BANDWIDTH MANAGEMENT FOR THE UNIVERSAL EDGE

By Bruce Thompson, Xiaomei Liu
Cisco Systems, Inc.

Abstract

To offer more services to end users in the cable network, bandwidth needs at the edge of HFC network are growing rapidly. The industry is moving from current architecture where each service has its own edge resources to a multi-service universal edge architecture to use the RF bandwidth more efficiently.

This paper will go into detail on the benefits of dynamically sharing resources across services. It will also describe a control/data plane architecture that can be used to share resources between these services. Lastly, this paper will provide examples of standardized protocol suites that can be used to implement the interface between the components of a distributed architecture that supports resource sharing across services.

OVERVIEW

The HFC plant is the point of convergence for all of the services that MSOs provide. In addition to Docsis based Internet Access and Broadcast Video, new services such as video on demand (VoD), network PVR and switched broadcast are now being offered. In current deployments, each service has a statically allocated portion of the RF spectrum. Dedicated QAM pools are allocated for broadcast, VOD, switched broadcast and DOCSIS based Internet Access services.

Using VOD and DOCSIS Internet Access services as an example, Figure 1 shows the deployment scenario where QAMs are dedicated to each service. For VOD services, VOD servers send content over gigabit Ethernet (GE) to video capable downstream (DS) QAMs which reach Setop boxes (STB) in the home. VOD servers are in control of allocating both video pumps and QAMs when VoD session requests are made by STBs.

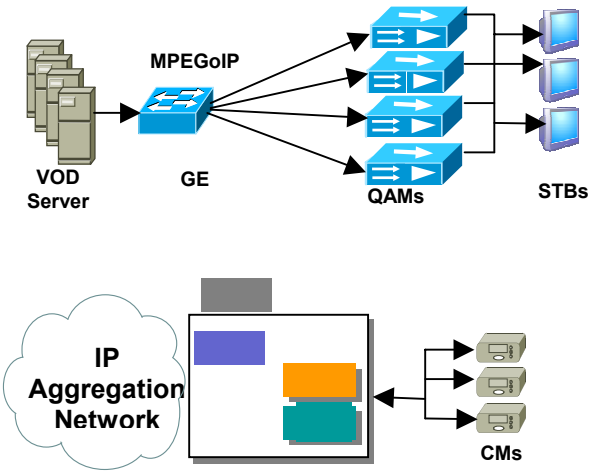

Figure 1. VOD and DOCSIS services in today's cable network

For DOCSIS Internet Access services, a CMTS serves a group of cable modems. The CMTS consists of a Docsis MAC layer processor, downstream (DS QAM) channels, and upstream (US) channels bundled into a single platform. The downstream QAMs embedded in the CMTS use a different portion of the RF spectrum than the QAMs dedicated to the VoD service.

With the increasing popularity of VoD services, high definition television, and DOCSIS Internet Access, the demand for HFC network bandwidth is ever increasing.

MSOs must use the bandwidth of the existing HFC infrastructure more efficiently to avoid having to upgrade plants as the need for bandwidth increases.

The demand for bandwidth has motivated MSOs to find ways to improve the efficiency of bandwidth utilization. The downstream for DOCSIS and other video services use QAM modulated MPEG-2 transport streams to carry the data. Because of the common transport encapsulation and modulation technique, a single set of MPEG-2 based QAM devices and associated RF bandwidth can potentially be shared across all of these services for more efficient bandwidth utilization. In later section of this paper, quantative analysis will be given to show the potential saving by sharing QAMs.

However, the current network architecture show in Figure 1 makes the QAM resource sharing impossible. First, each service is responsible for managing the QAMs dedicated to that service. Thus it is not possible to dynamically share HFC bandwidth between services. Secondly, the DOCSIS CMTS bundles both upstream and downstream together in the single logical device. This makes it difficult to share downstream RF resource between DOCSIS and video services.

A new architecture is now evolving which addresses the problems of the existing architecture. Figure 2 shows the new architecture that is capable of dynamically sharing downstream QAMs between VOD and DOCSIS services. In this new architecture, the downstream QAM is capable of both video and DOCSIS processing. In addition, the function of the DOCSIS CMTS is now broken into 3 separate components. They are the Docsis MAC processor, an upstream QAM, and a downstream QAM. These 3 components together are called the modular CMTS (M-CMTS). The components of the modular CMTS are connected using Gigabit Ethernet. This architecture makes it possible for the downstream QAM to accept both VOD and DOCSIS traffic. In later section of this paper, more details will be given on the data plane and control plane architecture which has the promise of sharing QAM resources.

