Approaches to Managing Edge Resources for Triple Play Services

Doug Jones, Chief Architect, BigBand Networks
Charles Hasek, Principal Engineer, Time Warner Cable

1 Introduction

An ability to deliver video, voice and data services to anyone at anytime, in a scaleable and economical manner, is the holy grail of the cable industry. In parallel, as service offerings grow to include richer libraries of on-demand content, expanded choices of entertainment-quality long-tail content, over-the-top user-created video and more, the need to abstract services from the underlying network is an imperative.

At a minimum, cable operators are going to need edge resource management solutions capable of the following functionality:

- Efficient management of digital bandwidth at the edge of the network in a manner which promotes resource sharing to defeat service silos and stranded resources;
- Flexibility and speed in the deployment of new services;
- Maximization of revenues based on service mix that applies business policy rules on top of resource management.

Two alternative edge management architectures are already emerged:

- Global Session Resource Management – a more distributed management scheme which typically uses both session managers and session resource managers that allocate and autonomously manage universal edge QAM resources;
- Edge Resource Management – a more centralized management scheme which typically uses external switched digital video and VOD session managers that interact through a single edge resource manager that manages edge assets.

This paper provides insights into the solutions being developed to achieve this goal, including both of the architectures described above. It explains their operation and explores there advantages and disadvantages of providing a solution to converged management of edge resources for the services of today and tomorrow.

2 Resource Management Applications

Until very recently QAMs used for digital simulcast, switched digital video and VOD, along with the DOCSIS® QAMs in CMTS chassis, were independent of each other and services, consequently, could not be shared across them. Edge resource management is intended to allow these services, and other ones, to be shared across a common UEQ (Universal Edge QAM).

In a typical SDV (switched digital video) or VOD deployment today the edge resources (e.g. number of edge QAM assigned to each service group) are usually fixed; however, better resource sharing can be realized if the resource can be spread across multiple edge QAMs due to the fact that different service groups are statistically unlikely to experience peak usage at the same time, especially as the service mix continues to grow. Edge resource management enables this to happen.

Another benefit of edge resource management is the ability to share QAM capacity across a variety of applications such as SDV, VOD, broadcast video, trick-play, high-speed Internet and digital voice. A combination of UEQs and edge resource management achieve synergies that result in better utilization of network assets, an improvement in services turn-up time and other advantages. Before looking at approaches to managing UEQs, however, it is worthwhile to briefly examine some other applications of edge resource management currently being considered. UEQs are also known as edge devices.

2.1 Policy Management

Policy management uses business rules to enforce the allocation of resources under certain situations. For example, in a scenario where two services, such as a pay VOD session and a free VOD session, are competing for the same edge resource, the policy manager can determine which session is assigned resources, based on a defined set of business rules.

2.2 PCMM Architectures

With the advent of the CableLabs specifications for an M-CMTS (Modular Cable Modem Termination System), the downstream DOCSIS RF signaling resources can be provided by the same Universal Edge QAMs used to provide video services. With a suitable UEQ, the resource management architecture can also be used to allocate data QAMs to the M-CMTS core, thereby allowing all the QAMs for all the services to be managed from a common pool of resources.

Within the DOCSIS scheme, these downstream QAM resources can be managed through PCMM (PacketCable™ Multimedia) interfaces. In the future the PCMM architecture could be harmonized with the video QAM resource allocation schemes in order to provide a holistic view of edge resource utilization for both video and data services.

2.3 Variable Bit Rate Services

On demand services used constant bit rate delivery because it was simpler to implement and allowed lower cost edge QAMs. However constant bit rate service is generally understood to provide a lower quality video experience compared to variable bit rate services and, with the advent of new digital technologies and ever more powerful processors, it is possible to provide variable bit rate services for similar, if not lower, cost than constant bit rate ones. Additionally, variable bit rate services have the potential to allow more sessions to fit within a given amount of QAM capacity, due to the savings that statistical multiplexing achieves (assuming, of course, that the edge device is able to perform some level of statistical multiplexing on the variable bit rate streams).

In variable bit rate implementations, the edge resource manager needs to take into account peak and average bit rates of the service. The system must have a methodology for conveying these additional bit rates to the session and resource managers in order to allocate the correct amount of edge resource capacity and to ensure the edge resource capacity is used to its greatest advantage.

2.4 Addressable Advertising

Advertising initiatives are emerging that promise to deliver specific ads to a specific households or even a specific viewer. While these addressable advertising solutions can use various methods to accomplish their goal, a common outcome is that all will require additional edge resources in order to deliver specific addressable advertisements to specific households. The addressable advertising methodology will, therefore, need to be tied into the edge resource management topology in order to account for the additional capacity required by these addressed ads.

3 Defining GSRM and ERM Approaches

A goal of this paper is to examine architectures that can be used for edge resource management. These include:

- Global Session Resource Management – uses a GSRM (Global Session Resource Manager) and one or more SRMs (Session Resource Managers);
- Edge Resource Management – uses an ERM (Edge Resource Manager) and one or more SMs (Session Managers).

