Programme ANR VERSO

Projet VIPEER

Ingénierie du trafic vidéo en intradomaine basée sur les paradigmes du Pair à Pair

Décision n° 2009 VERSO 014 01 à 06

du 22 décembre 2009

T0 administratif = 15 Novembre 2009

T0 technique = 1er Janvier 2010

Livrable 3.1

Content adaptation use cases and requirements
Architectural specifications of the content adaptation engine

Auteurs :

Pierrick PHILIPPE (Orange Labs)

Jean KYPREOS (Envivio)

Juillet 2010

Telecom Bretagne ; Eurecom ; INRIA ; France Telecom ; NDS ; ENVIVIO

Table des matières

Table des figures

Overview

Multiple devices

Devices have been considerably evolved for last years with the multiplication of new services to the customers. This has been done thanks to many factors like the video quality progress with more efficient media compression, the introduction of high definition services, the possibility for handsets to receive television services through multiple networks such as DVB-H, 3G, internet, Wireless, and so on.

In addition this evolution has been reinforced with the convergence of multiple services over internet and the delivery of new protocols.

This change is really noticeable for handsets devices with many resolutions varying from small resolution like “QCIF” to high definition, and with many intermediate sizes as tradeoffs for resolution and CPU power.

Video transport in the network

Increasing consumption of video data through many services like television broadcast over IP, user generated content, Video on demand, catch-up TV…, leads to a huge explosion of the network traffic.

Distribution of the video becomes really a challenge for operators to ensure a satisfying QoS to any customers, to feed heterogeneous terminal devices within heterogeneous infrastructures. Video data deliveries need also specific adaptation mechanisms to avoid traffic congestion or traffic burdening. Current adaptation techniques are described here after.

Transcoding in the network

When positioned at strategic nodes within the network, transcoding allows adapting a primary video format to dedicated resolutions or codec type in order to feed heterogeneous terminal devices or to support new video format either for storage costs reason or to face an unpredictable service request. This process has the advantage to avoid multiple unused encodings conveyed in the global network and therefore naturally reduces the bandwidth traffic.

In addition transcoding can be used for adaptation between heterogeneous networks, for example handsets devices may connect to internet via 3G or via a wireless connection.

HTTP streaming

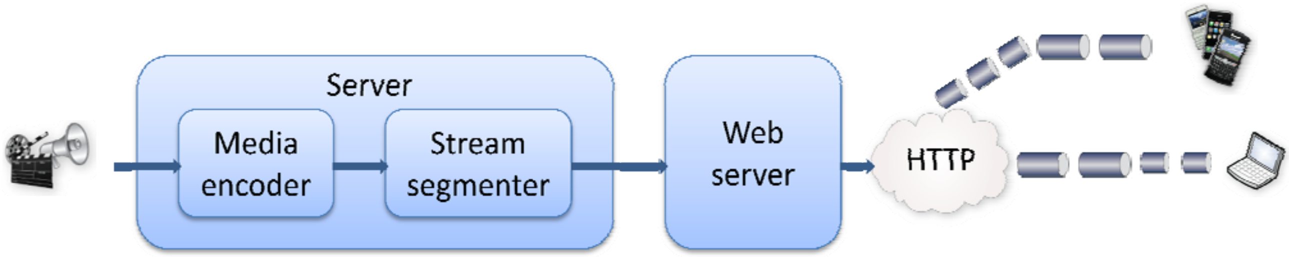
Traditional streaming relies upon a streaming session between the server and client which ends only when the user disconnects from the server (“stateful” protocol).

HTTP streaming is a hybrid media delivery method as it acts like streaming but it is based on HTTP progressive download. If a client requests some data, the server will send it but will not remember the client or its state (“stateless” protocol). Each HTTP request is handled as a completely standalone one-time session.

HTTP streaming is an adaptive media delivery which performs the media download as a long series of small progressive downloads.
The video/audio source is cut into small segments (“chunks” typically 2-10s long) and encoded to the desired delivery format according to different bit rates and resolutions: consequently each chunk is independent from the previous one and has a variable size.

The encoded chunks are hosted on a HTTP web server and are delivered on a linear and seamless fashion to the client requests. Chunks are encapsulated in a specific format before transport to the client, the main formats being “fragmented mp4” from Microsoft (smooth streaming) or “chunk TS” from Apple (adaptive streaming).
Note there is a concurrent technology from Flash which uses a proprietary protocol (RTMP) instead of HTTP.

The selection of the most appropriate bitrate/resolution depends then on the network conditions and on the user CPU device. The device player receives chunks, computes the cumulative playtime, and according to various strategies (for example temporal smoothing) requests the next bit rate to the web server. A new chunk at lower or higher bitrate is then delivered by the web server and the process reiterates.



Content adaptation in the distribution chain

Current practice

The figure below represents typical distribution architecture as operated today.

