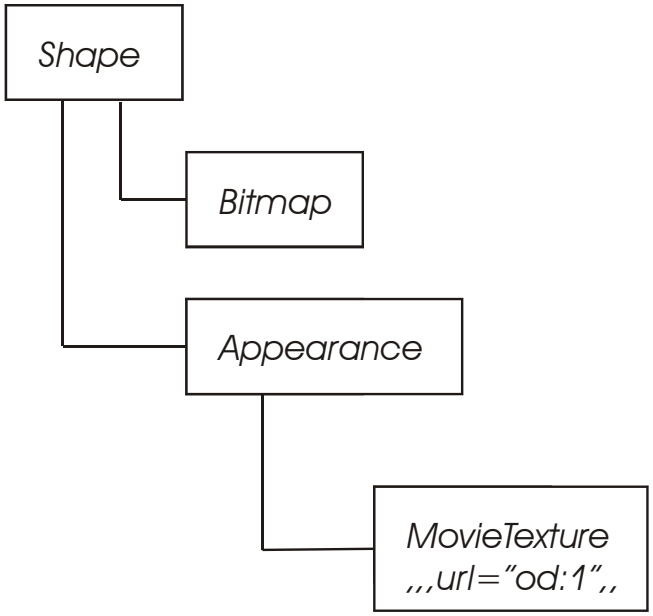

Figure 4. Details of Movie node.

XMT-1 description of Movie node:

```
<Shape>
 <geometry> <Bitmap/> </geometry>
 <appearance>
     <Appearance> <texture>
         <MovieTexture url="od:1" loop="false"
                 startTime="0.0" stopTime="-1.0"/>
     </texture> </Appearance>
  </appearance>
</Shape>
```

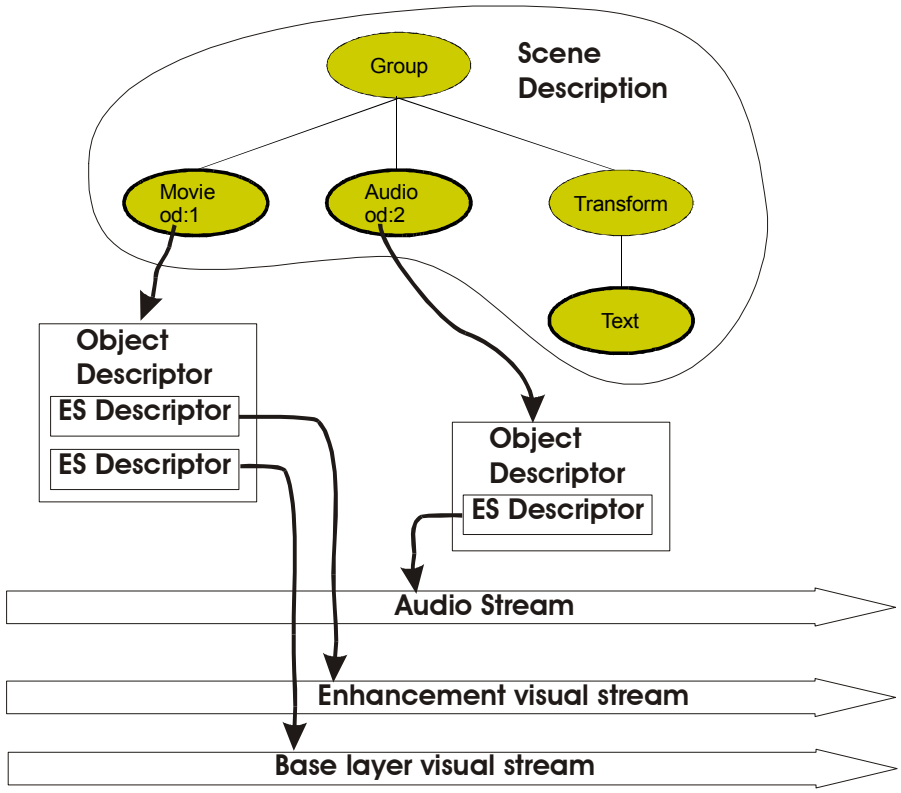

Figure 5. Stream Description in MPEG-4 (simplified).

1.5.2 Sub-objects and 2D Interlace

In the RL algorithm, each object is divided into 2 to 4 sub-objects, which are sent in separate streams. This is denoted 2D interlace. Figure 6 shows an example.

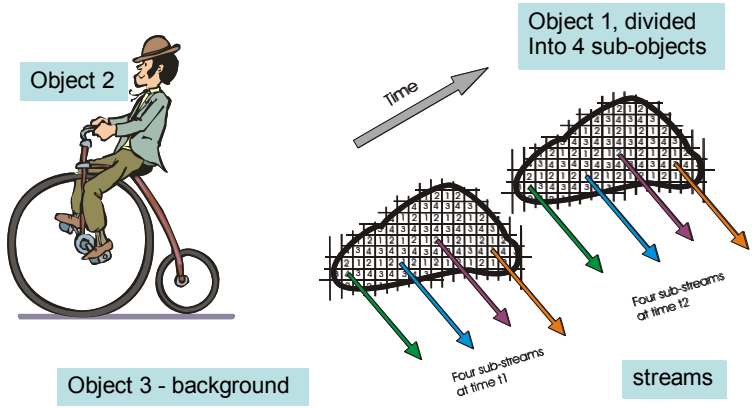

Figure 6. Object oriented scene with sub-objects.

1.5.3 Service description as part of scene description

In videoconferences based on SIP or H.323 centralized servers (MCU) normally handle setup of multi-user sessions. Another approach considered here is distributed control. This can be carried out by using MPEG-4 BIFS sub-scene descriptions. Ideally, when three persons in three different homes or offices have a virtual meeting, each person and background can be regarded as sub-scenes of the scene. The composition of sub-scenes into a scene could be handled automatically, or manually, see chapter 2.

1.6 3D Multi-view

1.6.1 Display techniques

The first section treats the displaying part of the system, while the next section has focus on the shooting of audiovisual content. In both cases it will be necessary to track the viewers, and interpret their behavior.

Natural seeing is a main issue in this report, and can be characterized by the parameters in Table 1 (SCH02):

Table 1: Natural seeing (source: (SCH02))

Amount of depth gradation	Infinite
Possible object distance	From zero to infinity
Color spectrum	Entire visible range
Size	Unrestricted in all directions
Freedom of movement	Unlimited
Number of possible observers	Only limited by viewer position
Different retina images	Always
Convergence	On focused object
Accommodation	On focused object
Moving parallax	Always

Multi-channel systems seems be good candidates for 3D television and DMP. The main drawback is the conflict between accommodation and convergence (SCH02), for distances between the viewer and the display of less than about two meters.

Table 2: Multi-channel Display Systems have the following characteristics 02 (source: (SCH02))

Amount of depth gradation	Principally infinite
Possible object distance	Clear viewing distance to infinity
Color spectrum	Limited by color depth of display unit
Size	Limited by size of display unit
Freedom of movement	Depending of number of channels, only limited by resolution of display unit
Number of possible observers	Many with variable viewing positions, depending on number of channels
Different retina images	Always
Convergence	**On fixed object**
Accommodation	**On screen surface**
Moving parallax	Always, depending on number of channels

Multi-channel display systems include several techniques: the lenticular grid technique, the barrier method and the Wavelength-selective filter array. Here, we first puts focus is on the lenticular method.

LCDs are often used in lenticular imaging because of low price, the flat construction and accurately defined pixels. A drawback is that pixel boundaries can be seen between the views as vertical bars, and another that each column of pixels shows discrete views, not a continuous view. Philips solves those problems using a slanted pixel design as shown below. Fraunhofer HHI solves the problems by tracking head movements and adjusting the lenticular display sideways (HHI04).

Philips delivers 3D autostereoscopic multiview displays, based on LCD displays combined with lenticular screens (BER97). Figure 7 illustrates the principle for a four view system.

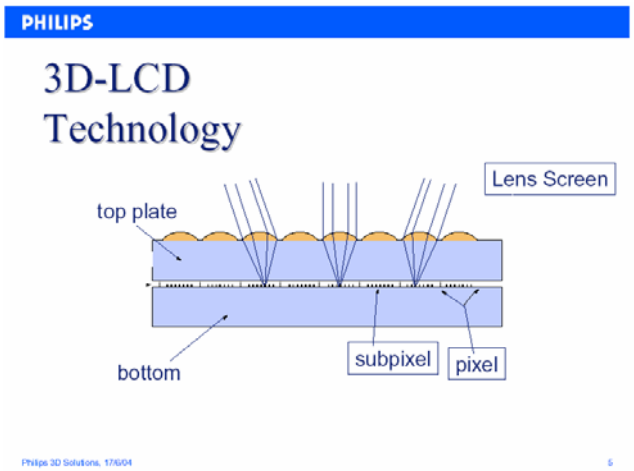

Figure 7. 3D Lenticular principle (source: (BER97)).