Although different, these architectures can achieve the same results in terms of managing edge resources, and the choice of which one to implement is a matter of philosophy or economics. Having two architectures can benefit the cable industry by promoting innovation and allowing choice to suit the needs of individual operators.

3.1 Global Session Resource Management

This approach uses a GSRM and one or more SRMs. Figure 1 shows the functions within an SRM which include both a SM and Resource Manager. The Resource Manager controls edge resources for the sessions which belong to the corresponding SM within the SRM. Session Management includes communicating with client software, generally residing in an STB (Set-Top Box), and setting up and tearing down sessions, examples of which include setting up and tearing down VOD sessions or channel changes with switched digital video. The Resource Manager communicates directly with the QAM in order to acquire or release resources for each session.

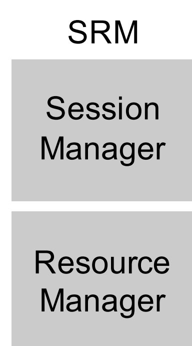

Figure 1: Functions of the SRM (Session Resource Manager)

Figure 2 shows the functions performed by the GSRM. The GSRM also contains both session manager and resource manager functions, but in addition to these includes a Global Resource Manager which allocates QAM resources to the other SRMs in the system. Allocation of resources to an SRM is generally made in bulk, meaning that some number of QAMs will be assigned to each SRM based on expected peak traffic loading. Once the resources are assigned, the SRMs manage those resources without further intervention with the GSRM. Even the SRM functions within the GSRM will manage its edge resources autonomously without further intervention with the Global Resource Manager.

Figure 2: Functions of the GSRM (Global Session Resource Manager)

The Global Resource Manager (and hence the GSRM) arbitrates requests for edge resources at multiple levels including system start up and when the SRMs may request additional capacity. To function at the highest level, the GSRM must be aware of how the resources are being allocated to the various services at all times in order to be able to best prioritize the use of those resources.

Figure 3 shows how the GSRM and SRM interact with each other and a UEQ while they are starting up, apportioning resources and arbitrating resource requests during steady-state operation. The arrows are meant to illustrate that these steps generally occur once during start up and the resource allocations are then static during the operation of the system. However, static is a relative term and the SRMs can at times of peak usage request additional resources from the GSRM.

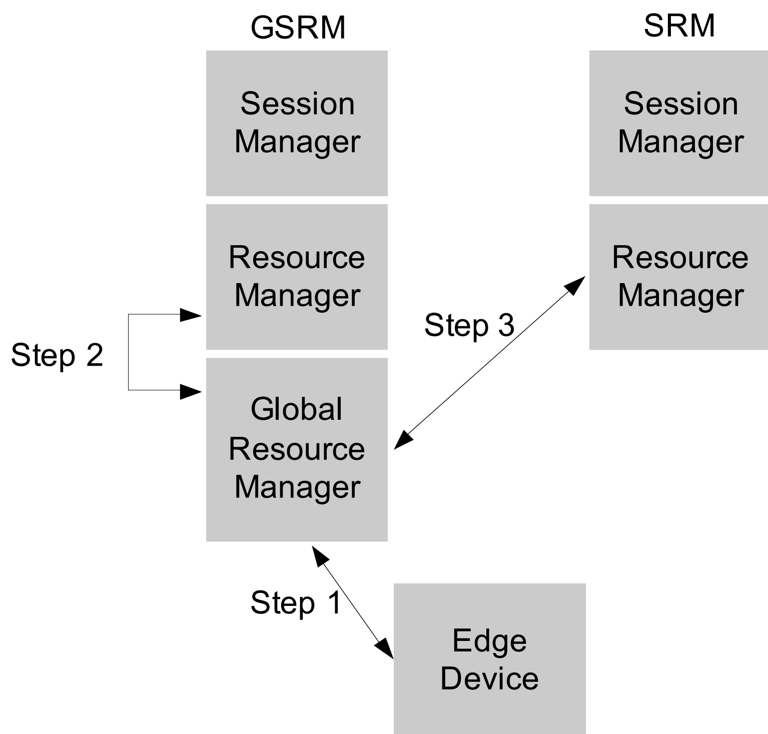

Figure 3: Steps associated with initializing the GSRM

By step 1 the Global Resource Manager has been configured with information about, or performs an auto-discovery process with the edge devices. In this way, the Global Resource Manager learns of the capabilities of the edge devices over which it has control.

In step 2, the Resource Manager associated with the GSRM starts-up and requests resources from the Global Resource Manager which then provides a bulk allocation of QAMs to that Resource Manager. Similarly, in step 3 the Resource Managers associated with other SRMs start up and request QAM resources from the Global Resource Manager. After the start-up procedure, each of the Resource Managers will have been allocated QAM resources. The Global Resource Manager can use various algorithms to allocate QAMs to the SRMs including static allocations, allocating all of the QAMs on the system, or only allocating some of the QAMs while holding some in reserve for times of peak usage.