A video content, either stored or resulting from a live event, is first encoded. Subsequently, it is made available to the end-users through a CDN (Content Delivery Network). BAS servers feed the last mile, presented here in the particular case of a DSL framework, before reaching the end users.

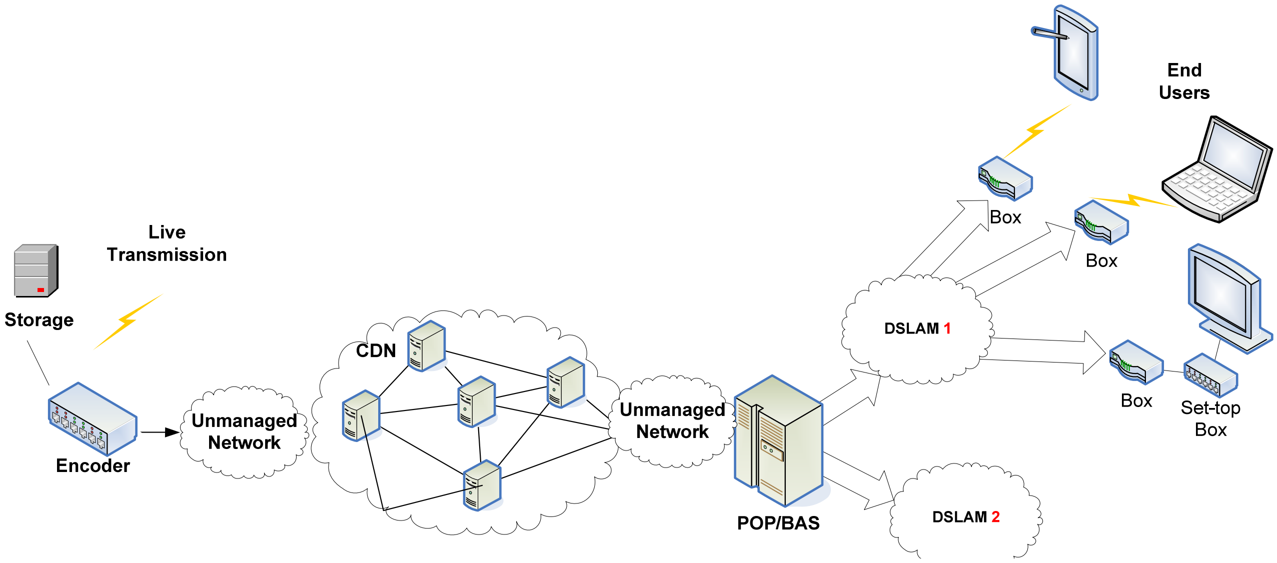


Figure 1: Typical Content Delivery System architecture

End users are connected with modem boxes to the network that feed their multimedia devices, wired or not. The multimedia devices can be of different kinds: connected devices such as a television connected to a set-top-box or a computer, or remote devices such as Mobile Phones, Portable Multimedia Players (PMP) or laptops wirelessly linked to the box in Wi-Fi.

In this schema, the only adaptation capability is offered through the encoding operation at the emission stage (the video bit rate can be changed) and the provisioning of caching capability in the CDN. Hence, using this typical distribution scheme, the adaptation capability is hardly able to anticipate audience variations and especially audience peaks. At least three possibilities are possible:

- the video bit rate is reduced resulting in a lowered quality of experience
- the CDN capability is over sized so as to face with the worst anticipated audience scenario, but this is costly considering CDN pricing practice
- the video distribution is cached and distributed at different rates, enabling bit rate switching in case of congestions: this duplication of streams has an impact on the caching capability and required bandwidth in the distribution network,

None of those solutions are effective in terms of price and efficiency, especially as far as the Quality of Experience is considered. Hence the alternative proposed here to enforce adaptation elements in the distribution scheme.

CDN connection

The content delivery network is a meshed structure. In this work package, we assume it is directly connected to some network elements, and operates with them. For example the CDN can be connected to the backbone network via some peering point or mesh servers present at different locations, such as in a point of presence (POP), broadband access server (BAS) or DSLAM. It can also use the caching ability of modem boxes, as studied in this project (WP4).