Instead of the normal rectangular pixels, Philips uses a slanted design where each pixel is a parallelogram as illustrated in Figure 8. The area between the pixels represents the black mask of the LCD. To understand the perception of multiview in this case, it is important to notice that the viewing angle Φ in Figure 9 corresponds to the position x, that is, a certain sub-pixel. If we go back to Figure 8, lines A and C represent viewing positions where view 2 and 3 can be seen. Line B represents a viewing position where pixels from both views can be seen. This technique nearly eliminates the black mask which can be seen when rectangular pixels are used. The explanation is that when we with one eye moves from A to C view 2 will fade out at the same time as view 3 fades in. Note that the same effect can be obtained by using rectangular pixels but slanted lenticular, cylindrical lenses.

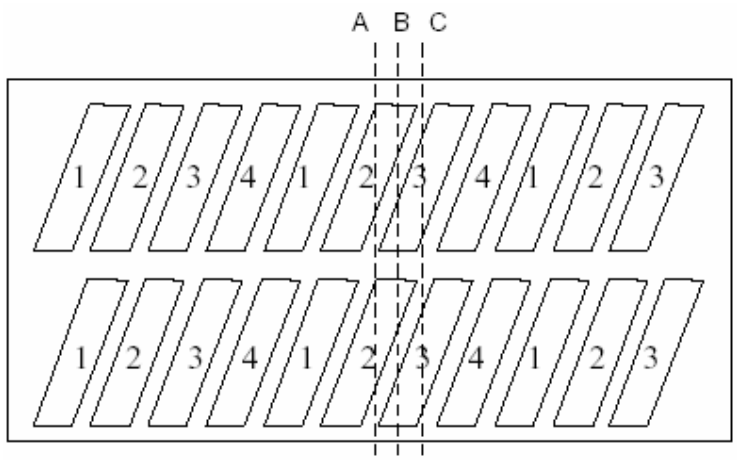

Figure 8. Slanted pixel layout (source: (BER97)).

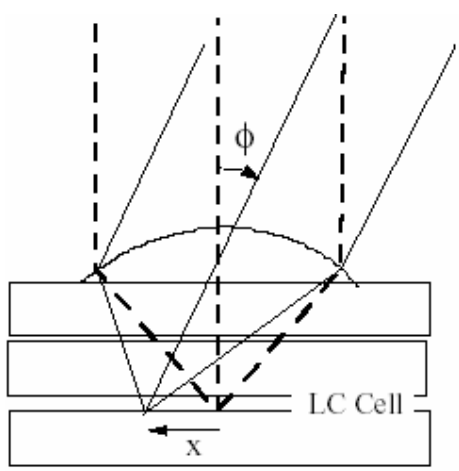

Figure 9. Details on position of sub-pixel and view (source: (BER97)).

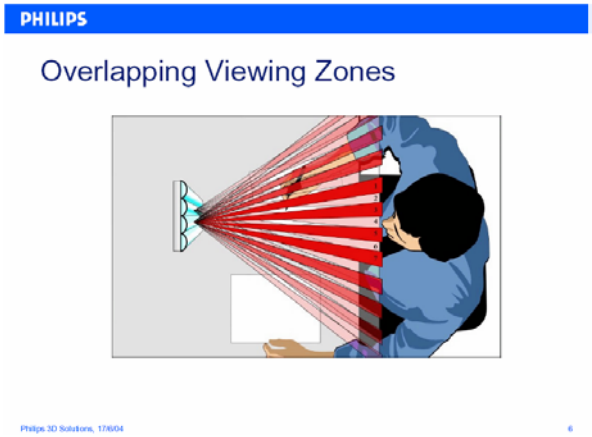

Figure 10. 3D LCD Lenticular display with overlapping viewing zones (source: (BER97)).

A large number of papers have been presented about 3D displays, see for example (SPI03), (WOE95).

Some objects in a displayed scene are more important that others, in the sense that viewers focus on one object at a time, and some objects are focused more often than others. Instead of shooting all objects with full resolution, the 'system' should track the viewers' eye movement and order the shooting cameras at the source to track and shoot the focused objects with enhanced resolution. The less important objects are shot with lower resolution.

New technologies

A promising new display technology is based on light-emitting organic materials. The OLED (Organic Light-Emitting Diode) offers brighter and more efficient pictures and cheaper displays than the LED (WEB04), (VAE03), (SHI03. Another great advantage is that the OLEDs can be built on substrates of different materials, also flexible plastic, which makes the displays 'rollable' or 'wrappable'. Small OLED displays are already on the market, and 'wallpaper' sizes are under development and will probably be available in volume in five to ten years.

It should be noticed that the accommodation - convergence problem as shown in Table 2, has been solved, se for example Fraunhofer, Heinrich Hertz Institute (HHI04).

1.6.2 Video segmentation and multiview shooting

Video segmentation is part of the research field called Computer Vision, and is about how to extract objects from visual scenes. This is normally carried out by analyses of video sequences. A large number of applications exist, for example, military tracking systems, security systems and surveillance systems.

Amongst others, Mitsubishi (MIT04) carries out research on video segmentation, and plans to integrate object detection, identification and tracking into MPEG-4 encoders and surveillance products.

The following quotation is copied from Mitsubishi's web page (MIT04):

> "After filtering the input video, markers are selected. Markers serve as the seeds of volumes. A volume is defined as the aggregation of video object planes of the same object in every frame of the sequence. Using the local color and texture characteristics, a volume is grown around each marker. The grown volumes are refined and motion trajectories are extracted. Self-descriptors for each volume, mutual-descriptors for a pair of volumes are computed from trajectories. These descriptors designed to capture motion, shape as well as spatial information of volumes. In the clustering stage, volumes are merged into objects by evaluating their descriptors. Iterative clustering is carried out until the motion similarity of merged objects becomes small. After clustering, an object tree that gives the video object planes for every possible number of objects is obtained".

An important application for DMP is face and eye recognition and tracking. In the 'Virtual Dinner' (Chapter 2) case, the 3D object in the display focused by the viewer A (at home A) at any time is defined as 'the most important object'. Via the interaction channel between the viewers A and B, a command is sent to B to shoot full temporal and spatial resolution of the focused object. At the same time, the object that viewer A focuses, has to be tracked at home B when the object moves around.
The Heinrich Hertz Institute uses an eye tracking system for there Free2C display, as described in 1.6.1.

Some objects in a scene are more important than others. This may be defined by 'the scene itself', or it has to be predefined or defined by the 'system' in real time. The importance of objects are also determined by the viewers and the objects they focus on, see the previous section. Important objects are tracked and shot with higher spatial and temporal resolution than other objects.

Note that motion estimation of object oriented scene objects is moved away from the coder (compressor) to the object and eye tracking systems.

A large number of cameras may be needed both for shooting multiple views, and also for shooting important object and less important objects.

For a theoretical introduction to camera motion estimation and projective/bilinear modeling of 3D objects, see (WAN00), (KOK03).

A large number of cameras may be needed both for shooting multiple views, and also for shooting important object and less important objects.

Figure 12 shows a multi-camera system for generation of stereoscopic views.

Figure 11. a) A blue ball and a red rod in front. b) seen from above

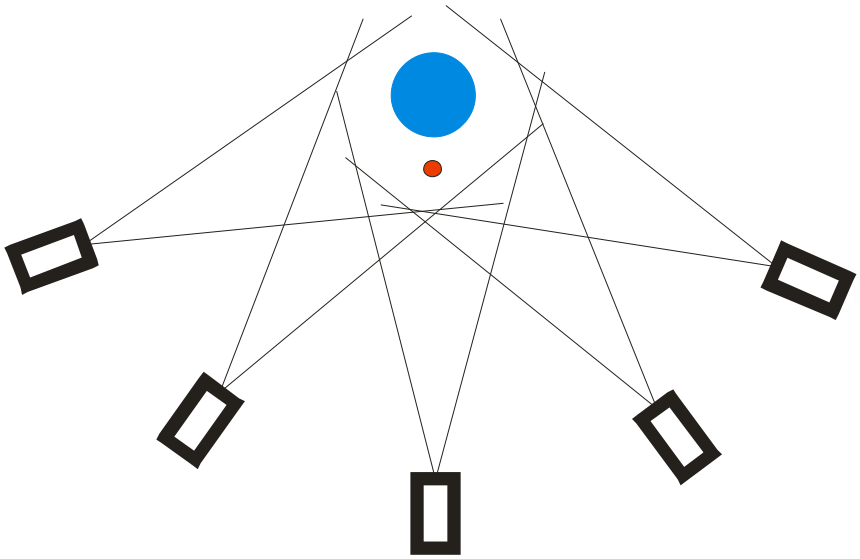

Figure 12. Five cameras shooting the ball and rod from different angels.

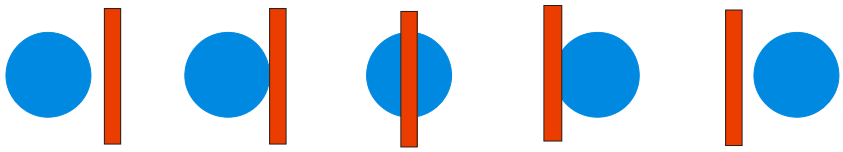

Figure 13. The pictures shot by the five cameras.