Figure 4 shows how the GSRM architecture operates during steady-state operation. The Global Resource Manager is still part of the GSRM, it has just been drawn to the side to better illustrate steady-state operation. When a session request is received by a SM it requests resources from its Resource Manager and, if successful, the Resource Manager assigns the resources on the Edge Device. The circular arrows are meant to illustrate how during steady-state operation that the SMs and Resource Manager are interacting in real-time with many transactions per second occurring and similarly the Resource Managers are interacting with the UEQs to use the assigned QAMs.

What is not shown are the session requests from digital set-top boxes to the respective session managers; however, these are ongoing.

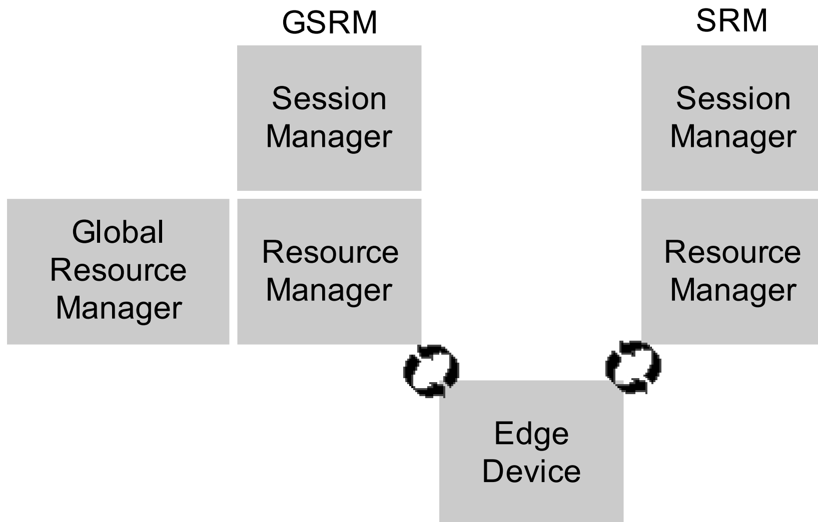

Figure 4: The GSRM in a steady-state condition

In a steady-state mode the SM can only create sessions within the QAM resources which have been allocated to it Resource Manager; if the Resource Manager has no resources available for a session, that session will be blocked and the user will receive an error message. In steady-state, the Global Resource Manager does not handle real-time requests as it allocated resources in bulk during initialization. Note also in the GSRM architecture that the QAM can have interfaces to multiple resource managers.

3.2 Edge Resource Management

This approach uses one ERM and one or more SMs, as shown in figure 5. The ERM and SM are considered separate functions that generally reside in separate network devices to allow each function to scale independent of the other. It is possible; however, for the functions to be implemented in a single physical device.

Figure 5: Functions of the SM (Session Manager) and ERM (Edge Resource Manager)

An ERM is the single device which communicates with and has knowledge of all edge resources, as shown in figure 6. At initialization, the edge devices and ERM communicate to share information about the available edge resources. The SMs neither communicate directly with the Edge Device nor do they allocate edge resources. The SMs only do session management and every time a session needs edge resources, that SM requests those resources from the ERM. When a session ends, the SM tears down the session and notifies the ERM that the resources associated with that session are no longer needed. The ERM in turn deallocates the resources from the edge device and these resources can now be used for another session.

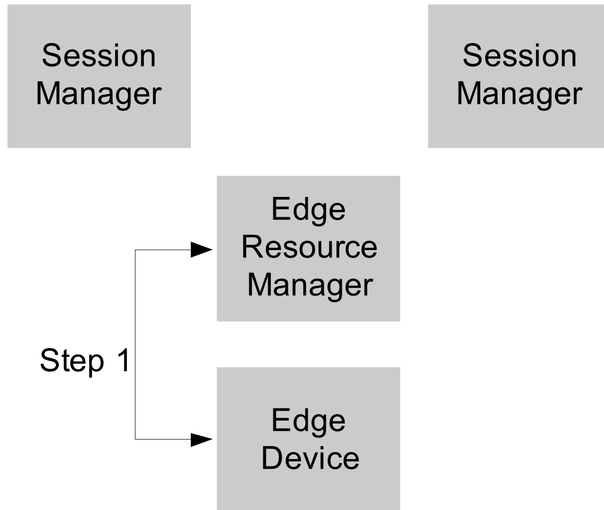

Figure 6: Steps associated with initializing the ERM

At steady-state the SMs are communicating with the ERM, and the ERM is communicating with the edge devices, as shown in figure 7. The circular arrows are meant to illustrate how this is a real-time with many transactions per second occurring. This figure illustrates how all the edge resources are managed by the ERM.