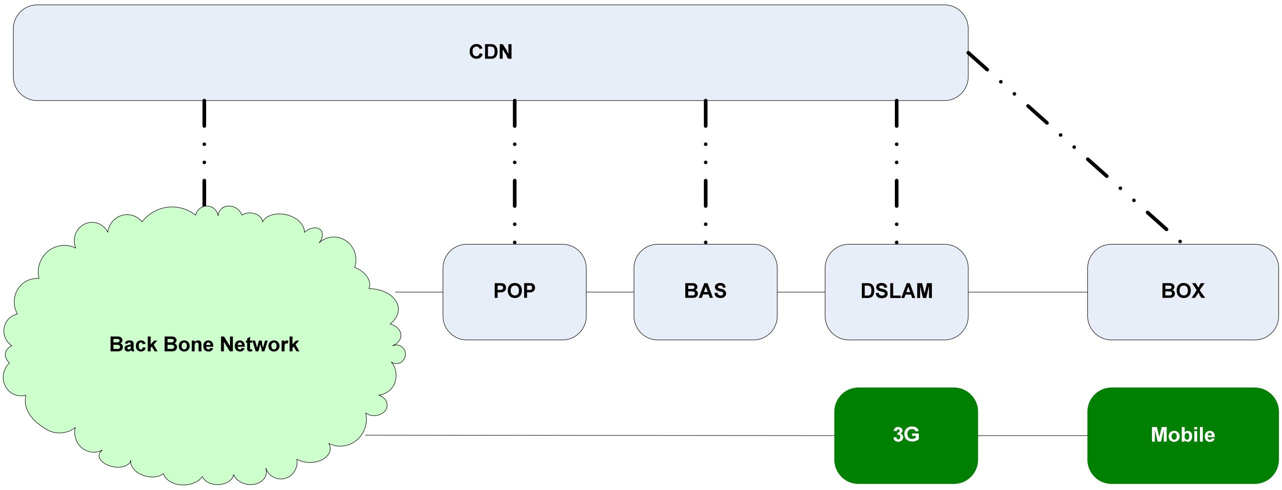


Figure 2: Simplified delivery architecture

Proposed adaptation architecture

Components elements

So as to be able to answer to highly variable audience scenarios, the following adaptation component is proposed for insertion in a legacy delivery network.

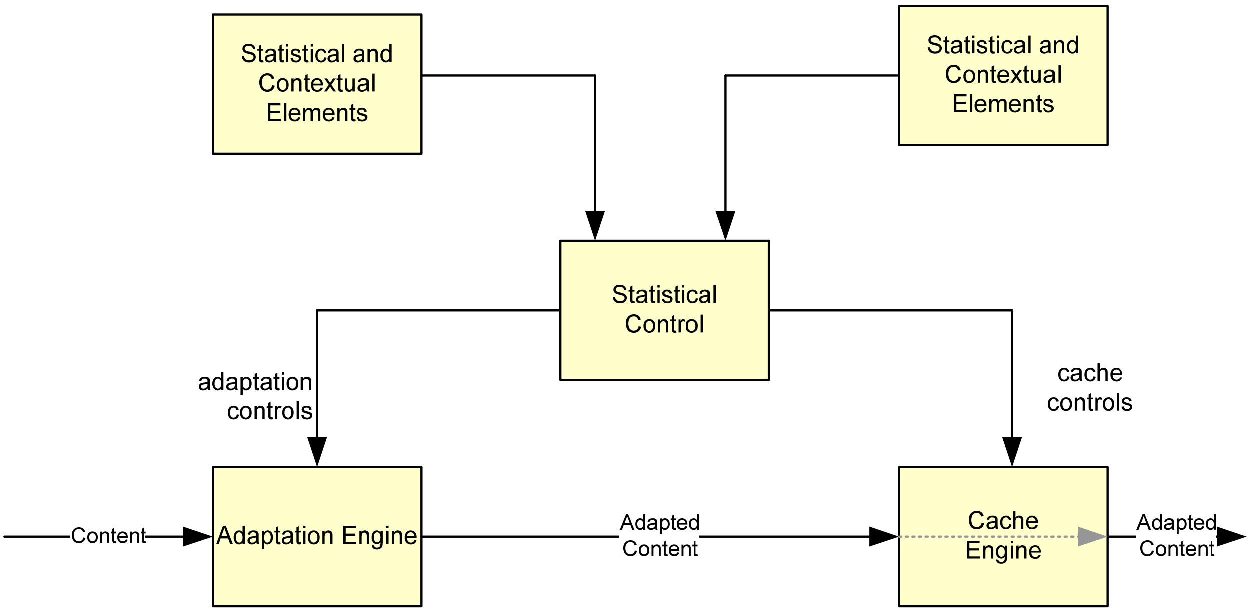


Figure 3: Adaptation Component

This adaptation component aims at operating on the cache engine and at adapting the multimedia content, thus enabling on the fly stream caching and adaptation. The adaptation is controlled by means of statistical and contextual components disseminated in the content delivery system.

Therefore, the adaptation elements building blocks are the following: statistical and contextual elements, statistical control, adaptation engine and cache engine. They are reviewed in the following paragraphs.

Statistical and Contextual Elements (SCE)

The role of the Statistical and Contextual Elements (SCE) is to collect static as well as dynamic information at different Content Delivery Systems (CDS) locations. Examples of relevant dynamic information are the current audience, the bandwidth usage, the cache occupancy etc. They are completed with static information such as the type of supported codecs, the network maximum bandwidth, the fact that a user is eligible to a premium offer etc.