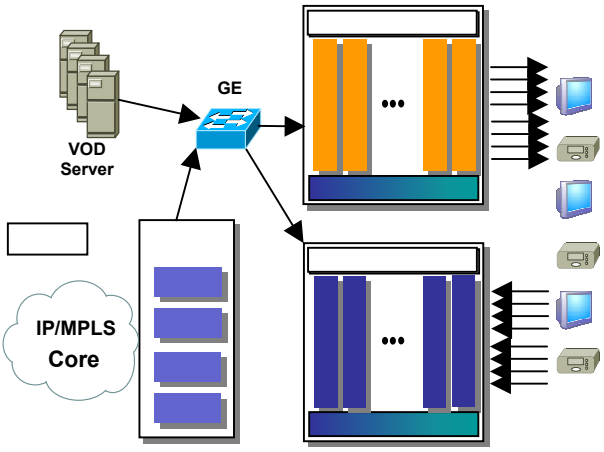

Figure 2. New architecture with universal edge QAM

BENEFIT OF UNIVERSAL EDGE

How much RF bandwidth associated QAMs can be saved when multiple service share the universal edge QAM? This section will do quantative analysis to show the potential significant savings. We will use switched broadcast, VOD and DOCSIS data services as examples.

The first factor that allows bandwidth savings is the fact that the busy hour associated with each of these services may not be at exactly at the same time or day. For example, while the busy hour associated with broadcast and on demand services are typically during the broadcast network prime time period (8:00 PM to 10:00 PM), the busy hour associated with internet access services may occur later in the evening after children have gone to bed.

The second factor that allows bandwidth savings is the savings associated with the semi random behavior of individual subscribers and the savings that can be obtained by taking advantage of the probability distribution of the behavior of a population of subscribers across services. The telephone industry has long used probabilistic models based on subscriber behavior to determine how much bandwidth needs to be deployed to allow a group of subscribers to gain access to the network with a high probability of obtaining network service. A well known model for this type of telecommunication traffic design and analysis is called the Erlang model.

Multi Service Erlang Analysis

From mathematical point of view, Erlang model has provided further evidence that the second factor mentioned above achieves bandwidth savings. The Erlang model shows that the efficiency level of a resource such as telephony trunks or RF bandwidth in a cable plant increases as the number of subscribers that share that resource goes up. When RF resources are shared across multiple services for a given population of users, the effect is to essentially increase the number of subscribers that are sharing that RF bandwidth. This more efficient usage of the RF bandwidth allows less RF spectrum to be allocated to the combined set of services that would be the case if resources were not shared.

While most Erlang calculators are specific to telephony, the calculation itself is applicable to any form of service where the arrival rate of new requests during the period over which the calculation is run (the busy hour) can be assumed to be random. Since the exact timing of how subscribers make VoD requests is not synchronized to external events (such as the advertised beginning of a television show) the arrival rate for VoD requests can be assumed to be random just as it is with telephony. Another factor that makes the Erlang calculation applicable to video is that the calculation is independent of call hold time. In telephony, the call hold time is the amount of time an average call lasts. It is typically a couple minutes. In VoD, the equivalent of call hold time is the amount of time the typical subscriber spends watching a movie. While this time is likely much longer for video than for telephony, the Erlang calculation is independent of this factor.

Since Erlang B calculators perform calculations in the context of telephony requirement, the variables of a typical Erlang B calculator must be translated into appropriate units relevant to other services such as video and Internet Access. Erlang B calculators are typically used for call center analysis and are readily available from many sources including the Internet. An Erlang B calculator has 3 variables associated with it. The calculator typically allows the user to specify 2 of the variables and it calculates the third variable.