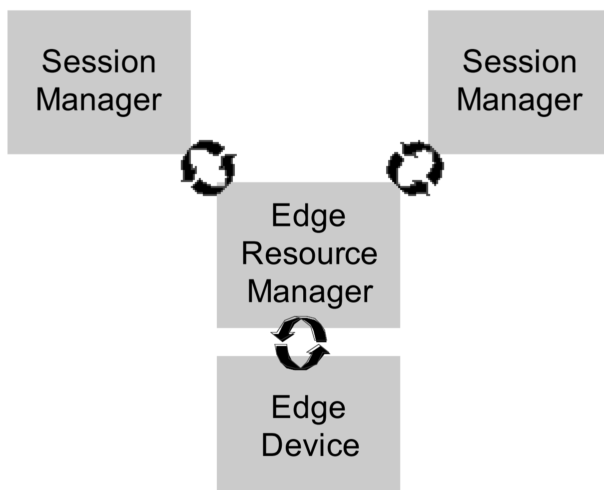

Figure 7: The ERM in steady-state mode

In steady-state mode, the ERM will place a video session on whichever QAM it deems appropriate. The resource allocation algorithm applied by the ERM could either be static, that is, it uses pre-determined QAMs for each service, or the ERM could place sessions from different services onto the same QAM channel.

3.3 Comparing GRSM and ERM Proposals

Both the GSRM and ERM approaches work, can be customized of a particular set of business rules and cable applications, and are being deployed in the field today.

Within a GSRM architecture, the SRM is constrained to just the edge resources it has been allocated; however, a particular implementation may allow an SRM to request additional edge resources from the GSRM during periods of peak operation. If the GSRM allocates additional edge resources to the SRM, that SRM may manage within those resources as it sees fit, and provide an opportunity for innovation within the session managers of emerging services. Constraining edge resource usage too tightly may restrict innovation in how new services develop to generate revenue and meet the needs of operators and customers.

Within the ERM architecture, the SM can only request edge resources; all of the edge resources are allocated and managed by the ERM which allows for centralized management of all edge devices. Centralized management could allow for more efficient allocation of edge resources between different services such as Switched Digital Video and VOD. This efficiency, however, could come at the cost of more tightly constraining implementations and innovation in the SMs.

Ultimately, both architectures are competing to provide similar solutions while pushing each other in terms of innovation. The choice is up to the operator as to which solution best fits their needs.

4 GSRM and ERM Use Cases

4.1 Overflow Operation

In order to compare and contrast the operation of the two proposals, a use case is explored to show how edge resources are managed with either proposal. Since operation during off-peak usage is fairly straight-forward, this use case will consider the operation during peak usage when SMs or SRMs are requesting more sessions than anticipated.

Figure 8 shows one way to implement GSRM and how QAMs can be allocated between two services with a third set of QAMs into which either service can overflow if allowed by the GSRM.

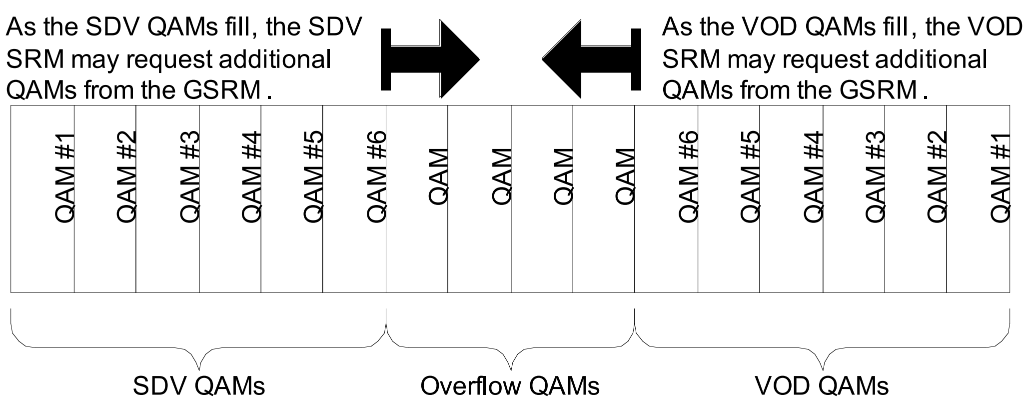

Figure 8: Basic operation of a GSRM

When initializing a GSRM system, the SDV and VOD SRMs are allocated a particular set of QAMs, which the respective SRMs manage. If an SRM experiences a peak in session requests and needs additional QAM capacity, that SRM can request additional capacity from the GSRM. If the GSRM has those resources available and grants them to the SRM, the SRM can then begin using those QAM resources to deliver services. In this way, the system assigns resources to reflect real-time demand – this is useful in scenarios where the services have peak busy rates at different times during the day. Additionally, the system could manage QAMs which have different capabilities, for example, QAMs which have the ability to perform session-based encryption could be managed in one pool and QAMs which do not have encryption resources could be managed as a separate pool.