SCE are knowledge elements having a potential influence on the adaptation. They are collected in numerous locations, emanating from the user or automatically collected in the network.

Statistical control (SC)

Based on collected SCE a Statistical Control (SC) pilots the two engines: the adaptation and caching capability through dedicated control signals. A SC maybe connected to multiple SCE emanating from different network locations and gathers the delivered information so as to decide on the potential actions required at the cache and adaptation level.

Adaptation Engine (AE)

The Adaptation Engine (AE) transforms the video content. Examples of transformation can be transcoding, i.e. a change in the codec kind, bit rate or resolution, but also the modification of random access point (e.g. to enable fast content browsing) and digital right management. Different transformation can be operated during an adaptation stage, such as the cascade of DRM decrypting, transcoding and finally re-application of the DRM. The adaptation controls pilot those transformation and their organization.

Note that an adaptation engine potentially requires a lot of processing resources; if transcoding is considered the adaptation element may require intensive computational capacity. It is also worth mentioning that if DRM is considered then it may be required for IPR (Intellectual Property Rights) reasons to adapt the DRM in private areas. For those reasons, the adaptation engine cannot be inserted everywhere in the content delivery system.

Cache Engine (CE)

The Cache Engine has the storage capability in order to replicate the (possibly adapted) incoming content. Hence the content will be made available closer to the end-user. Caching requires storage capacity and may consequently be a costly element that cannot be spread in the content delivery system.

Adaptation Architecture

The proposed adaptation architecture is therefore flexible: it can be simplified for example when no caching capability is needed and if the context emanates from a single source. Alternate sources of simplification come from the capability of the CDS elements: for example a modem box has a limited computational power and is probably not able to perform a transcoding operation.

Therefore, these limitations and simplifications have an impact on the overall Content Delivery System architecture when those adaptation elements are inserted. An illustrative architecture of the CDS with adaptation capability is presented figure 4.

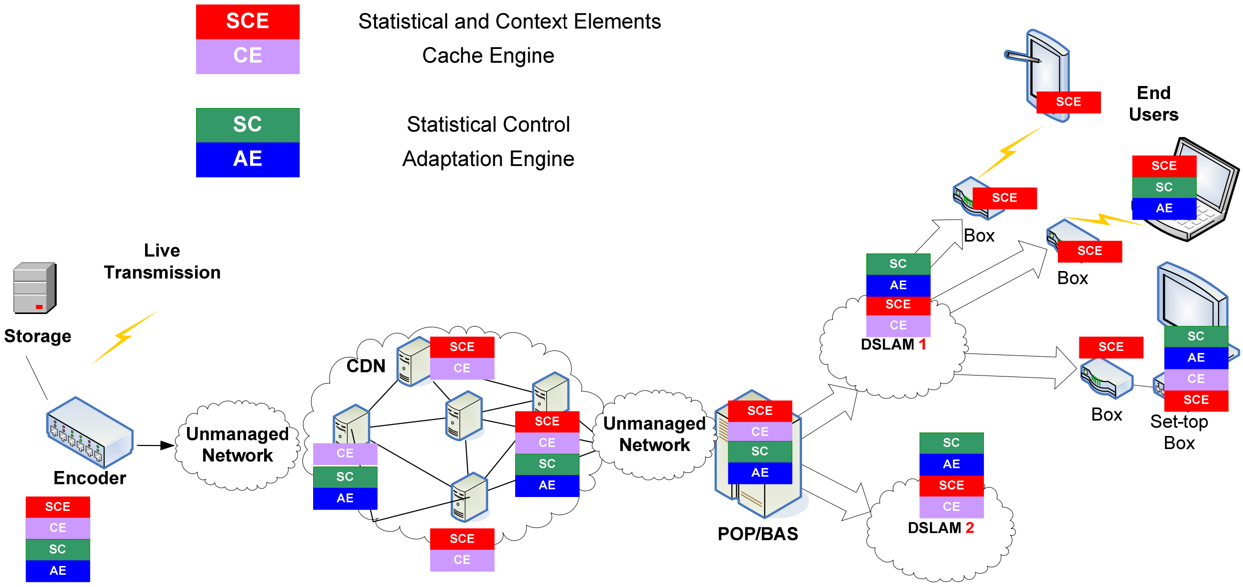


Figure 4: Content Delivery System with adaptation capability

Statistical and Contextual Elements (SCE) can be considered as lightweight elements, since they are passive elements sampling information. Hence they can be placed at numerous locations in the CDS. We consider inserting those elements at the following locations:

- Content production, e.g. where a content popularity could be measured
- CDN, where a CDN work load and cache occupancy can be reported at different places,
- DSLAM and POP/BAS Server, in order to report the QoS at the server level
- Modem/Box, where the QoS can also be reported depending on the user activity.
- Set-top box, its capability can be reported, for example in terms of codec support
- Device, information concerning the user can be collected at this level (Pay Per View, etc.), also device bandwidth capability or image resolution can be informed.

