EVOLUTION OF VIDEO ON DEMAND ARCHITECTURES

Weidong Mao, Kip Compton
Advanced Engineering, Comcast Cable

Abstract

This paper describes the architectural evolution of Video On Demand (VoD) infrastructures in supporting open interfaces and expandability to future on demand services. Key technology areas that are essential for the next generation on demand architecture are discussed.

With the advent of new technologies in Gigabit Ethernet IP networking and high performance storage/streaming as well as software, it becomes feasible to evolve the current VoD architecture from a vertically integrated and proprietary system to an open and modular architecture. Significant economic advantages are realized as a flexible and open environment makes it possible to deploy new technologies much more efficiently, facilitating the rapid proliferation of new services while controlling the associated capital and operating expenses.

This paper presents a reference architecture that describes logical components and interfaces. In addition, it provides a detailed discussion on several key areas of architectural evolution. They include:

- *Standardized asset propagation and management interfaces*
- *Better utilization of resources via dynamic session and resource management*
- *Secure content delivery via session based encryption and/or pre-encryption techniques*
- *Edge device management and configuration*
- *Shared infrastructure with other services, including IP-based streaming media services*
- *Expansion to multiple services such as networked PVR and interactive advertising*

ISSUES WITH EXISTING SYSTEMS

Cable operators currently offer Video On Demand (VoD) services through the interactive video systems that feature tight integration and customization across several system components, such as: asset management, session and resource management, billing and entitlement, network transport, and set-top client applications.

While today's monolithic system architectures helped cable operators bring a compelling product to market very quickly, there are several issues in the current VoD architecture that must be addressed going forward:

- Open Interfaces: In many of today's VoD architectures, the interfaces and protocols between different components are poorly defined and proprietary. Without standardized and open interfaces, significant effort is required to integrate any new vendor into the architecture. Earlier standardization efforts such as Time Warner's ISA and CableLab's VoD Metadata project have addressed some of these issues. However, the interfaces related to key components, such as Session and Resource

Management and Edge QAM, remain largely un-addressed.

- Multiple Service Support: Today's architectures are typically customized for a very limited set of services (e.g. Movie on Demand and Subscription VoD). Unfortunately, a significant re-engineering effort is required to support the addition of new services, like Switched Broadcast Video, or networked PVR. An extensible, on-demand platform that allows multiple services to share the same underlying infrastructure will create significant cost efficiencies and make it possible to provide new services more quickly and easily.

- High Performance, Scalability, and Redundancy: The existing VoD system deployments are typically designed for 10% simultaneous usage. Content is typically duplicated on each video server system at each server location. In addition, the network and edge resources are typically "hardwired" or associated with a specific server. To support services such as HDTV on demand and networked PVR, which will produce higher bit rates and viewing concurrency, a high performance and highly scalable distributed approach with fail-over capability is needed.

The following sections provide a VoD reference architecture and discussions around how today's VoD architecture might evolve to address these issues.

REFERENCE ARCHITECTURE

Figure 1 provides a Next Generation VoD reference architecture that supports Video On Demand services and can be expanded to support other on demand services such as Switched Broadcast Video or networked PVR.

The architecture is partitioned functionally into a number of logical components. Each component is defined in such a way that the interchangeable module implementing the common interfaces can be introduced to work with the rest of the system. It is possible that implementations may integrate several components into a single product or solution.

Each logical entity described in the reference architecture may represent one or many physical entities in an actual implementation. For example, multiple servers may be utilized to provide load balancing and scalability for the Session Manager (SM) component.

The On Demand Client is typically located at the customer premises in a digital set-top box. Any gateway server that is communicating with other headend components on behalf of the digital set-top box will be considered part of the On Demand Client. All other components are located at cable operators' master headend, secondary headend, or remote hub, depending on the specific deployment configuration and network topology.

Key logical components include:

Asset Distribution System (ADS) – distributes assets from content providers' or aggregators' premises to the network operators.