Erlang B Analysis for Video Services

The 3 variables in an Erlang calculator are: busy hour traffic (or Erlangs), blocking factor, and capacity measured in number of lines. Busy hour traffic (BHT) is the number of hours of call traffic during the busiest hour of operation in the system. For VoD services, this can be determined by multiplying the number of homes in a service group, the percentage of homes subscribed to the service, and the engineered peak usage rate for the service. The blocking factor for VoD services specifies the percentage of time that VoD requests will be allowed to fail due to lack of QAM bandwidth. Note that the blocking factor for VoD is usually specified to be very low since it is undesirable to disallow service to a subscriber. The number of lines is the value that we are solving for in the Erlang calculations shown in this paper. For

telephony, this is the number of telephone lines that must be installed to support the specified traffic at the given blocking rate. For video, the number of lines can be translated to the number of video streams that you need QAM bandwidth for. To turn video streams into a bandwidth value, we assume that each video stream requires 3.75 Mbps of bandwidth. To determine the number of QAMs required, we then divide the resulting bandwidth by the bandwidth per QAM (38 Mbps) and round to the next higher integer.

Given the above factors, the full formula for determine the number of QAMs required in a service group for VoD services is:

BStream = BW per VoD Stream = 3.75 Mbps
BQAM = BW per QAM = 38 Mbps
Homes = homes per service group
SR = Subscription Rate
PR = Peak Usage Rate
BF = Blocking Factor

BHT = Homes * SR * PR

# of QAMs =
roundup(
ErlangB(BHT, BF) * BStream / BQAM), 1)

Note that the above Erlang analysis can also be used for switched broadcast services.

Erlang B Analysis for Internet Access

Erlang B analysis can also be used to model traffic associated with an Internet Access service. While the traffic patterns associated with Internet Access are different than telephony or video, you can still model the Internet Access service as one where subscribers are randomly making requests and the service provider is trying to provide a user experience where the subscriber gets a minimum bandwidth for a certain percentage of the time. The percentage of time that the subscriber does not get this minimum bandwidth can be considered the blocking factor for the Erlang calculation. The blocking factor for Internet Access can be quite high since the effect of a "blocked" user is that his Internet Access service appears slower than the minimum rate. Another factor that must be taken into account is that when a subscriber is using their Internet connection, they are not always making requests that require bandwidth. We must take this factor into account when calculating BHT. We call this factor the Internet usage factor. The Erlang calculations for Internet Access in this paper will use an Internet usage factor of 20% or .2.
Given the above factors, the full formula for determine the number of QAMs required in a service group for Internet Access services is:

IUsage= Internet Usage Factor = .2
BSub = BW per Internet Subscriber

BHT = Homes * SR * PR * IUsage

# of QAMs =
roundup(
ErlangB(BHT, BF) * BSub / BQAM), 1)

Multi Service Erlang B Example

In the following example, we apply the Erlang analysis to a cable plant with service usage patterns that are typical in today's network. For VoD services, the example shows with 500 homes per service group, a 20% subscription rate, a 10% peak usage rate, a 0.001% blocking factor and peak usage time of 8:00PM.

For the internet access service, we assume 2000 homes per service group, a 30% subscription rate, a 20% peak usage rate, a blocking factor of 1% and a peak usage time of 10:00PM. Finally, to calculate the number of QAMs needed for Internet Access we will assume a minimum rate per subscriber of 1 Mbps.

The non-peak hour usage rate is assumed to be half of the peak usage rate for each service.

Table 1 shows the RF bandwidth requirement and QAM resources needed per 2000 subscribers if these services use statically allocated QAMs. The RF bandwidth calculation must take into account the sum of the peaks of each service. The calculations were done using the Erlang analysis described above.

Table 1. Current RF bandwidth requirement without resource sharing

Service	Usage (%)	Blocking (%)	BHT (hour)	BW (mbps)	QAM
DOCSIS	20	1	24	35	1
VOD	10	0.001	10	101.25	12
Total					13

Note from Table 1 that the amount of QAMs required for VoD and broadcast are much greater than those needed for Internet Access services. Because of this, dynamic resource sharing does not provide much benefit with this type of usage pattern.

Table 2 shows a likely future usage pattern that will become common as the need for Docsis bandwidth grows. The basic assumption here is that the amount of bandwidth that the MSO sells the subscriber for Internet Access service will increase from 1Mbps to 4 Mbps. An example change that will drive the need for higher Docsis bandwidth is the evolution of Web based Video over IP to higher screen resolutions. In this scenario, the video usage is the same as in Table 1, but the following Docsis usage patterns apply. The increased use of Docsis bandwidth will drive down the size of the serving group for Docsis to be identical to that of VoD. In this future example, we assume an Internet Access service with 500 homes per service group, a 30% subscription rate, a 20% peak usage rate, a blocking factor of 1% and a peak usage time of 10:00PM. From table 2, it is clear that savings can be achieved if the peak usage times for Docsis and VoD services are not the same.