At this stage, many questions arise considering the location of these elements in the distribution chain. The main question is the added value of these components and their cost in a realistic environment. These aspects need to be carefully evaluated to as to appropriately disseminate the adaptation elements in the content delivery system.

Another point need to be considered: are all these elements accessible? This highly depends on the envisioned architecture and development of this project. For example, it may be valuable to insert adaptation elements in the CDN but this is likely impossible if it is operated by a private organization. Caching capability could be inserted in modem boxes in order to benefit from dCDN approaches, but this capability need to be investigated to assess whether this is economically feasible or not.

This architectural considerations need to be evaluated in the light of different application scenarios, highlighting the weaknesses of the current Content Delivery Systems and the expected benefit of adaptation capability insertion. Next session deals with this by considering some realistic use cases having different impact on the underlying distribution architecture.

Use cases study

Introduction

In this section we study some simple application scenarios. The objective here is to illustrate the key adaptation elements utility and behaviour for improving the media distribution chain. Here we mainly focus on the Quality of Experience (QoE) improvement. The adaptation elements role and their possible locations in the distribution chain will be considered in the light of the studied use cases. This will clarify the possible architecture for the content delivery systems and impact WP1.

A particular focus is made on the expected benefits with respect to the current practices as operated in legacy distribution systems. The following axes will be considered:

- QoE (example congestions are minimized)
- number of potential clients (extend the number of supported devices)
- infrastructure cost (avoid oversized / underused infrastructure)

Three simple scenarios are considered in this section.

- "World Cup" (Highly Popular event)
- Movie VoD and Catch-Up, which are similar in terms of network elements
- User Generated Content (UGC)

These scenarios share the same video set up as presented in the next section.

Video distribution setup

In this study, we target a rather high quality service in which video is distributed mainly on wired devices. Typically we target HDTV services for the top quality devices maximizing the quality. Lower rates are made available in case of network congestion or unavailable bandwidth.

Examples are

- 8 Mbps higher quality (top quality)
- 4 Mbps medium quality
- 2 Mbps lower quality (lower quality)

Additional levels could be inserted such as to accommodate in case of limited bandwidths capability for example.

Popular live event: world cup football match

The first scenario investigated here studies the transmission of a live popular even, such as a football match for a homogeneous set of TV devices. It is assumed that the underlying network is unmanaged, which means that the network capacity is not manageable and that no traffic engineering is possible.

The fact that we deal with a live event means that the caching capability of the network will be reduced at its minimum so as to avoid transmission delays. It is also very probable that the content is protected by a DRM scheme.

As a consequence, due to the huge audience targeted with this program, sufficient provisioning is required in the network capability and encoding set-up: the required bandwidth is likely to reach high peaks during the event. Figure 5 illustrates the audience evolution and the bit rate adaptation required to adapt to the limited capacity of the network. In order to face with the demand, the requested bit rate will be lowered, and switching toward the lower rates during the maximum audience period will be required in order to face with the multiple bottlenecks in the network.

This effect can particularly be noticed in a mobile environment, in which an access point will be solicited by many portable devices during the live event: the congestion point will likely arise at the 3G cell level.

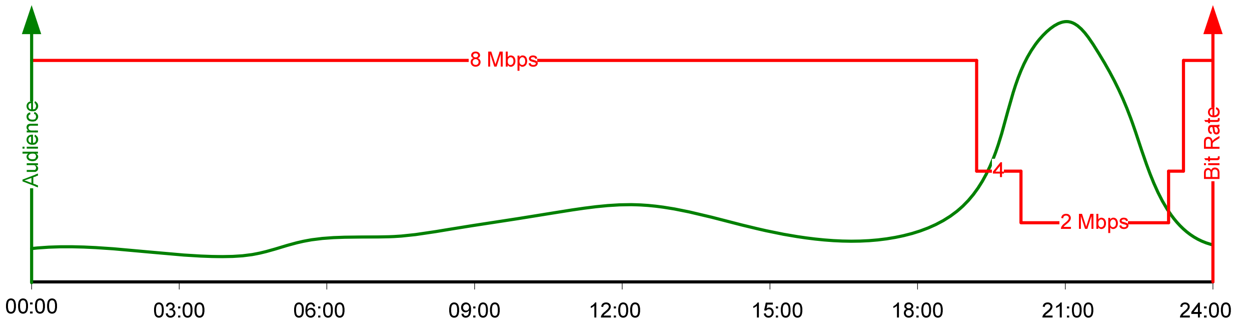


Figure 5: Illustrative audience and rate adaptation

We notice that the network is under used most of the time and that the different bit rates are rarely used at the same time: the provisioned network and encoding resources are underused, there are wasted most of the time. We see here the interest in adapting the content in the network, especially upstream from the congestion points.