An important problem is to convert existing 2D videos to 3D. This can be accomplished through depth analyses. In addition to a normal RGB picture, a black-white picture where the grey tone indicates the depth is calculated. The quality of the converted 3D video varies, depending on the amount of relative movement of objects. This principle can also be used for representing a 3D scene with RGB and D (depth) (KAW02).

1.7 Compression, coding

In DMP systems scenes are object oriented, and objects are further divided into sub-objects and sent in separate streams as described in 1.5.2 and 1.6.2. This also implies that some objects can be tracked and shot with low time delays.

Normally, video compression also involves motion estimation. In our case, the motion estimation is carried out in the video segmentation process (necessary in order to track moving objects). That means that compression can be applied within sub-object snapshots (or the samples packed into single packets), and a low processing delay can be obtained.

In some cases, for example a background scene with moving objects that are not important, can be treated as one object. In this case the time delay is not critical and the object can be compressed using standard MPEG-4 compression (with motion estimation).

It is not a goal to obtain constant bit rate from the coder. On the contrary, to obtain a time varying resolution a number of parameter in the coder can be set adaptively, resulting in a variable bit rate.

Many compression algorithms use the Discrete Cosine Transform (DCT) together with Huffman coding. The DCT can be applied on 2D blocks of pixels, or on 1D bit strings. Assuming YUV video components as input to the coder, Y and U,V can be represented with parameters that can be adjusted to get the wanted resolution.

Compression of sub-objects can be performed in two ways (at least):

1. sample all sub-objects simultaneously, send in separate packet streams immediately, decode and display when received.
2. sample sub-objects interlaced, send in separate packet streams.

To speed up the processing and avoid delays use DCT-look-up tables (LUT) and Huffman-LUT.

The Wavelet transform is similar to the Fourier transform, but uses two parameters, that is, the transformed data is represented in the time-frequency space. The Wavelet transform makes level scalability easy (MAR02a) (MAR02b), (JAI99), (CYC02), (XIN00).

Resent research shows that video codecs using the wavelet transform, improves the performance of coding compared to MPEG-4 and H.26L/H.264.

1.8 Resolution & traffic control

With a given network transport capacity, and stochastic input, IP networks will from time to be saturated, and packets have to be dropped. In addition, new requests for admission to the network can be denied.

Another philosophy that can be applied to share the limited capacity resources, are temporarily to reduce the content resolution when the traffic peaks. The content resolution can be reduced in a controlled way by either dropping selected packets in routers, and by down-scaling of contents of traffic sources. At the same time the maximum end-to-end time delay can be guaranteed.

1.8.1 The RL Resolution & Traffic Control Algorithm

The detailed description of the RL-C algorithms, an enhancement of the RL algorithm described here, is given in the next section. The bases for the algorithm is the traffic shaping concept, first proposed in (RON81) in 1981. The traffic shaping concept included traffic classes, measurement and forecasting of traffic in the network, feedback control with reporting high loads back to the traffic sources, and shaping the traffic. Several years later the concept was applied to ATM traffic control. The RL algorithm was further elaborated by Rønningen and Lie in several papers in 2002 and 2003 (RON02a), (RON02b) (LIE03).

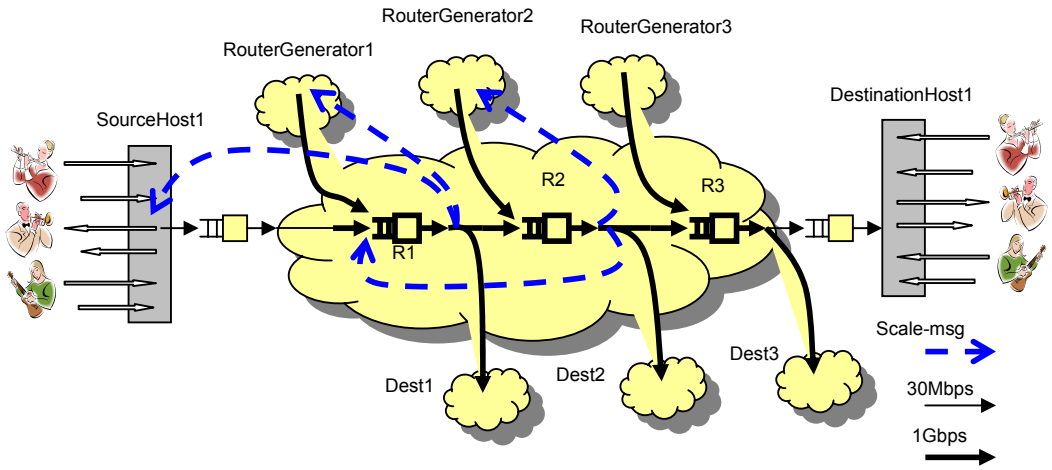

Figure 14. Simulation model

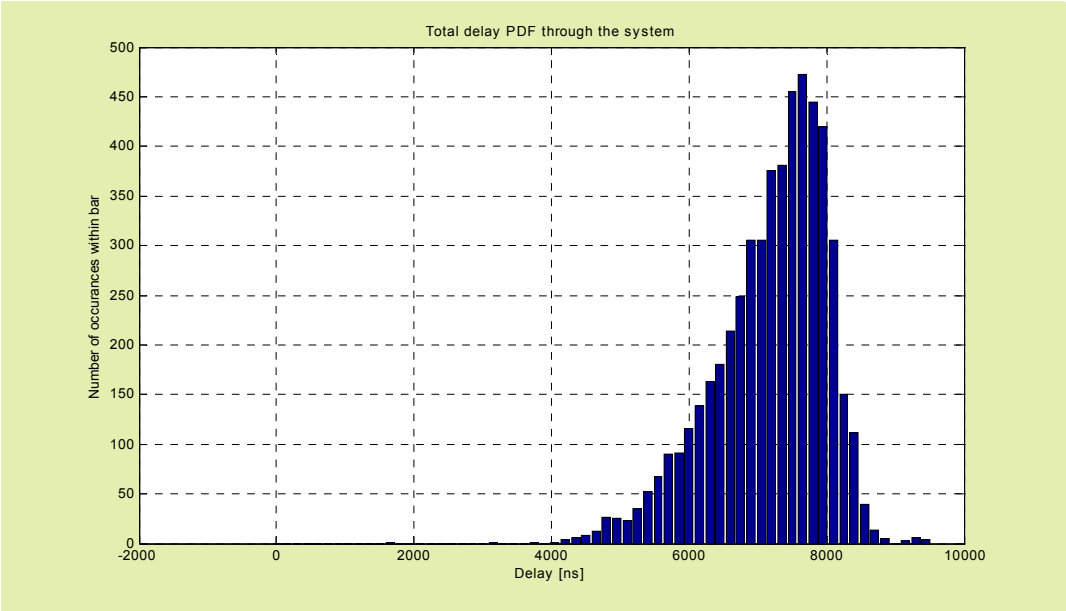

Figure 15. Simulated end-to-end time delay, histogram

The simulation shows, as expected, that the end-to-end delay can be guaranteed. However, no formal tests have been conducted to show to what extent and when the time-varying resolution of audio-visual content is acceptable.

1.8.2 Other traffic control algorithms

A large number of traffic control algorithms have been proposed. The selection below was largely developed in the 1990s and 2000s.

ABE – Alternate Best Effort, is a queuing scheduling proposal that trades loss for latency in giving fairness to links carrying both TCP and UDP traffic (HUR01).

ECN – Explicit Congestion Notation, packets are marked instead of dropped. The destination node signal the marking back to the source which acts accordingly (FLO97), (RAM01).

AIMD – Additive Increase, Multiplicative Increase, the control method used in traditional TCP. A stochastic analyses is presented in (DUM02).

TFRC - TCP Friendly Rate Control, gives TCP friendliness in the long run, but the variance of data rate is less than for AIMD (HAN03).

DCCP – Data Congestion Control Protocol, can switch between AIMD and TFRC. DCCP is connection oriented and runs on top of UDP (KOH04).

RED – Random Early Detection, incoming packets are randomly dropped with a probability based on a cost function of average queue length (RED04).

REM – Random Exponential Marking (ATH01).

AQM – Active Queue Management, regulators applied in routers and terminals in packet networks with mixed TCP and RTP/UDP traffic, in order to obtain fairness (LIE04a), (LIE04b), (KIM03), (HOL01).

XCP is new transport protocol described by (KAT02) which generalizes the ECN, and decouples utilization control from fairness control. Instability problems are avoided.

1.9 Service setup and address allocation

The SIP/TCP/IPv6 protocols will be used for set-up, and the SIP URL identifies services, scenes and sub-scenes. Parameters in the RTP/UDP/IPv6 protocol header identify the same for content transfer. IPv6 addresses will be allocated in reserved way, and the SIP and RTP protocols will be adapted for this application.

2 The Virtual Dinner Scenario

The Virtual Dinner scenario is based on principles described in 'Background' and further elaborated here. The collaboration between people and groups of people at home will be based on virtual reality. The most important is to approach natural perception. To obtain that, autostereoscopic multiview 3D and object based scenes are applied. Objects will be treated as sub-objects. The spatial and temporal resolution may vary from object to object and with time, and the minimum pixel size is given by that pixels shall not be visible as individual points at a viewing distance of 50 cm. The picture boarders should be outside the viewer's scope. The 'system' interprets the number of persons and their focus points in the display and decides the actual views that are sent at any time. To perform this, very fast end-to-end response is required.