Table 2. Example future RF bandwidth requirement without resource sharing

Service	Usage (%)	Blocking (%)	BHT (hour)	BW (mbps)	QAM
DOCSIS	20	1	6	52	2
VOD	10	0.001	10	101.25	3
Total					5

If QAMs are dynamically allocated between Docsis and VoD, the combined service group can be provisioned for each service peak independently. Dynamic allocation will ensure that the correct number of QAMs is allocated to each service is it reaches its peak usage.

Table 3 and Table 4 shows the bandwidth savings that can be obtained in the using the usage data above if RF bandwidth is dynamically allocated to each service.

Table 3 shows the RF bandwidth requirement at 8:00 PM when the VoD service is running at its peak rate while the DOCSIS service are running at its non-peak hour rate.

Table 3. RF bandwidth requirement at 8:00 PM with dynamic resource allocation

Service	Usage (%)	Blocking (%)	BHT (hour)	BW (mbps)	QAM
DOCSIS	10	1	3	32	1
VOD	10	.0001	10	101.25	3
Total					4

Table 4 shows the RF bandwidth requirement at 10:00 PM when the Docsis service is running at its peak rate while the VoD service is running at its non-peak hour rate.

Table 4. RF bandwidth requirement at 10:00 PM with dynamic resource allocation

Service	Usage (%)	Blocking (%)	BHT (hour)	BW (mbps)	QAM
DOCSIS	20	1	6	52	2
VOD	5	.0001	5	105	2
Total				142.5	4

From the above tables, we can see that dynamic resource sharing between VoD and DOCSIS services requires 4 vs. 5 QAMs to be deployed in the serving group which results in a 20% reduction of plant bandwidth used by these services. The 20% saving comes from the difference in peak hours between the different services.

While not shown in the above example, additional savings can be obtained by sharing the same service group with switched broadcast services. In this case, additional savings can be obtained by using a single QAM pool for both VoD and Switched Broadcast services. These additional savings occur because the number of effective users sharing the same pool of QAM resources is increased. The savings is essentially due to the law of large numbers which is what is represented through Erlang analysis.

DATA AND CONTROL ARCHITECTURE

The clear separation of data plane and control plane components makes it easier to put a common resource manager to manage the resources associated with multiple services. In this section, a data plane architecture that can be used for resource sharing will be discussed first followed by a control plane architecture that can be used for dynamic resource sharing.

Data plane architecture

There are multiple ways to achieve the resource sharing among different services. Figure 3 shows an example data plane architecture. In this architecture, a VOD server, a real time broadcast encoder, a CMTS core, Downstream QAMs, and Upstream QAMs are all inter-connected through a Gigabit Ethernet network. The GE can switch traffic from any components to any components. With this architecture, DS QAM resources are shared among all three services. In other words, DS QAM is capable of both video processing and DOCSIS data processing.

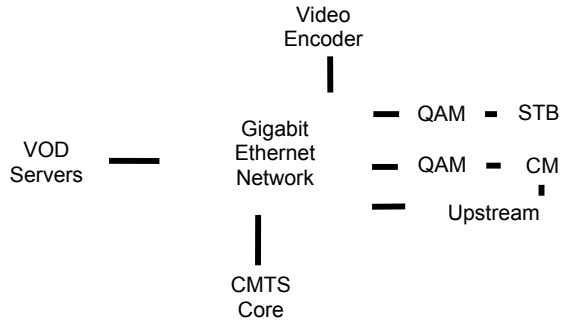

Figure 3. Data plane architecture

This architecture is highly scalable. As more services are added, only the server related to the service needs to be connected to the Gigabit Ethernet. If the QAM bandwidth needs to be increased, additional QAM resources can be shared among all existing services.

Control plane architecture

To achieve resource sharing, a common logical resource management unit needs to exist to coordinate the resource allocation of different services. A component called the session manager is then responsible for

determining the classes of resources required for a session request and communicating with the resource managers responsible for allocating those resources.Figure 4 shows an example control plane architecture that can be used for dynamic QAM allocation. In Figure 4, a control component called edge resource manager is introduced. Edge resource manager is responsible for monitoring the DS QAM resource and allocating QAMs for each service.