Therefore a proposed solution would be to insert the adaptation elements proposed above in order to optimize the media distribution. To face with congestions at the POP/BAS/DSLAM server transcoding tools could be inserted at those steps in order to reduce the bit rate when required by the amount of audience. As such the bit rate can be locally adapted to the local congestions the quality reduction would not affect all network points. If it happens that most AE perform exactly the same operation, then the adaptation could be shifted to a higher point in the network, e.g. at the head end server: the initial encoding bit rate being changed as it is shown on figure 6.

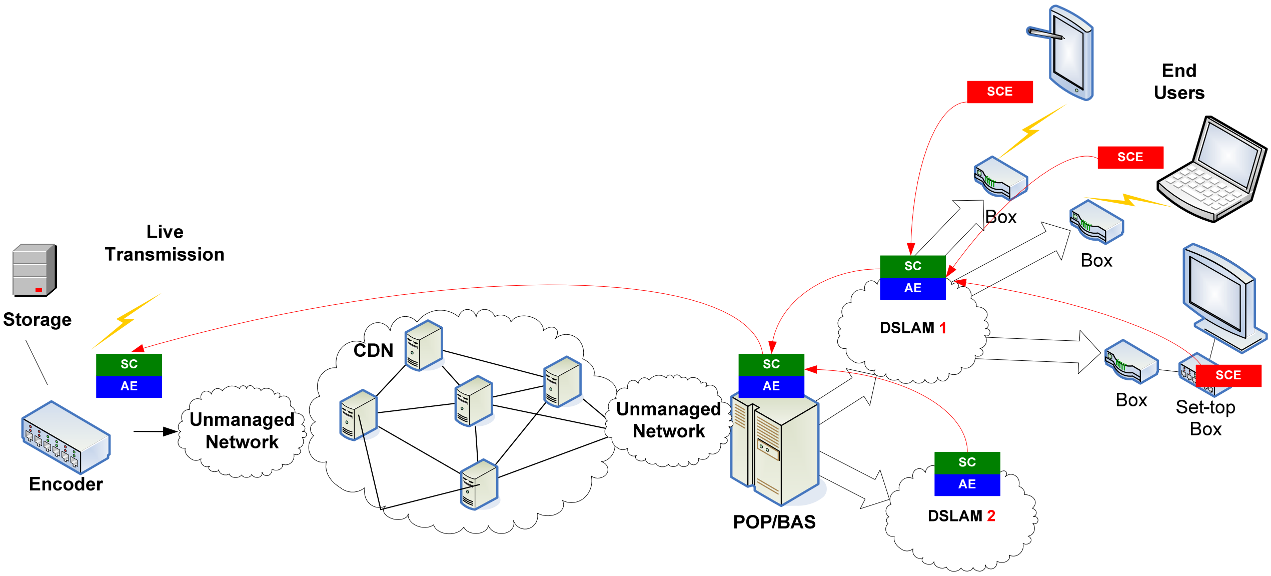


Figure 6: Live event adaptation scenario

All this points, identified from the adaptation point of view need to be confronted with the network analysis in order to measure the benefit of the adaptation points. They also need to be confirmed, from an architectural point of view by the experts from this project. Particularly, we raise the following points to be solved jointly with the WP1 experts:

- How the benefit of adaptation can be evaluated and quantified?
  - e.g. avoid multiple caching, save bandwidth, ensure unexpected network congestions
- Where adaptation could take place?
  - Should we concentrate the adaptation elements or spread them as much as possible at different locations?
  - How do these elements communicate?

Additionally, we raise some questions to be addressed by this work package:

- How to handle adaptation elements
  - How to handle DRM adaptation?
  - How to limit the delay overhead (A 30s or 1mn delay can be a drawback for live events).

- How to avoid unnecessary transcoding, implying quality losses and reducing the quality of experience (QoE).

Video on Demand

In this use case the Video on Demand (VoD) scenario is investigated. It assumed that a catalogue is proposed to clients that would have to pay in order to access to the program as a consequence a high level of QoE is expected. It is also assumed that many different devices can potentially have access to the proposed programs.

A particular case of VoD is the catch-up TV service: it allows the users to choose the program or the TV show they want to watch from an archive of programs broadcasted usually shortly before. The audience generated by the catch up TV is far different from the VoD, but the technical elements are quite similar.

The current practice under these assumptions would be to prepare the content in multiple formats and store all the versions in the CDN, such as to face with an unexpected consumption of the program. Multiple encoding comes at a price and caching numerous formats in the CDN has also impact on the infrastructure cost. As a consequence, there are inherent restrictions in the number of devices supported by these systems.

Content adaptation can be a solution in order to rationalize the number of supported devices. One can imagine caching an initial format and add adaptation capability to the CDN in order to recode the content in a requested format. This new format being temporarily cached, expecting that this format will be sufficiently popular. User types could also be differentiated whether the content they wish to access to: e.g. a different pricing could be done for HD and SD program.