Continuing this use case, figure 9 shows a scenario where one service can even overflow into the resources assigned to another SRM. When this occurs, the VOD service is peaking at such a rate that its SRM has not only requested (and been assigned) all the overflow QAMs, but the VOD SRM has requested additional resources and based on business policy rules has been assigned QAMs which had previously been assigned to the SDV SRM. To accomplish this, the GSRM would have to de-allocate resources from the SDV SRM. In turn, the SDV SRM would have to move any SDV sessions from those QAMs which have been re-tasked to the VOD service to other SDV QAMs. The GSRM is the ultimate manager of edge resources and the SRMs have to obey its orders and do what is necessary.

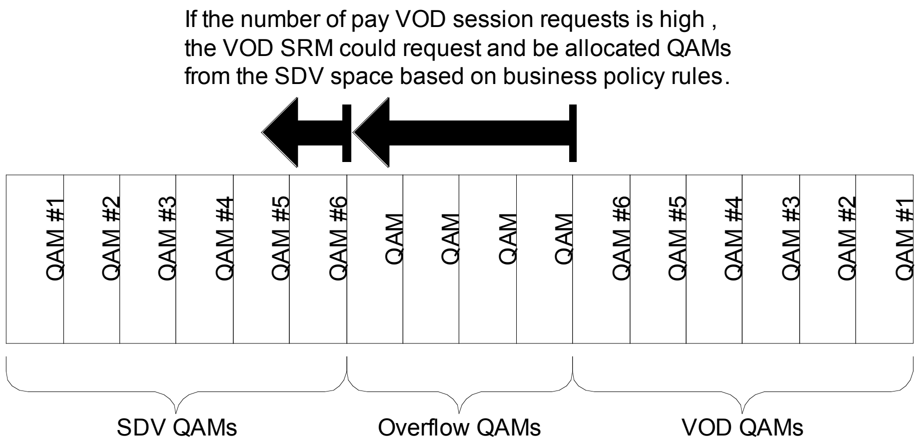

Figure 9: GSRM Overflow into another Service

In both the cases of figures 8 and 9, the system needs a way to reclaim unused capacity when it is no longer needed by an SRM. Either the SRM can determine it has excess unused capacity and return that capacity to the GSRM, or the GSRM can poll the SRMs for their resource utilization and reclaim resources as needed.

While the use cases illustrated in both figures 8 and 9 were described for the GSRM architecture, they could as easily be implemented within an ERM architecture. The ERM would just assign SM requests to a particular QAM or set of QAMs on a real-time basis and would manage the overflow of one service into a QAM or group of QAMs. If one service needed additional resources, the ERM could instruct an SM to free-up QAM resources by either moving those sessions to anther QAM or just tearing down sessions. For example if the VOD SDV needed additional resources, initially the switched digital video SM could tear down heuristics sessions, however, ultimately the ERM could be responsible for managing edge resources to maximize revenue and could prioritize certain sessions over others based on business rules just as the GSRM can.

4.2 QAM Sharing Operation

Additionally, both the GSRM and ERM architectures can allow the mixing of services within a single QAM which can be thought of as the allocation of a partial QAM channel to a particular service and the rest of that QAM channel to a different service. For example, VOD and switched digital video sessions share many similar characteristics and can be shared on the same QAM channel in this emerging environment. Sharing the services on the same QAM channel can result in both better QAM utilization by not stranding the capacity associated with "partial QAMs" and with a better quality of experience for viewers by allowing a service with might otherwise be blocked to be placed onto any QAM where there is available capacity.

Figure 10 is an illustration of how a 256 QAM channel can deliver 38.8 Megabits per second of digital information which can support 10 CBR video sessions, each with a rate of 3.75 Mbps. This configuration is in general use today for both VOD and switched digital video where the 3.75 Mbps includes both video and audio information. In addition to the ten video sessions, the QAM channel will have about 1.3 Mbps of capacity left for overhead which could be used for a number of purposes including SI (Service Information) and other information specific to the QAM multiplex.