2.1 The user's quality requirements ten years ahead

It is always hard to predict what user requirements will be ten years ahead (history shows), but the perceived naturalness should be dramatically improved compared to existing systems (HTDV in 2D, RAVE in 3D). It is assumed that the display will cover a corner of the living room, from floor to ceiling, up to 2,5 + 2,5 m wide, and parts of the ceiling. The system should provide up to at least 25 views (TBD), where each view shall have a spatial resolution of 0.25 x 0.25 mm pixels to be individually invisible at 50 cm. Important objects can have a temporal resolution down to 5 ms and full spatial resolution, while a static background object can be updated every 100 ms to 1 second and use pixel size from 0.5 x 0.5 to 1.0 x 1.0 mm. The number of sub-objects per object will be up to 16, sent as independent streams. The YUV luminance and chrominance components shall have at least [12,10,10] bits accuracy, and 4:2:2 sampling, for important objects it should be possible to use [16,12,12] and 4:4:4.

It is likely that a user wants to leave the living room and seamlessly continue the media session in the garden using a mobile 3D display. This can be of a flexible type (WEB04) and should have a size 1.6 m wide and 0.9 m high. The requirement to the contrast ratio and the luminance of this display is higher than for the indoor type. Else, the requirements to the display resolutions should be as that for the living room case.

Furthermore, the user may want to drive to another physical place without interrupting the media session. In this case the user cannot drive a car (use a bus, have another driver, program the car to find its path to the destination in a safe way), but put focus on the media session. The requirements to the display are still 3D autostereoscopic, but one to three views should be sufficient. The pixels should be smaller than for the living room case, because the viewing distance to the display will be shorter, say 30 cm.

2.2 The Virtual Dinner

The Virtual Dinner service allows colleagues, friends, family members to have dinner together, even if they are not in the same city. This is a type of enhanced video conference, where the requirements of naturalness are severe, as described in the previous section.

Consider a situation where three researchers have dinner and discuss research problems. One researcher lives in Trondheim, Norway, while the other two live in Padova, Italy and Poznan, Poland. Generally, there can be more than one person in all places. Persons may come in and go out of the scene, or walk around the room, but will most of the time sit or stand, talking. The faces and hands of the persons talking are regarded important objects. From time to time, the whole body may be an important object. The food and plates are also important objects. There may be other important objects as well, a painting on the wall, a dog, etc. The dining rooms are background scenes.

The behavior of the scenes may be as follows: Researcher A in Trondheim comes into his dining room, sits down in his sofa and requests a Virtual Dinner with researcher B in Padova, also sitting in his sofa. The system finds two faces and a plate with food researcher B. Researcher A goes out for a plate of food , comes in again, and sits down. The system finds the plate and the food. They start eating and talking. Researcher B stands up and walks sideways in the room. After a few minutes they need to talk to researcher C in Poznan, and set up a three party DMP. After eating, B leaves the room. A asks the system to disconnect B. Researcher C has to leave home and go to his office. But he wants to continue the session with A while traveling to the office.

Characteristics of objects of the Virtual Dinner

When the researchers are at home the objects of the scenes can be characterized as follows:

Background objects
YUV - [12,10,10] bits, 4:2:2 sampling
8 sub-objects, pixel size 0.25 x 0.25 mm
3 views (adaptive)
Compression 100x, H.264-like
Update every 500 ms over air (5 ms in receiver)
size 3000 x 2000 mm
Uncompressed 12.7 Gbps, compressed 127 Mbps

Body Object1 approaches fast towards dedicated cameras
The 'system' tracks the object
YUV - [16,12,12] bits, 4:2:2 sampling
16 sub-objects, pixel size 0.25 x 0.25 mm
3 views (adaptive)
Compression 5x, DCT-Huffman or Wavelet, on sub-objects
update 5 ms over air
Size 200 x 600 mm, increasing to 600 x 1800 mm
Compressed min. 6.45 Gbps, Compressed max. 58 Gbps

Body Object2 moves sideways
The ’system’ tracks the object
YUV - [16,12,12] bits, 4:2:2 sampling
16 sub-objects, pixel size 0.25 x 0.25 mm
3 views (adaptive)
Compression 5x, DCT-Huffman or Wavelet, on sub-objects
update 5 ms over air
Size 300 x 900 mm
Compressed 14.5 Gbps

Face Object
The ’system’ tracks the object
YUV - [16,12,12] bits, 4:2:2 sampling
16 sub-objects, pixel size 0.25 x 0.25 mm
3 views (adaptive)
Compression 5x, DCT-Huffman or Wavelet, on sub-objects
update 5 ms over air
Size 200 x 300 mm
Compressed 3.22 Gbps

Hands and arm Object
The ’system’ tracks the object
YUV - [16,12,12] bits, 4:2:2 sampling
16 sub-objects, pixel size 0.25 x 0.25 mm
3 views (adaptive)
Compression 5x, DCT-Huffman or Wavelet, on sub-objects
update 5 ms over air
Size 200 x 300 mm
Compressed 3.22 Gbps

Plate and food Object
The ’system’ tracks the object
YUV - [16,12,12] bits, 4:2:2 sampling
16 sub-objects, pixel size 0.25 x 0.25 mm
3 views (adaptive)
Compression 5x, DCT-Huffman or Wavelet, on sub-objects
update 10 ms over air
Size 200 x 300 mm
Compressed 1.61 Gbps

Characteristics of objects of the Virtual Dinner, mobile researcher
When the researcher C goes to his office his display shows the objects of the scene as follows:

Background object
YUV - [10,10,10] bits, 4:2:2 sampling
4 sub-objects, pixel size 0.25 x 0.25 mm
3 views (adaptive)
Compression 100x, H.264-like
Update every 100 ms over air (10 ms in receiver)
size 500 x 280 mm
Uncompressed 1.344 Gbps, compressed 13.44 Mbps

Body Object1 approaches fast towards dedicated cameras
The 'system' tracks the object
YUV - [16,12,12] bits, 4:2:2 sampling
16 sub-objects, pixel size 0.25 x 0.25 mm
3 views (adaptive)
Compression 5x, DCT-Huffman or Wavelet, on sub-objects
update 10 ms over air
Size 20 x 50 mm, increasing to 200 x 280 mm
Compressed min. 26.9 Mbps, Compressed max. 1.5 Gbps

Face Object
The 'system' tracks the objects, shoots with dedicated cameras
YUV, [16,12,12] bits, 4:2:2 sampling
16 sub-objects, pixel size 0.25 x 0.25
3 views (adaptive)
Compression 5x, DCT-Huffman or Wavelet, on sub-objects
update 10 ms over air
Size 100 x 100 mm
Compressed 269 Mbps

Traffic generated by researcher A, sent to researcher B

Researchers A, B and C generate different traffic. Researcher A generates about 7 Gbps to start with. When he gets on his feet and walks out of the room, the date rate increases rapidly to a maximum value of close to 60 Gbps, and then decreases down to the background object traffic, which is 127 Mbps. When he returns with his plate, the data rate increases to maximum, and when he sits down the face, arms & hands, and plate & food dominates the data rate, about 8 Gbps.

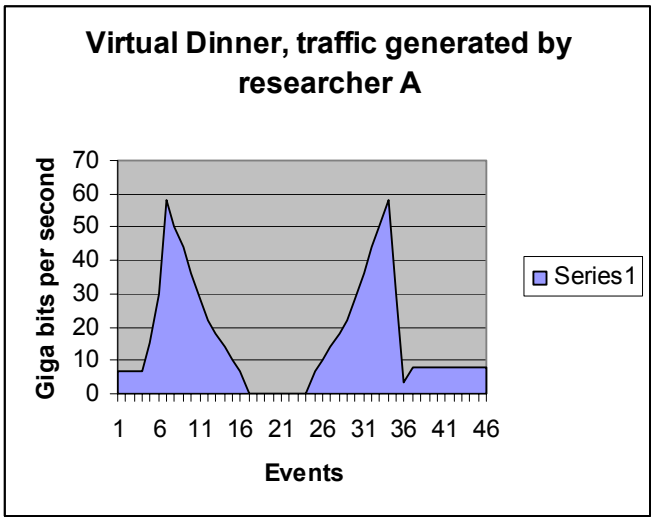

Figure 16.

3 Traffic generated by advanced scenes

(to be included in a later version)

4 Backbone networks

4.1 A European fiber packet network architecture

The propagation delay between any two users in Europe can with optimal topology, be less than about 30 ms. In order to study practical cases, Europe has been divided into regions each of less than 85 millions of users, where the delay typically is less than 10 ms within each region.