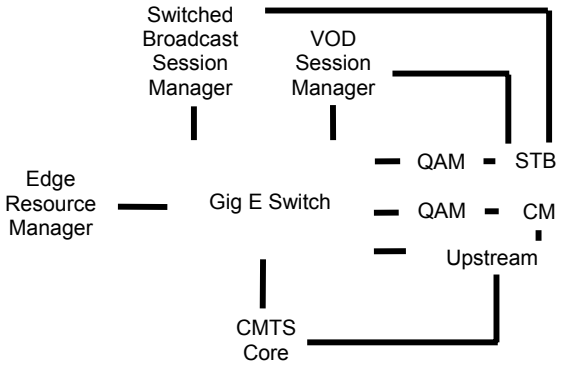

Figure 4. Control plane architecture

The control QAM resources the edge resource manager must communicate with session plane components from each service. In Figure 4, these components are the VOD session manager, and the switched broadcast session manager. The VoD session managers is responsible for accepting user requests from Set Tops for VoD sessions while the Switched Broadcast Session Manager is responsible for accepting channel change requests from Set Tops. Each of these session managers request bandwidth from the edge resource manager as part of the process of instantiating a session.

The CMTS core is responsible for the managing a DOCSIS mac domain. It will request QAM bandwidth from the edge resource manager as part of the process of setting up or modifying the bandwidth associated with a Docsis MAC domain.

The control plane architecture provides two additional functions to the system. The first is QAM discovery while the second is dynamic QAM allocation.

Service discovery protocol allows the Edge Resource Manager to dynamically detect when a QAM comes in or goes out of service. When a new QAM is added or taken out of service, the resource manage will be notified immediately about the resource change. The edge resource manager also maintains a database maintains the mapping of QAMs to service groups.

The second function added is dynamic resource allocation signaling. Each session manager signals to the Edge Resource Manager to allocate or deallocate QAM bandwidth. The Edge Resource Manager the returns information of allocated QAMs to each session manager.

This control plane architecture introduces several benefits for the system. First, it simplifies provisioning and management, which in turn reduces the operational expense. In addition, it can improve availability by dynamically reallocating QAMs when a QAM failure is detected. Finally, the separation of session management and resource management make it possible to dynamically allocate QAM bandwidth across services. This provides for more efficient use of existing HFC plant bandwidth.

CONTROL PROTOCOLS

As mentioned in previous section, the control plane supports both service discovery and session signaling. Based on the different requirements for these two functionalities, different control protocols can be selected.

For service discovery, RFC 3219 (TRIP) can be used with minor modifications to suit the needs of cable networks. TRIP is

Telephony Routing over IP protocol which deals the problem of translating telephone numbers into session signaling address of a telephony gateway in VOIP system. When modified for an HFC plant, TRIP allows a QAM to dynamically announce properties about itself to an edge resource manager. These properties include attributes such as the frequency the QAM has been configured for, the HFC service group the QAM is connected to, the amount of bandwidth that is available for the edge resource manager to allocate from, etc.

Dynamic resource signaling could be implemented with a protocol such as RTSP. RTSP is an HTTP based client / server protocol that provide a simple state machine that can be used for resource allocation. RTSP can be used by a session manager to request qam bandwidth from the edge resource manager. The request for bandwidth is encoded in an RTSP Setup message. It includes information such as the amount for bandwidth required for the session / service and the HFC serving group that the bandwidth needs to be allocated from.

After getting the SETUP request, edge resource manager uses its QAM selection algorithm to search for a best QAM to use for this session request. If QAM resources are available, the resource manager will notify the session manager about the QAM that was selected for the session. If no resources are available to satisfy this session request, the session manager will get an RTSP response with an indication of why the request failed.

CONCLUSION

This paper describes the trend in the cable industry to move to a distributed architecture where RF resources for different services can be shared. From the quantative analysis of this paper, it is clear that resource savings can be achieved by dynamically sharing resources among different services.

This paper further describes a possible architecture to achieve resource sharing and related control plane supports. At the time of writing this paper, the cable industry is actively working on standardizing this architecture and related data / control plane interfaces.

REFERENCES

1. Erlang calculator: www.erlang.com/calculator
2. Telephony Routing Over IP (TRIP), RFC 3219, IETF, Jan 2002
3. Real Time Streaming Protocol (RTSP), RFC 2326, IETF, April, 1998