Asset Management System (AMS) – validates and manages the life cycle of assets and their associated metadata.

Real Time Source – generates assets from real time encoders and / or broadcast feeds.
Billing System – manages customer billing and service subscriptions.

Entitlement System (ES) – manages entitlements and transactions.

Navigation Server – presents assets and service offerings; manages navigation from the subscribers.

Purchase Server – receives purchase authorization check from the subscribers and validates via the Entitlement System.

On Demand Client – provides interfaces with the headend components and enables end user applications.

Asset Propagation Manager – manages asset propagation across multiple streaming servers.

On Demand Resource Manager – manages resources required from the Streaming Servers.

Streaming Server – stores and outputs contents and manages stream control.

Session Manager (SM) – manages session life cycles for on demand video services requested by subscribers.

Conditional Access System (CAS) – performs Conditional Access for the on demand video services.

Encryption Resource Manager – manages encryption configuration for each session.

Encryption Engine – performs encryption of video services associated with the session. It can be located anywhere between video server and edge device.

Network Resource Manager – manages resources required in the transport network for each session.

Transport Network – transports video services from streaming servers to the edge.

Edge Resource Manager – manages resources required at the edge for each session.

Edge Device – performs re-multiplexing and QAM modulation.

Network Management System (NMS) – provides network management for all the components in the headend.

SESSION AND RESOURCE MANAGEMENT

In order to achieve the goal of managing and sharing resources dynamically across all services, the reference architecture separates Session Management and Resource Management. The Session Manager (SM) is responsible for managing the life cycle of individual sessions for each on demand service. The Resource Managers (RM) manage shared resources (streaming, network, encryption, and edge) on the system. The Session and Resource Management (SRM) reference architecture and interfaces are described in Figure 2.

Typically, the SM will perform the following functions:

- Communicate with the subscriber device regarding session setup, session status, and session tear down.
- Interface with the corresponding Purchase Server to authorize the session requested by the subscriber.
- Allocate the resources required for the session by negotiating with the resource managers for appropriate server and network components.
- Dynamically add, delete, or modify the resources associated with the session to support integration of multiple on demand services.
- Manage the Quality of Service for the session.
- Manage the life cycle of the sessions.

One of the main functions of the SM is to obtain the required resources for each session by negotiating with resource managers for the relevant server and network components. These tasks typically include the following:

- Interface with On Demand Resource Manager to determine the Streaming Server resources such as asset location, allocated streaming server and output port. (Interface S3)
- Interface with Encryption Resource Manager to determine encryption resources required for the session. (Interface S4)
- Interface with Network Resource Manager to determine the uni-directional path that will transport the requested video stream to the edge devices that cover the service group in which the subscriber resides. (Interface S5)
- Interface with Edge Resource Manager to determine the resources used at the edge devices such as bandwidth required and MPEG tuning parameters. Digital set-top boxes can then tune to the MPEG program that carries the requested content. (Interface S6)

The session and resource management process may follow any of several possible implementation models: In a centralized model, the SM will retrieve and aggregate the topology and resource information from each resource manager. The SM then updates this information as necessary, assigns resources, and communicates with each resource manager to configure those resources.

In a distributed model, each resource manager is responsible for maintaining and updating the topology and resource state of the devices it manages as well as allocating the resources for the session on behalf of the SM. In this model, the SM collects the choices provided by each resource manager and selects an appropriate combination of resources to enable the session. Several optimizations of this model are possible to reduce latency and increase the throughput of session and resource management.

MULTIPLE VoD SERVERS

Another key area of evolution in the VoD architecture allows multiple video servers to operate in the same headend environment. This enables multiple video server vendors using the common open interfaces and helps facilitate the over-subscription of VoD storage and streaming resources within or across service groups. A reference architecture and interface diagram using the multiple servers is shown in Figure 3.