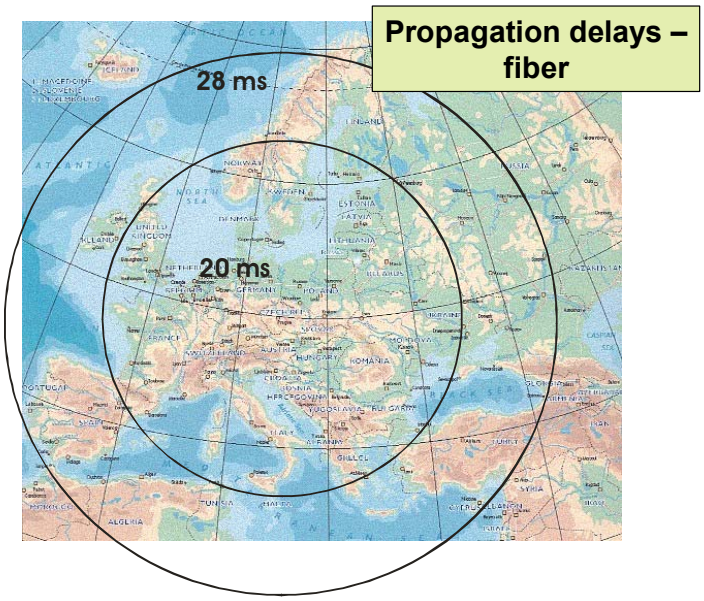

Figure 17. Propagation delays (minimum) in Europe

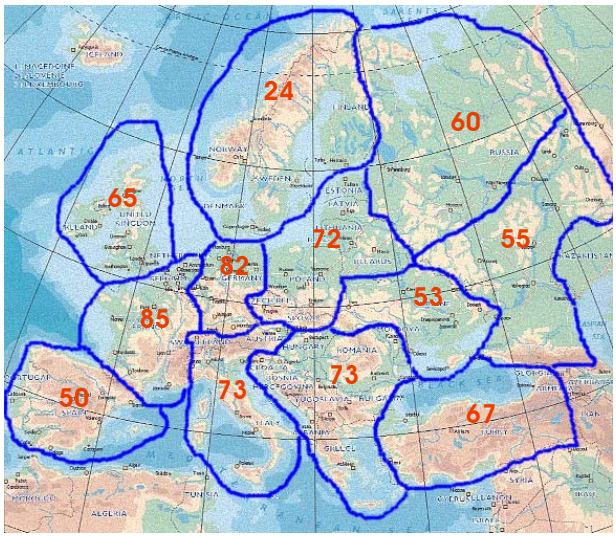

Figure 18. European regions

Europe is divided into 12 regions (Turky included), each of less than about 85 mill people and with less than 10 ms propagation delay. The video network can be built with seven node levels as shown in section 4.3. The capacities into and out of a node will be say 512 x 100 Gbps.

4.2 RTP-UDP-IPv6 applied for adaptable resolution and traffic control

The headers of the three protocols are combined into one header that is processed in parallel. Some of the standard parameters are redefined for special purposes in order to increase the performance and to reduce the complexity of the network (RON04d).

The combined RTP/UDP/IPv6 header

The combined header consists of the standard parameters, but some of them are used for special purposes here in order to increase the performance and to reduce the complexity of the network. The parameters allocate bits and bytes in the order as shown below.

The IPv6 header

- **Version** - 4-bits, protocol version = 6, drop other
- Class - 4-bits, priority, TC1, TC2, TC3
- **Flow Label** - 24-bits, optional
- **Payload Length** - 16-bit unsigned integer, the rest of the packet following the IPv6 header, in bytes. Fixed for RTP/UDP/IPv6 packets
- **Next Header** - 8-bits, identifies the type of header immediately following the IPv6 header. RTP/UDP or RTSP/TCP
- **Hop Limit** - 8-bits unsigned integer. Decremented by 1 by each node that forwards the packet. The packet is discarded if Hop Limit is decremented to zero. Set to 32.
- **Source Address** - 128 bits. The IP address of the initial sender of the packet.
- **Destination Address** - 128 bits. The IP address of the intended recipient of the packet (not necessarily the recipient if an optional Routing Header is present). ScaleMsg has a reserved address = xyz.

IPv6 Extension headers:

- **Routing** - Extended routing. A routing header is used to indicate the previous node IP address
- **Fragmentation** - Fragmentation and reassembly, not used for RTP/UDP/IPv6
- **Authentication** - Integrity and authentication; security, TBD
- **Encapsulation** – Confidentiality, TBD
- **Hop-by-Hop Option** - Special options that require hop-by-hop processing

The UDP header

- **Source Port,** 16 bits, specifies the port of the sender.
- **Destination Port,** 16 bits, the port this packet is addressed to.
- **Length,** 16 bits, the length in bytes of the UDP header and the encapsulated data. The minimum value is 8.
- **Checksum,** 16 bits, not used

The RTP header

- **V**-Version, 2 bits, private
- **P**-Padding, 1 bit, option
- **X**-extension, 1 bit, option
- **CSRC count**, 4 bits
- **M**, 1 bit, option
- **PT**-Payload type, 7 bits, private use, set to 1111111
- **Sequence number**, 2 bytes, reserved for 2D interlace coding and packet dropping, see sections above.
- **Timestamp**, 4 bytes, optional
- **SSRC** – synchronisation source identifier, 4 bytes, used for identifying Sub-object streams
- **SCRC** – contributing source identifiers, 0-60 bytes, TBD

4.3 Addressing

SIP URL provides a global addressing scheme, and the following IPv6 addresses are allocated:

Node		*IP address,*	*bits*
		(Format prefix)	3
• GlobalNode,	**G**	(TLA)	13
• EuropeNode,	**E**	(reserved)	8
• DistrictNode,	**D**	(NLA1)	12
• CityNode,	**C**	(NLA2)	12
• VillageNode,	**V**	(SLA)	16
• FireNode,	**F**		16
• FireLink,	**FL**		48

4.4 The RL-C Resolution & Traffic Control Algorithm

The RL-C algorithm is a variant of the original RL algorithm (RON02), and has to be implemented in all packet switches and traffic sources. The RL-C algorithm goes as follows:

With period t_0 do

Read the link queue length x_i and the number of dropped packets z_i.

If $x_i < x_{\max}$ then

calculate the scale factor $sf_i = \frac{\mu}{\lambda_i}$ where μ is the link capacity and

$\lambda_i = \frac{x_i - x_{i-1}}{t_i - t_{i-1}} + \mu$, is the arrival rate. *)

If $x_i = x_{\max}$ then **)

calculate $sf_i = \frac{\mu}{\lambda_i}$ where $\lambda_i = \frac{z_i}{t_i - t_{i-1}} + \mu$

For i = 1, 2,, r periods calculate the weighted scale factor

$s_i(j) = c_i sf_i + s_{i-1}(j)$, c_i are constants decreasing exponential backwards in time, and where $s_o = 0$.

When $i = r$ do

- $s_r(j) = s_r(j) * k$, where k is a parameter (affects the time to stable state).
- read incoming Scale Messages, ScM(*s(q,)*), where q is the neighboring node connected via link j.
- find s*(j) = max[s(j), s(q)]*
- when j is this node, calculate *s(p) = (CntUp(p) *s(j)/*$x_{\max}$*)* where *CntUp(p)* is the part of all incoming packets to j sent by p, for all neighbor sources p.
- send ScM(*s(p)*) and reset *CntUP(p)*
- set drop-limit $x_{\max}$ <TBD>

After a delay t_d , ScM(*s(p)*) reaches a Traffic Source,
that in turn scales the traffic:

- From look-up table LUT(*s(p)*), find the sub-object scale profile
- The LUT may contain non-linear functions.

A more sophisticated use of *s(p)* will be that the source takes the history into account, that is, samples from periods r-1, r-2,,,, r-n to find the next values of *s(p).* The formulas above can be reused, but now with subscript *i* = *r*, and as the regulator in a traditional control loop.

ScMs use the same combined RTP/UDP/IPv6 header as other packets in the network.

*) In time period $t_i - t_{i-1}$ the change in queue length is $x_i - x_{i-1}$ (no packet drop), and according to Little's formula (LIT61) $x_i - x_{i-1} = (\lambda_i - \mu)(t_i - t_{i-1})$, where λ_i is the average arrival rate in the period, and μ is the link capacity (packet rate).

**) If x_i passes the $x_{\max}$ limit more than once in the period, the time from the last passage to the end of the period is used to calculate sf_i .

4.5 Example network

To support the Virtual Dinner scenaro, the following network and usage are assumed: Trondheim (path m, F node to E node) and Padova (path a) send a large number of packets to Poznan (path n, E node to F node). The capacities [Gpps – Giga packets per second] between nodes are given in Figure 19. The network is part of the European network, but only the paths of interest are shown in the figure. The traffic considered in the example is aggregated from several sources. The traffic source model is simple, packet rates are constant within an interval, and change in discrete steps.

Since the contribution from the propagation delay already is critical for several DMP applications in Europe, the queuing delays in nodes should be small compared to the propagation delay. It is assumed that the maximum waiting times in F, V and C nodes are 0.1 ms, in D nodes 0.5 ms, and E nodes 1 ms. When the outgoing link capacities are known, it is simple to calculate the maximum queue length. At time t = 0 packets start arriving with constant rate λ and depart with rate c, the link capacity. After time t the number of packet in queue is $x = (\lambda - c)*t$ For example, when the output link capacity from an E node is 12 Gpps, the input 24 Gbpps and the time 1ms, the maximum queue length is 12 M packets.