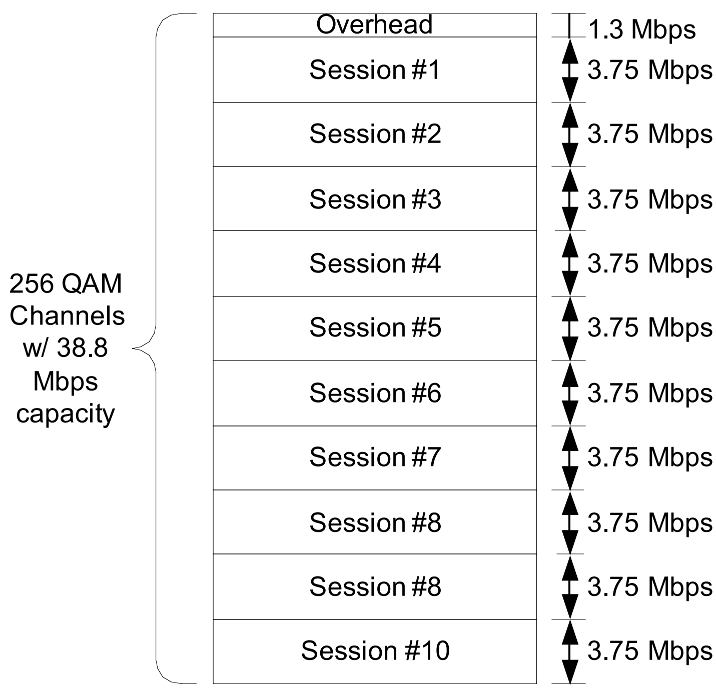

Figure 10: Video sessions in a QAM channel

In Figure 10, any of the sessions could be VOD or switched digital video. A Universal Edge QAM should not place restrictions on what types of services can be offered through it. And while not reflected in the diagram, a UEQ channel should also support the capability to be partially assigned for use with data services by an M-CMTS. Sharing a QAM channel between video sessions and data is still an emerging use case and the resource sharing control plane still needs definition, but UEQ implementations should not preclude such a configuration.

Figure 11 shows how in a GSRM architecture resources associated within a singe QAM channel can be shared between services. In this case, both the switched digital video and VOD SRMs have been allocated a single 265 QAM channel. This particular figure illustrates a case where the SDV QAM has completely filled with ten sessions and the VOD QAM is only partially filled with five sessions. The switched digital video SRM has communicated with the GSRM to request additional resources and the GSRM has fulfilled this request by assigning to the SDV SRM part of the resources that had originally been associated with the VOD SRM (and the GSRM has taken these resources away from the VOD SRM). Now when the 11th SDV session is attempted it will be allocated onto a QAM which is also delivering VOD sessions.