The On Demand Resource Manager is responsible for allocating and managing the resources that are required from the Streaming Servers. Upon the session setup request from the client, the Session Manager (SM) will request resources from the On Demand Resource Manager (via Interface S3), in conjunction with the resources of other components in the overall system. The resources allocated by the On Demand Resource Manager may include:

- Asset location: This includes the locations of the requested asset that has been determined by the propagation service. This information may be retrieved from the Asset Propagation Manager (via Interface R1).

- Server resource: This includes the availability of the Streaming Server that contains the asset and covers the service group in which the requested subscriber resides (via Interface R2).

- Network resource: This includes the network resource allocated at the selected Streaming Server output port (via Interface R2). They may include the UDP port number and IP address that carries MPEG SPTS.

The Session Manager (SM) will need to negotiate with the On Demand Resource Manager as well as the resource managers for any other components which allocate resources to enable streaming video from any server to any edge. For example, the asset files may not be available in the streaming server that is connected to a selected network path to the edge, creating the need for an alternate server and network path. In this case, the SM will need to negotiate with the On Demand Resource Manager and other resource managers to reconcile the differences.

Single or multiple Streaming Servers may be deployed across the network. Streaming Servers may be deployed at a centralized headend or at distributed remote hubs or both. The choice of deployment architecture will ultimately be driven by a number of factors, such as: operational feasibility, network transport availability, scalability, content caching and propagation, and the overall cost.

The Asset Propagation Manager is responsible for propagating the assets coming from the AMS to the appropriate Streaming Servers. The policies in the Asset Propagation Manager may be determined by a number of factors. For example:

- Storage capacity: determine if there is enough storage for content files.
- Content duplication: determine whether the content needs to be duplicated in a distributed manner.

The interface between Asset Propagation Manager and Streaming Server (Interface A3) is defined so that Streaming Servers from multiple vendors can be introduced to work within the same propagation service framework. It is essential that this interface hides the internal implementation of each Streaming Server's storage system. The interface may include parameters such as the required storage capacity, the available storage capacity, service group coverage, and whether to duplicate a content file.

The On Demand Client interfaces with the Session Manager for session signaling (Interface S1) and with the Streaming Servers directly for stream control (Interface C1).

SHARED EDGE RESOURCES

Another key area of evolution of the VoD architecture is the availability of a common, open interface that enables multiple services to share multiple edge devices on the same network at the same time. This can be achieved by introducing standardized interfaces between the Edge Resource Manager, the Session Manager, and the Edge Devices.

The Edge Resource Manager is responsible for allocating and managing the resources that are required from the Edge Devices. Typically, the Edge Resource Manager needs to know the topology of the service groups that Edge Devices are serving. When the Session Manager requests resources for a specific session, the Edge Resource Manager needs to determine which Edge Device to use, specify input UDP port and IP address to be utilized, define output MPEG program parameters, and set the output RF frequency.

Other functionalities of the Edge Resource Manager may also include bandwidth management and quality of service. For example, additional edge bandwidth may be required from a QAM to dynamically add content to an existing session. Quality of Service can also be provided by using techniques like MPEG bit rate reduction.

The main function of the Edge Devices is to receive the multiple MPEG Single Program Transport Streams carried over UDP/IP from the IP transport network, multiplex them into MPEG Multiple Program Transport Streams, and generate QAM modulated signals. The other features of Edge Devices may include:

- MPEG PID (Packet ID) and / or TSID (Transport Stream ID) remapping
- PCR (Program Clock Reference) re-stamping
- Statistical multiplexing
- Bit Rate Reduction

SESSION BASED ENCRYPTION

Session based encryption is the ability to dynamically encrypt on demand content in real time using a Conditional Access System (CAS) that the set-top box supports. This makes it possible to enable multiple conditional access systems to support legacy as well as next generation set-top boxes.