Users in Trondheim and Padova start sending with a packet rate of 15 [Gpps] each. Following the path m, after 1 ms the packets reach the D node. Since the capacity between D and E is only 12, packets are queued, and after 1.5 ms packet dropping starts in D. At the end of the forecasting interval (1 ms) D sends ScM down to C, which immediately forecasts its load and sends ScM down to V. After about 1 ms the ScM reaches F and its traffic sources which scales down to 12 (how?). The time before the overload situation in D is registered in F, is 1 + 1 ms propagation delay + 1.0 ms forecasting delay max waiting in D = 3 ms. After 4 ms (max) the traffic into D is reduced to 12.

The maximum queue length in D is reached after 0.5 ms, and is x_{max} = (15 – 12)*0.5 = 1.5 M packets. Table 3 shows a special case, when the forecast interval starts at the same time as the step in packet rate.

The propagation delay between D and E nodes is 5 ms. Between E and E it is 10 ms. After 17.5 ms the traffic 12 from Padova and Trondheim reach the E node of path n, and offers 24 to the E node in path n, which can carry only 12. With a forecasting interval of 5 ms for E nodes, the overload situation is reported back to the E nodes of paths m and a after 16 + 2.5 + 10 ms. It takes additional 6 ms before the ScMs reach the sources. That is, a total of 34.5 ms. Note that if the destination user is overloaded, it takes additional 12 ms, together about 46.5 ms to slow down. Meanwhile packet are dropped in a controlled way, so that the destination user can reconstruct the content by interpolation, giving a period of about 46.5 ms where the content quality is temporarily reduced from outstanding to very good. The perception of such quality reductions shall be tested.

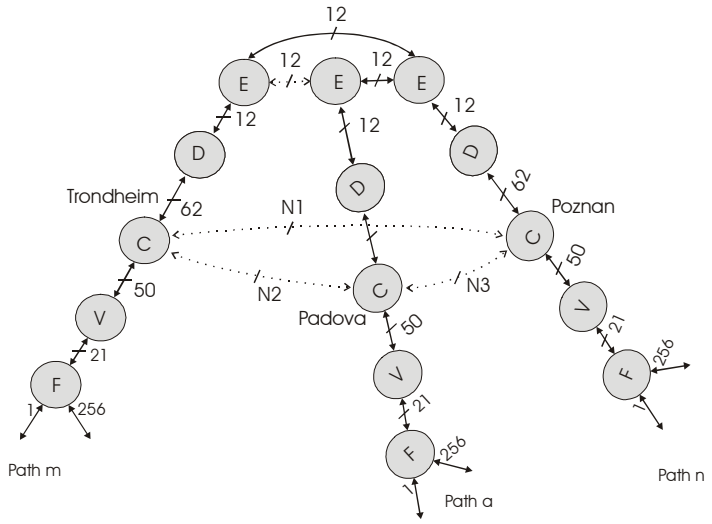

Figure 19. A video network supporting the Virtual Dinner

5 Access networks

5.1 Fiber to the home - FTTH

When capacities higher than 10 Gbps are required, users need individual optical fiber cables from the packet switching node. The capacity is assumed to be scaled in steps of N x 10 Gbps, N = 1, 2, 3,,, . In ten years, it is expected that fiber link with capacity 100 Gbps is available at an affordable cost.

5.2 Extended IEEE1394 access to packet networks

An alternative is an enhanced ‘1394 with capacity of the busses extended to 12.8 Gbps and at least 10 km lengths, using optical fiber cables and new IC technology (Intel). It is possible for each user to reserve up to 80% of the capacity (10.2 Gbps), and the reservation can take less than 100 microseconds, including 50 microseconds propagation delay of the fiber cable, when the source packet generators directly reserve the capacity.

The motivation for using extended ‘1394 as access net is it’s bus structure, which in many cases gives much less cable meters, than other topologies. With a reservation system like ‘1394, fairness can easily be taken care of. The introduction of long fiber cables to residences, opens up the possibility to introduce fast packet switches at home, or wireless ‘1394.

6 Packet switch architectures

Packet switches can in principle be realized using input queuing only, output queuing only, or a combination of input and output queuing. In the middle there is principally a space switch, the switch fabric, which can be realized in several ways, e.g. a crossbar switch or a shared-memory switch. Many commercial packet switches are based on output queuing only.

6.1 Short review of packet switch state-of-the-art

A number of papers have compared Input Queue (IQ), Output Queue (OQ) and Combined Input Output Queues (CIOQ) (CHU99) (KAR87) implementation of queues for packet switches. A qualitative summary can be given as follows:

When FIFO (First in First Out) IQ is used, and the switch fabric is non-blocking, then each queue and path can be modeled as a G/D/1/k queue (k is the maximum number of packets that can wait in queue), and there is no point having higher capacity of the fabric data path than the output link. When there are packets waiting in the queue, the channel utilization will be 100%. From queuing theory we know that this cannot be obtained over time for a general arrival distribution. However, when the packet length is short (say 1500 bytes) and the data path data rate is in the Gbps range, the mean queuing delay is in the microsecond range for a channel utilization of about 99 % (RON02a). In practice, the maximum queue length will be limited. An arriving packet will then experience a waiting time that is equal to the number waiting in queue multiplied with the deterministic service time (+ the residual service time of the packet in service). This means that the maximum time through the switch can be guaranteed. Unfortunately, the FIFO queuing discipline is not the most efficient. Packets waiting in queue could have been served, if the corresponding output link was idle. This is the HOL (head of line) blocking.

On the other hand, when OQ is used and the fabric path date rate is equal to the capacity of each link, an output queue can be busy or full when a packet arrives, and the packet gets lost. The queuing system is still of the G/D/1/k type. This can be improved by using a fabric of higher data rate for each path, the so-called 'speedup of the switch', or increase the number of parallel paths through the switch to OQ buffers.

If we introduce CIOQ, the efficiency of the switch can be increased (CHU99), because arriving packets can be buffered on the input when the output queue is full or busy. In practice, the number of buffers for each queue will be limited, and in cases with highly unbalanced traffic, queues will eventually run full, and packets have to be dropped. In this case the arriving packet sees two queues in series. A lot of streams are merged and split through the fabric, it should be investigated whether the input traffic to the output queue approximately has Markovian characteristics, so that the two queues can be considered as part of a simple open Jackson network. If this assumption holds, the waiting time through the switch is a sum of two waiting times that can be modeled analytically.

6.2 Guaranteed end-to-end time delay and acceptable resolution variation

It is easily shown that the end-to-end time delay can be guaranteed, and that the propagation delay is the limiting factor. Assume a propagation delay between two European cities of d milliseconds, and a network hierarchy of nodes giving l links maximum. Furthermore, assuming a constant packet length and a constant link capacity, delay h for sending a packet is constant. There is a single queue with limited waiting room N for each link. If the queue is full when a packet arrives, the packet is dropped. When the last packet in the queue arrives, it has to wait until all packets in the queue are sent out plus the remaining service time of the packet being sent out. The maximum remaining service time is h (uniformly distributed $(0,h)$). The total maximum delay is then $l*(N*h + h) + d = l*h*(N+1) + d$. The design question is then to decide N and the number of parallel links to have an acceptable time delay, and at the same time an acceptable time varying resolution, when the traffic input is given.

6.3 A packet switch implementing the RL-C algorithm

The switch to be proposed here shall implement a modified version of the RL traffic control algorithm, called RL-C, see 4.4. Two traffic classes TC1 and TC2 shall be supported (RON02a). The second class TC2 ensures that packets always arrive correctly to the destination, but the time delay is not critical (TCP end-to-end controlled traffic). Queues can be implemented as FIFO queues, and the number of buffers should be high, so that the probability of dropping packets can be kept low.

TC1 has severe time delay requirements ('real time streaming'), but carefully selected packets can be dropped in a controlled way. The CIOQ scheme (CHU99) is here extended with controlled dropping of packets at the front and the back of input queues, and at the front of output queues. And to reduce the probability that the next packet to be switched shall see a busy output, several packets n can be transferred in parallel to output buffers for each output queue, and later be forwarded out serially. The RL traffic control algorithm operates on the output queues. The architecture is shown in Figure 20. Each input and output has a double queue, one for each traffic class. TC2 are served only when there are no TC1 packets to serve. However, if a TC2 packet is being served and a TC1 packet arrives, the TC2 packet is allowed to finalize before the TC1 class takes over. At a later stage, searching the queues for specific packets and sorting for priority will be considered.

A packet is always delayed by the propagation. Between regional EuropeNodes in Europe, it is assumed to be less than 10 milliseconds. Between a EuropeNode and DistrictNode the propagation is normally less than 5 milliseconds, while between any two other nodes the it should be less than 2 ms. The play-out time of a packet is L/C, where L is the packet length and C is the link capacity (bits per second).

When a packet arrives to the switch, the data are converted from serial to parallel (64, 128 or higher data bus width). The Packet Header (PAHE) Process then analyses the packet header. The analysis takes some time, and the packet is delayed until this is finished. PAHE inserts a self-routing label (12 bits) in front of the packet header, and

the packet finds its path to the correct output queue according to the IP destination address and traffic class indication. The self-routing label also includes the IQ number (12 bits) used by OQRL to update rate counters for each neighbor node, and to route ScM.