The adaptation envisioned for this use case is summarized on figure 7.

Caching and adaptation capability have been inserted in the CDN which is informed, based on contextual information emanating from the users, and able to process the audiovisual content.

For this use case different aspects need to be clarified and confirmed by experts from the different work packages.

How the benefit of the proposed adaptation scheme can be evaluated or quantified? The reduction of encoding formats is a plus, and the ability of handling any device type (even obsolete or emerging devices can be supported) can be counterbalanced by the adaptation cost.

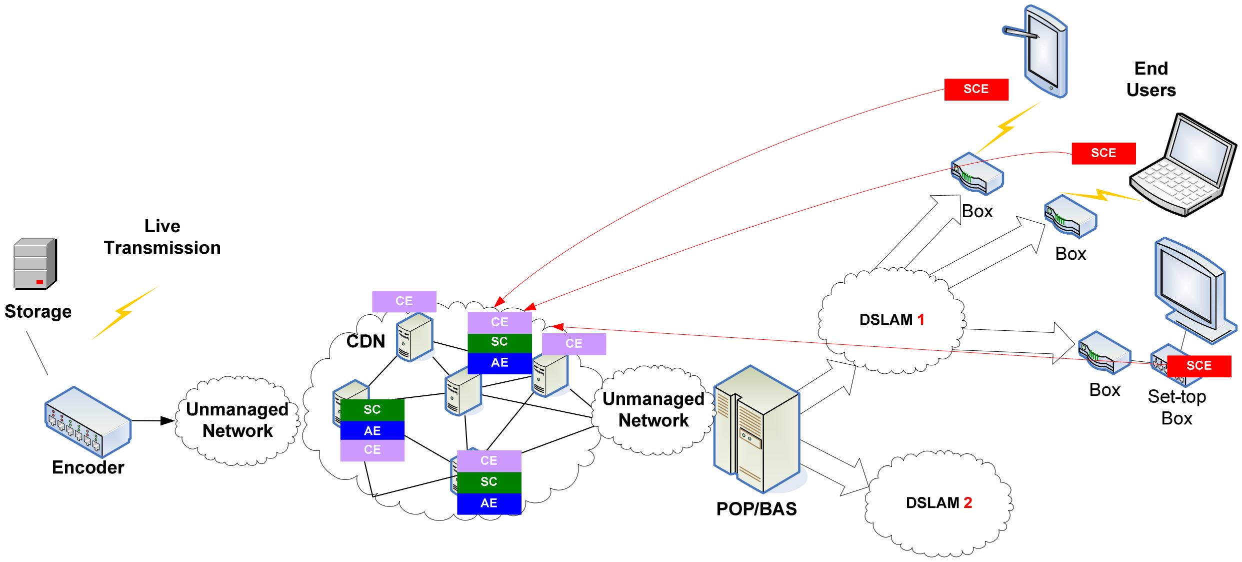


Figure 7: VoD adaptation scenario

In this scenario as well arise the question of the adaptation optimal position in the delivery chain. For example, the CDN looks like a promising place, but the question of accessibility is present. The question of DRM is important too, since VoD content is accompanied with some protection scheme, caching cannot be done in open areas.

The communication between the adaptation elements is also to be considered notably in WP1.

Clarifications are needed for the development inside this work package (WP3) as well: typically the way to handle DRM adaptation need clarification. Also on the fly transcoding may impact on the QoE. Adaptation provisioning needs also to be considered in order to quantify the density of these elements e.g. in the CDN.

User Generated Content

Third scenario investigated in this report, a User Generated Content (UGC) is made available from a user. It is assumed that this content may become extremely popular. We take as an example a funny private sequence eligible to a "Video Gag" kind television programme or a breaking news event, for which a short movie was produced by an individual. The content therefore becomes public, no DRM being attached on this item, and is supposed to be spread through an unmanaged network in a distributed fashion.

The current practice would be to send the video to a centralized service such a Dailymotion or YouTube, based on a private CDN that spread the content encoded for a limited number of devices.

Referring to figure 8, in a first phase, it is supposed that the content is transferred from the recording device to a user personal storage device connected to the modem. This storage device could be e.g. user's set-top box directly connected, through an Ethernet connection, to the Internet. We assume that the set-top box has a programmable coding capability such as to be able to adapt the content if required. The content can also be encoded directly on the user's computer if this device is available at user's end.

Once the content is available at the user's end, the user posts a description and a link corresponding to the generated content on a website. Interested people may access to the content. If a particular format is required, the coding capability is able to deliver the original content in an alternative format. This alternative format is potentially cached at the original user side or at the new user end, so as to be made available to others.

Depending on the popularity success, some of the cached formats may be duplicated in a higher network point, such as a CDN for a wider audience support. The CDN would not be used if the content is of minor importance, which is different from the current practice in which CDN stores any content, popular or not.