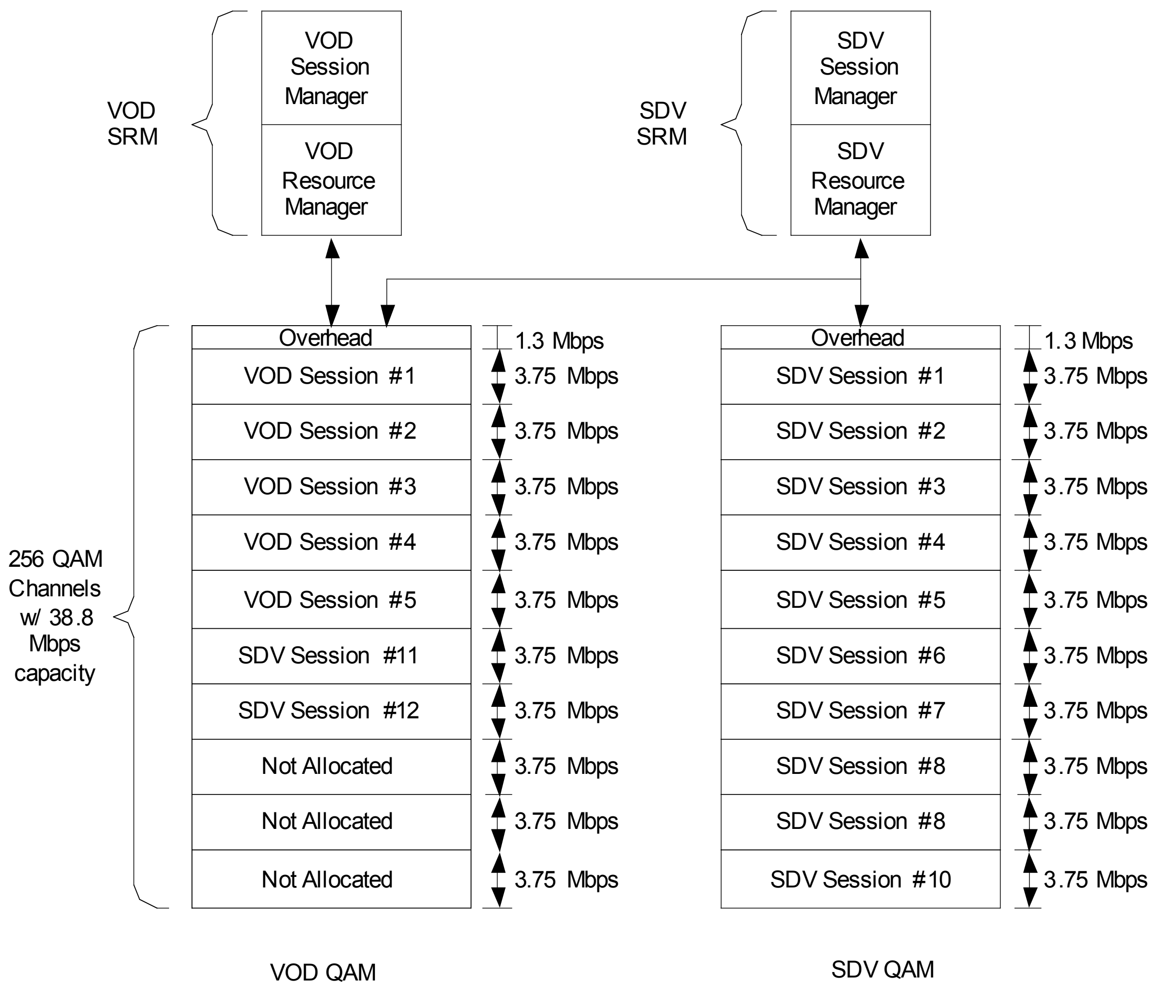

Figure 11: GSRM QAM channel allocation

Similarly, figure 12 shows the case for an ERM architecture where that ERM manages all the edge resources. In this case when there is an 11th SDV session to be added, the ERM can place additional SDV sessions on any QAM with available resources since it is the ERM which has the view of how all the resources have been allocated.

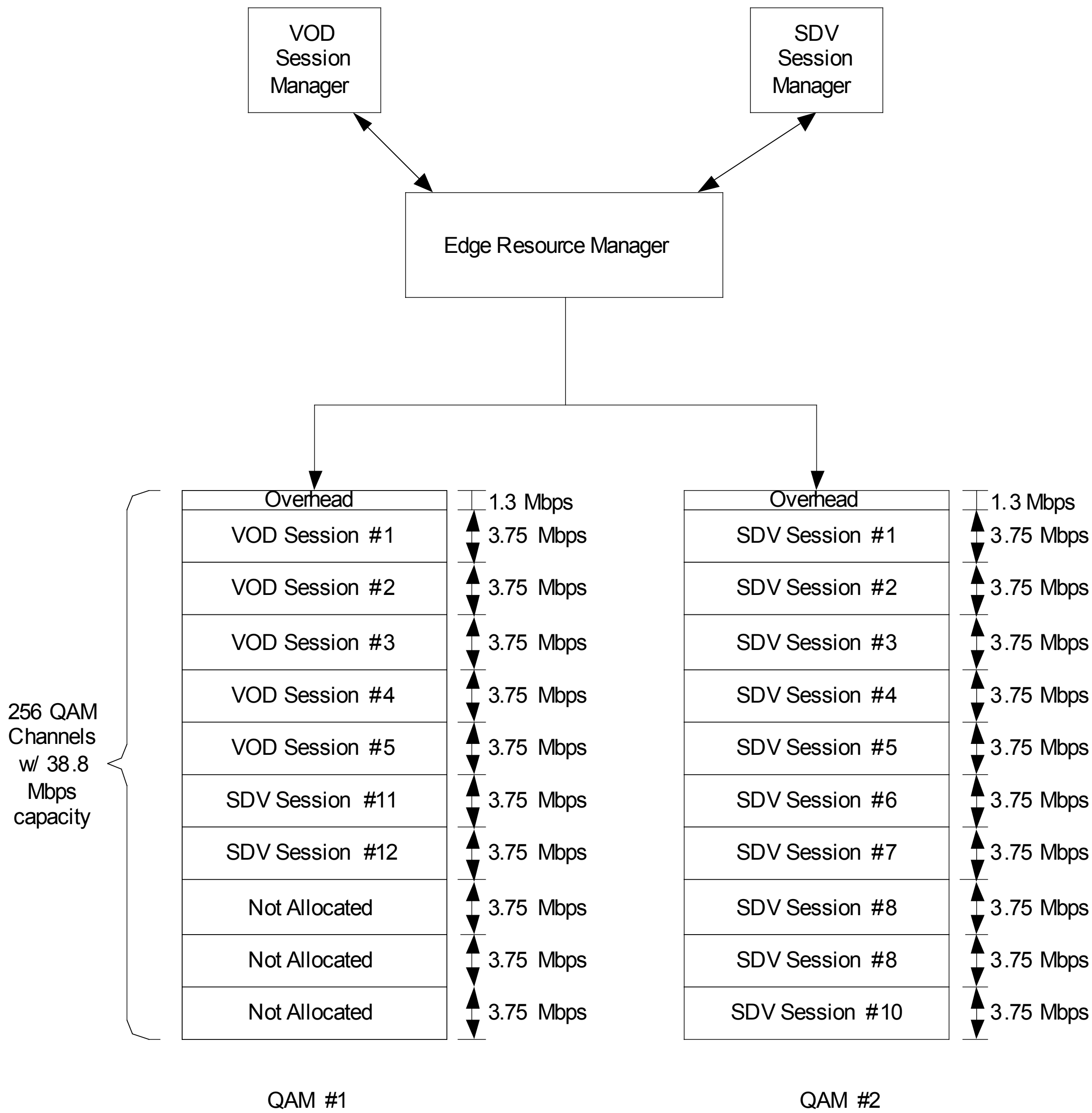

Figure 12: ERM QAM Channel Allocation

With the ERM, the QAMs do not belong to one service or another, rather, the ERM allocates resources within QAM channels to best utilize the available resources. While figure 12 shows that VOD sessions were initially allocated on QAM #1 and SDV sessions on QAM #2, another implementation may have simply mixed both types of sessions across all available QAMs as a first come, first served basis. How the sessions are mapped to QAMs is result of the bandwidth allocation algorithm running on the ERM.