When a packet arrives, the IQID (Input Queue In Drop) process checks the IQ length, and if it has reached a certain figure, and the packet priority (given by view and sub-object importance) permits dropping, the packet is dropped. The probability for dropping here should be very small compared to OQID, see below. At this time there are a number of buffers left for high-priority packets. It is of course a very low but non-zero probability that many high-priority packets arrive in sequence, so dropping cannot be entirely avoided.

When a packet has moved to the head of the line of IQ, it may experience that there are no output buffers available. The IQOD (Input Queue Out Drop) process drops low-priority packets, but just waits if the packet has high priority. (It would be more efficient if the queue were searched for packets that can find an idle output buffer). The probability for dropping here should be very small compared to OQID.

The OQID (Output Queue In Drop) process is needed when a number of inputs try to overload the same output. The packet priority is checked, and the packet is dropped if it has low priority. The counters CntUP(p) are updated for arrivals from each neighbor node p, indicated by the self-routing label. The number of packets dropped are counted using counters $z_i(j)$ and for the total $z_{iTOT}(j)$.

When the network node is a FireNode, the OQID has to know the traffic contribution from all users on all Firewire buses. The SSRC parameter of the RTP* packet header uniquely identifies the object of a scene. The packet arrival rate from each sub-object is then registered by means of the value of the counters CntUP(SSRC) during a time interval. The self-routing label indicates which Firewire bus p the packets arrive from. The counters are stored in a two level index table, IT(p, SSRC).

The Play-out Process, POUT, fetches packets from the OQ in a FIFO order, releases the buffer, and sends the packets out serially.

The Out Queue RL Process, OQRL, implements the RL-C Resolution & Traffic Control Algorithm, which requires that scene objects are coded according to the 2D interlace scheme as illustrated in Figure 6. Each sub-stream can be used to compose a complete picture, but with low spatial resolution. If all streams are received without loss the maximum spatial resolution is obtained. By giving at least one sub-stream high priority, packets from that stream will not be dropped. If packets from one of the other sub-streams are dropped, the missing pixels can be found by interpolation. The temporal resolution of the object is also illustrated with the time axis in Figure 6. Note that the temporal resolution can vary from object to object.

The RL-C algorithm is handled by OQRL and Traffic Source. The RL-C algorithm is described in detail in 4.4.

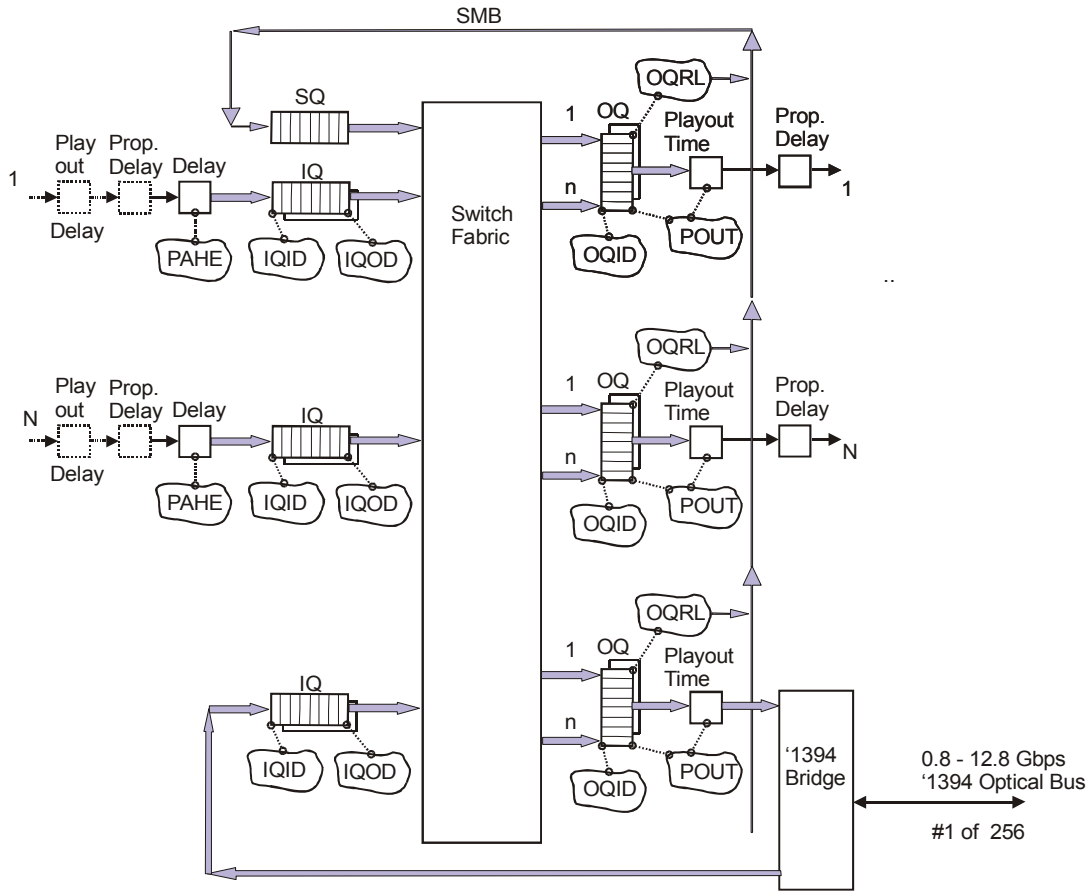

Figure 20. A packet switch architecture with input and output queues.

7 Simulations

Real-time Simulator for the RL-C Resolution & Traffic Control Algorithm
The RL-C algorithm adapts temporal and spatial video resolution to the traffic load in IP networks. Then end-to-end delay can be guaranteed. To investigate how human beings perceive this time-varying resolution, a real-time simulator of a possible IP network applying the RL-C algorithm, and with realistic multimedia traffic, shall be developed.

The simulation model should leave out details that are not important for our case. Aggregated traffic sources can be designed to substitute parts of the network adjacent to the routes and nodes we are studying. The modeling will be based on futuristic scenarios for the traffic. The simulator shall scale the traffic generator outputs according to the RL-C algorithm. In the testes as described in Ch 8, high-resolution, stored 3D videos, will be played out and scaled according to traffic load. The simulations do not have to be 'real-time', but can off-line produce a sequence of ScaleMessages that controls the multimedia quality level when played out from a server. The other approach is to let the sequence of RL ScaleMessages be produced in real time.

8 Testing time-varying resolution of scenes

Assessment of the RL Resolution & Traffic Control Algorithm
The RL-C algorithm adapts temporal and spatial video resolution to the traffic load in IP networks. Then end-to-end delay can be guaranteed. To investigate how human beings perceive this time-varying resolution, a complete visual/psychological test shall be developed.

Based on existing futuristic scenarios and simulations (see Ch 7), video resolution variations can be estimated. From this, test scenes can be shot using high-resolution 3D cameras, and after editing (Adobe tools), the test videos can be projected on a high-resolution 3D screen for test persons. Before display, the resolution will be scaled in real-time from input as described in Ch 7. It is important that the viewer's experience is as close as possible to a real situation. The first tests will apply single-view, stereoscopic display (polarized pictures and glasses). Later, multi-view, auto-stereoscopic techniques (e.g., lenticular) will be considered. The tests shall be based on refinements of methods developed by ITU, DVB, ISI/IEC, MPEG, and others. See (PER02) for description of video testing.

Video object resolution controlled by eye tracking
Eye tracking can be used to sort out which objects in a scene viewers focus on. The resolution of those objects should then be sent with higher resolution than less important objects. The resolution parameters are important for resolution and traffic control in RL-C packet networks. Studies of viewer focus on objects in film and TV content are also highly relevant (ELV04). Known eye tracking theories shall be described (XXXzz). Practical tests measuring the variation of resolution and video traffic, using existing tracking software (YYYzz), shall be planned and carried out. A traffic model for the video generation shall be proposed.

References

(ATH01) S. Athuraliya et. al., REM: Active Queue Management, *IEEE Network, 15*, Issue: 3, May-June 2001.

(BJO03) Thomas Wiegand, Gary J. Sullivan, Gisle Bjontegaard, and Ajay Luthra: Overview of the H.264/AVC Video Coding Standard, *IEEE Transactions on Circuits and Systems for Video Technology*, July 2003.

(CHU99) Shang-Tse Chuang, Ashish Goel, Nick McKeown, Balaji Prabhakar Matching Output Queueing with a Combined Input Output Queued Switch.. Stanford University. Infocom '99, New York, USA.

(CIS04) http://www.cisco.com/en/US/products/sw/iosswrel/ps5207/products_call_flows09186a00801c5d23.html

(CRO04) M Croll, P Bohler, L Haglund. European evolution toward HD. IBC Conference Publication, September 2004.

(CYC02) H.L.Cycon, D. Marpe. High-Performance Wavelet-based Video Coding Using Variable Block- Size Motion Compensation and Adaptive Arithmetic Coding, , Proc. 4th IASTED International Conference on Signal and Image Processing (SIP'02) Kawai, USA, ACTA-Press, pp. 404-409, Aug. 2002.