The encryption of digital services can be achieved by using the Entitlement Control Messages (ECM) and Entitlement Management Messages (EMM). ECMs are used to secure the control words that are required to scramble the packets. EMMs are used to enable specific users to retrieve the ECMs that are required to decode the control words and de-scramble the packets.

In the case of pre-encryption, ECMs/EMMs are generated a manner that enables a group of digital set-top boxes to access content that has been pre-encrypted and stored at the server ahead of time. In case of session based encryption, ECMs/EMMs are generated and assigned to a particular session. The content has to be scrambled on the fly at the Encryption Engine.

Whether the content needs to be encrypted may be determined by a number of factors. Content providers can require the asset to be encrypted by enabling the "Encryption" field in the corresponding asset metadata file as defined by Content Specification 1.1. Network operators can also require the specific service to be encrypted. In addition, the system shall be able to identify which CA system to encrypt the content in case of multiple CAS headend.

The Encryption Resource Manager is responsible for managing the Encryption Engines and provisioning the encryption resources required by sessions.

The Encryption Engine performs real time encryption of the MPEG-2 packets carrying on demand content. It can be located anywhere between Streaming Servers and edge devices. For example, the encryption engine may be embedded in a multiplexer or edge QAM device. In order to perform the session based encryption, the Encryption Engine needs to retrieve the appropriate parameters such as the ECMs from the corresponding CA system.

MULTIPLE SERVICES

In order to enable multiple services over the same on demand platform, a service architectural model is proposed (see Figure 4). In this model, each service will be able to use one type of the Session Manager. Several different types for Session Managers can be defined to further optimize the protocols and actual implementation for a variety of applications.

Networked PVR

Networked PVR (nPVR) services allow the subscriber to watch broadcast programming on demand and interact with a live broadcast programming (e.g. pause or rewind). To effectively achieve this goal, the network operator must record, manage, and stream broadcast programming content in real-time based on each subscriber's requests.

The reference architecture and associated interfaces can be extended to support these features by adding a few more capabilities. In particular,

- Real time asset ingest must be supported to the Streaming Servers. In particular, this includes the ability to import metadata, like programming schedule information, into the Asset Management System.
- The segmentation of the digital video programming (start and end of the programming segments) needs to be addressed. An operationally friendly scheme is also required to address programs that start late or overrun their original schedule.
- The subscriber will be able to perform asset query, purchase authorization request, and session setup / teardown for time-shifted content, just as any video on demand service.
- For live broadcast, the subscriber will automatically trigger a Networked PVR session by issuing a command such as Pause or Rewind.

The Networked PVR services impose significant challenges on the performance of the overall system. For example, the Streaming Server will be required to handle large amount of real time stream ingest. The Session Manager and Resource Managers will be required to manage a large number of simultaneous sessions in cases such as a popular live broadcast. The architecture and interface design must take these issues into consideration.

Interactive Advertising

The Next Generation On Demand Video Architecture opens new opportunities for providing innovative interactive advertising. For example, advertisement can be inserted at the beginning of a VoD session. The advertisement can be either determined statically based on the asset metadata or dynamically targeted to a particular subscriber based on a set of business rules.

From an architectural perspective, there are several areas of interest surrounding interactive program insertion services. They include: where the digital insertion will happen, how it can be done, and what determines the advertising play list.

A digital program can be inserted either at the Streaming Server location or at the Edge. Insertion at the Streaming Server provides an integrated approach and can leverage the existing storage and streaming infrastructure. Insertion at the Edge will allow separately managed Ad content servers to interface with

Edge Devices, eliminating the need for a given Ad to be stored on the same server as the content it is being inserted in.

In both cases, CableLabs Digital Program Insertion standards can be used to define the cueing messages that are required for splicing MPEG-2 streams. In addition, digital program insertion into encrypted streams should be handled properly.