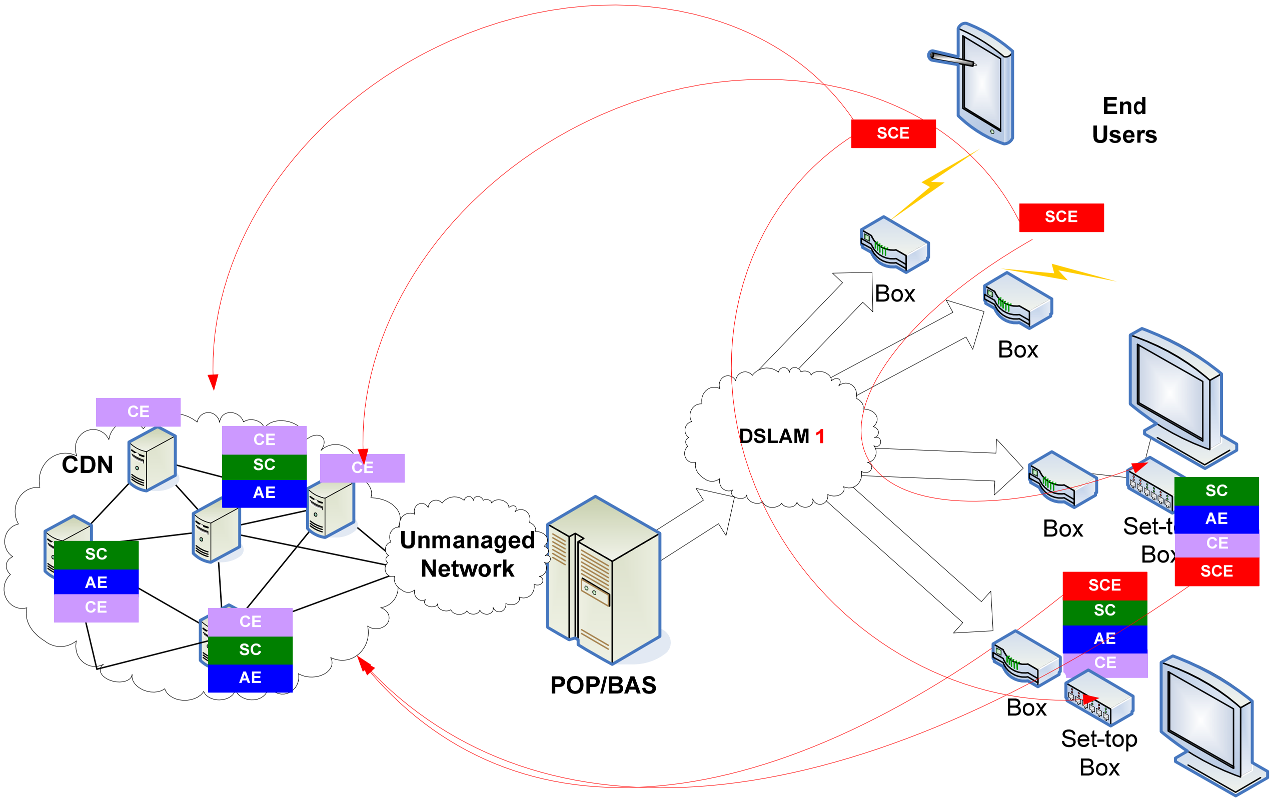


Figure 8: UGC adaptation scenario

Hence in this scenario we face with the following particularities: the adaptation must satisfy any device/ any network so as to be able to deliver the content to anybody in a distributed architecture. The adaptation will help at facing with an a priori unknown popularity using CDN caching resources only when necessary. It is beneficial for the long tail of UGC: those numerous contents with a limited popularity.

For this scenario, many aspects need to be clarified as far as the architecture is concerned.

As for the previous use cases the adaptation benefit must be evaluated: for this scenario, it can e.g. be a reduction of the caching cost in a CDN for unpopular contents.

The relation with a dCDN (distributed CDN) architecture is obvious in this scenario, hence WP1 can clarify architecture elements and interactions there.

Still the question of adaptation adequate location comes here. The adaptation capability, at the user end level, in connected devices, need to be clarified, as well as the communication protocols.

Conclusion

The current document proposes a new architecture of a generic content adaptation made of multiple components. To understand the role and functionality of the different parts, three reference use cases have been studied, although it is naturally possible to extend these scenarios to more complicated ones.

As described in section 2 the CDN may occur at multiple points in the global chain. In this study we let the possibility to have different networks architecture where the role of CDNs and their localizations in the network will be clarified in the WP1 according to economical and markets features.

Another point to be taken in account is the management of the DRM which has a strong impact on the final network architecture, as far as media handling will have to manage scrambling and descrambling operations.

The next deliverable for this work package will detail the technical specifications for the realization of a content adaptation engine suitable for the architectures identified in this report. This specification will be the initial step validating the proposed adaptation architecture.