(DAG04) Dagens Næringsliv, 18sep04

(DUF02) Jean-Claude Duford. BIFS: Scene Description. In The MPEG-4 Book, edited by F Pereira and T Ebrahimi. Prentice Hall 2002.

(DUM02) Vincent Dumas, Fabrice Guillemin, and Philippe Robert
A Markovian analysis of additive-increase multiplicative-decrease algorithms
Source: **Adv. in Appl. Probab.** 34 (2002), no. 1, 85–111

(ELV04) J M Elvemo. Xxxxxxx, 2004

(ENV04) 4Caster MPEG-4 encoder series, http://www.envivio.com/products/etv/register.html

(FLO97) Floyd, V. Jacobson, Random Early Detection gateways for congestion avoidance, *IEEE ACM Transactions on Networking, 1*(4), August 1997.

(HAN03) Handley, M., Floyd, S., Pahdye, J., and Widmer, J.
RFC 3448, Proposed Standard, January 2003.
TCP Friendly Rate Control (TFRC): Protocol Specification.

(HEL03) T E Helgesen, M Karpinski, Videoconference on a set-top box. Master thesis, ITEM, NTNU 2003.

(HHI04) Web page, Fraunhofer HHI, 23004

(HOL01) C. V. Hollot, et. al., On Designing Improved Controllers for AQM Routers Supporting TCP Flows, *Proc. of IEEE Infocom,* 2001.

(HUR01) P Hurley e.al, in IEEE Network, vol 15, issue 3, May-June 2001.

(ITU93)
http://www.itu.int/rec/recommendation.asp?type=products&parent=T-REC-h

(ITU003a)
http://www.itu.int/rec/recommendation.asp?type=products&parent=T-REC-h

(ITU03b)
http://www.itu.int/rec/recommendation.asp?type=products&parent=T-REC-h

(ITU03c)
http://www.itu.int/rec/recommendation.asp?type=products&parent=T-REC-h

(JAI99) S. Jaisimha, EP Ong, WC Wong, T. Miki and S. Hotani
Improving The Robustness of The MPEG4 Wavelet Transform Image Coder. Department of Electrical Engineering, National University of Singapore. Microelctronic Systems Application Division, Institute of Microelectronics, Science Park, Singapore. Multimedia Signal Processing Laboratory, NTT DoCoMo, Kanagawa, Japan, 1999.

(JMF04) Java Media Framework,
http://java.sun.com/products/java-media/jmf/index.jsp
http://www.mindspring.com/~lindend/corejmf/

(KAR87) M Karol, M Hlychyj, S Morgan. Input versus output queuing on a space-division switch. IEEE Transaction of Communications, vol. 35, pp. 1347-1356, Dec 1987.

(KAT91) Steven Katz. Film directing shot by shot. Visualizing from concept to screen. Michael Wiese Productions, 1991.

(KAT02) D Katabi, M Handley, C Rohrs. Congestion Control for High Bandwidth-delay Product Networks. SIGCOMM'02, Pittsburg, Pennsylvania, August 19-23, 2002

(KAW02) M Kawakita, T Kurita, S Inoue. HDTV AXI Camera. IBC Conference Publication, Amsterdam 2002.

(KIM02) M Kim and S Wood. Extensible MPEG-4 Textual Format. In The MPEG-4 Book, edited by F Pereira and T Ebrahimi. Prentice Hall 2002.

(KIM03) K. B. Kim, S. H. Low, Analysis and Design of AQM based on State-Space Models for Stabilizing TCP, *Proc. of the*

American Control Conference 2003, Volume: 1, 4-6 June 2003.

(KOE02) R Koenen. MPEG-4 Overview. ISO N4668, 2002.

(KOH04) Eddie Kohler, Mark Handley, Sally Floyd, Jitendra Padhye Datagram Congestion Control Protocol (DCCP) ICIR

(KOK03) R Köker, S Çakar, C Öz: "Moving Object Detection and Target Prediction in Video Image Using Computer Vision".IJCI Proceedings of International Conference on Signal Processing, Vol. 1, No. 2, September 2003.

(LIE03) Lie, L. A. Rønningen, Distributed Multimedia Plays with QoS guaranties over IP, *Proc. of IEEE Wedelmusic'03,* 14–17 Sept., Leeds UK, 2003.

(LIE04a) A. Lie, O. M. Aamo, L. A. Rønningen, On the use of classical control system based AQM for rate adaptive streaming media, *Proc. of 17th Nordic Teletraffic Seminar,* Norway, August 2004.

(LIE04b) Arne Lie, Aamo, L A Rønningen. OPTIMIZATION OF ACTIVE QUEUE MANAGEMENT BASED ON PROPORTIONAL CONTROL SYSTEM. CIIT, Virgin Islands, Nov 2004.

(LIT61) J D C Little. A proof of the queueing formula L = λ W. Operations research, 9, 1961.

(MAR02a) D Marpe, H L Cycon: "High-Performance Wavelet-Based Video Coding Using Variable Block-Size Motion Compensation and Adaptive Arithmetic Coding". www.iasted.org/conferences/2002/hawaii/program-358f.htm

(MAR02b) D. Marpe, H.L.Cycon, T.Wiegand. Design of a Highly Efficient Wavelet-Based Video Coding Scheme, , Proc. SPIE, Vol. 4671, pp. 1133-1141, Jan. 2002.

(MIL01) Gerald Millerson. Video Production Handbook, Focal Press, 2001, Third edition.

(MIC04) The coding standard VC-1 for HD, a competitor to H.264, developed by Microsoft. Proposed as an SMTP standard.

(MIT04) www.merl/projects/Video-segmentation

(MIT04) Mitsubishi home page

(PER02) F Pereira and T Ebrahimi. The MPEG-4 Book. Prentice Hall PTR, 2002.

(QUI04) QuickTime, http://www.apple.com/quicktime/download/

(RAM01) K. Ramakrishnan, S. Floyd, D. Black, The Addition of

Explicit Congestion Notification (ECN) to IP, *IETF RFC3168,* September 2001.

(RED04) http://www.icir.org/floyd/red.html

(RON82) L A Rønningen. Input Traffic Shaping, Proceedings of the International Teletraffic Congress, Montreal 1982.

(RON99) L A Rønningen. The Combined Digital Satellite Broadcast and Internet System. Telenor Satellite Services, February 1999.

(RON02a) L A Rønningen and A Lie. 'Performance control of high-capacity IP networks for colloborative virtual environments'. Paper, IBC 2002, Amsterdam Sept. 2002.

(RON02b) L A Rønningen and A Lie. 'Transient behaviour of an adaptive traffic control scheme'. Paper, EUNICE '2002/IFIP WG6.7 Workshop, Trondheim Sept 2002.

(RON04a) L A Rønningen, Adaptive Audio-Visual Resolution Control and Traffic Control in Packet Networks, Stanford University, May 2004

(RON04b) L A Rønningen. Natural seeing. (to be published)

(RON04c) Observations made at IBC 2004.

(RTP96) RTP. A Transport Protocol for Real-Time Applications, RFC1889

(RTS98) RTSP – Real-Time Streaming Protocol, RFC2326, 1998

(SCH02) A Schmidt, A Grasnick. Multi-viewpoint Autostereoscopic Display from 4D-Vision. Proceedings of SPIE Vol 4660, 2002

(BER97) C van Berkel, J Clarke, Characterisation and optimisatin of 3D-LCD module design. Proceedings of SPIE vol 3012, 1997.

(SDP02) http://standards.ericsson.net/sip/rfcs/rfc3266.txt

(SHI03) Joseph Shinar. Organic Light-Emiting Devices. Edited by. Springer Verlag, 2003.

(SIP02a) SIP: Session Initiation Protocol, IETF, RFC3261, 2002

(SIP02b) H.323 versus SIP: A Comparison, 2002. http://www.packetizer.com/iptel/h323_vs_sip/

(STR02) Ola Strand. Distribuert multimedia samhandling. Master thesis, NTNU, June 2002.

(VAE03) Kathleen M. Vaeth. OLED-Display Technology. in Information Display, Vol 19, No. 6, pages 12-17; June 2003

(WAN00) Y Wang, J F Doherty, R E Van Dyck: "Moving Object Tracking in Video". Dept of Electrical Engineering, The Pennsylvania State University/ National Institute of Standards and Technology, Gaithersburg, USA. 2000
(VRM97) ISO/IEC 14772-1:1997. The Virtual Reality Modeling Language, www.vrml.org/Specifications/VRML97

(WOE95) M Woepking:" Viewing comfort with stereoscopic pictures: An ebyxperimental study on the subjective effects of disparity magnitude and depth of focus". Journal of the SID, pp. 101-103, 1995.

(WEB04) Webster E Howard. Better Displays with Organic Films. Scientific Americal, Feb2004

(XIN00) G Xing, J Li, S Li, Y Zhang: "Arbitrarily Shaped Video Object Coding by Wavelet. Institute of Mathematical Science, Peking University, Beijing. Microsoft Research China, Beijing. ISCAS 2000, Genova, Switzerland 2000.

(XXX04) Tests of distributed playing of jazz music, (TBD)

(XXXzz) Eye tracking (TBD)

(YYYzz) tracking software (TBD)