Switched Broadcast Video

The architecture can be extended to support Switched Broadcast Video services. A switched broadcast system only sends the digital broadcast video streams that are being watched to their corresponding service groups. In addition, a subscriber can join an existing multicast that is available to its corresponding service group.

In more precise terms, Switched Broadcast Video is a tool to save bandwidth rather than a new service. From the subscriber's perspective, he or she still receives the same broadcast video service when switched broadcast is used. In fact, an ideal switched broadcast implementation would make it impossible to tell that the stream was switched at all. If each one of the digital broadcast channels is being watched by a subscriber in the same service group, the Switched Broadcast Video approach does not yield any bandwidth savings. However, the more likely situation is that several channels in each service group would go 'unwatched' at any given time. This "concentration ratio," (i.e. the percentage of all available channels that are watched by each service group) is the primary driver of any bandwidth efficiency that can be realized through a Switched Broadcast Video solution.

One way to extend this architecture to support Switched Broadcast Video is to utilize the Session Manager to manage broadcast sessions. For each channel change, the subscriber will send a message to the Session Manager, which, in turn, determines if the requested channel is already being sent to the subscriber's corresponding service group. If the requested channel is already being sent to the service group, the session manager will instruct the subscriber to join the existing broadcast session. A new broadcast session is assigned if the requested channel is not already being sent to the service group. The Session Manager will negotiate with the Resource Managers to allocate resources required for the session. The Edge Device needs to dynamically retrieve the MPEG single program transport stream that carries the requested broadcast program (likely via IP multicast) and generate the MPEG multiple program transport stream. As part of the session setup response message, video tuning parameters, such as frequency and MPEG program number, are sent back to the subscriber to access the requested broadcast channel.

Switched Broadcast Video imposes specific requirements on the performance of the overall system. For example, the broadcast session setup/channel change latency needs to be minimized to achieve desired channel change response time. In addition, frequent channel hopping in peak time can cause significant upstream traffic required to carry session messages. It is necessary that the architecture and interface design take these issues into consideration.

Streaming Media to PC

It is desirable to share the on demand video service infrastructure among multiple services and multiple devices, including the Streaming Media services to PC and other video enabled devices.

There are several aspects which should be considered when expanding the architecture to support shared Streaming Media platform.

- Shared asset distribution and asset management system.
- Shared session and resource management.
- Shared streaming servers.
- Shared entitlement and billing system.

Future digital set-top boxes may support advanced codecs, such as MPEG-4/AVC or Window Media 9 in addition to MPEG-2. They might also be able to receive video content over IP/DOCSIS. The VoD architecture evolutions discussed in this paper can be extended to support these capabilities in the next generation all digital platform.

SUMMARY

This paper describes a next generation VoD architecture model and its associated interfaces that will enable multiple services to be provided through multiple servers, multiple edge devices, and multiple customer premise equipments using the same underlying infrastructure. Open interfaces among the various components are the key to achieving this goal. Work is on-going to finalize the interface specifications and creating a deployment migration strategy from today's VoD architectures to the next generation framework.

REFERENCES

[1] CableLabs Video On Demand Content Specification Version 1.1, April 15, 2003

[2] CableLabs Asset Distribution Interface Specification Version 1.1, April 15, 2003

[3] Pegasus Session Setup Protocol, Version 1.0 and Version 2.1, Time Warner Cable

[4] Pegasus Lightweight Stream Control Protocol, Version 1.0, Time Warner Cable

[5] Pegasus Interactive Service Architecture Version 1.3, Time Warner Cable

[6] ISO/IEC 13818-6: MPEG-2 Digital Storage Media – Command and Control (DSM-CC)

[7] ISO/IEC 13818-1: MPEG-2 System

[8] ISO/IEC 13818-2: MPEG-2 Video

[9] IETF RFC 2326, Real Time Streaming Protocol (RTSP)

[10] Insertion Cueing Message for Cable Television System

[11] ANSI/SCTE 30 2001, Digital Program Insertion Splicing API