5 Looking Forward

5.1 Other Last-Mile Networks

Edge resource management facilitates the efficient use of resources and will be applicable beyond the cable HFC network. For example, the cable industry has recently invested in wireless spectrum where last-mile bandwidth is provided by wireless antenna and the amount of available spectrum will be limited by federal regulations. In this environment, the last-mile wireless spectrum should be managed to maximize return on investment through a combination of policy and edge resource management.

In the event cable moves to other architectures for the HFC network, such as IPTV, or migrates to other networks such as PONs (Passive Optical Networks), edge resource management will remain applicable.

Basically every last-mile network has a finite amount of resources which are available for delivering services. The last-mile network always seems to be either the most expensive to upgrade or is otherwise limited in capacity through laws or regulation. Maximizing the revenue associated with any last-mile network, present or future, will utilize edge resource management.

5.2 Network Resource Management

The concept of edge resource management can be extended to all network resources as shown in figure 13.

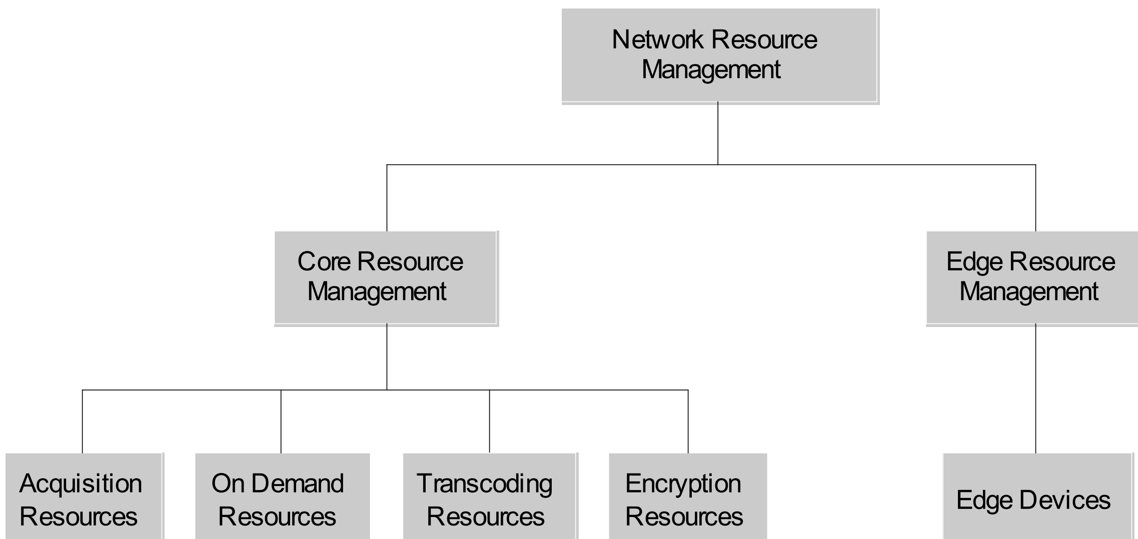

Figure 13: Network Resource Management

Network resource management defines specific network functions and places those resources under a common management umbrella which allows the network to allocate resources to specific services based on business policy rules. In addition to the edge resource management which has been discussed, core network resources could include such functions as acquisition, on demand, transcoding and encryption resources. Each of these resources could be applied as needed to personalize a service for a viewer or a device which is consuming that service.

5.3 Combining Video and Data Edge Resource Management

The examples given in this paper have mainly dealt with video where individual sessions are of a convenient constant bit rate so they pack nicely into a QAM channel. With data, the sessions have bursts and are much less well defined than a video session. But with the advent of the M-CMTS specifications data can utilize QAM resources that have been disaggregated from the M-CMTS Core. The ability to utilize spare QAM resources on an as needed basis for additional data services could be an important service as broadband services continue to gain acceptance.

6 Conclusion

There are two emerging architectures for edge resource management, and both provide equivalent services although their implementations are different. The intent of this paper is to define these two architectures and illustrate how they operate while also showing how they can both be used to manage edge resources and share services across a common universal edge QAM.

A goal of edge resource management is to ensure that edge resources are managed in a consistent fashion across multiple services. Additionally, those edge resources should be monetized to the greatest extend possible and provide the ability to dynamically assign and repurpose services to different QAM channels as needed to fit an ever changing mix of services.

Having two architectures compete to achieve these goals will benefit the cable industry as each architecture strives to develop new and innovative solutions to common issues. Overall the consistent management and sharing of edge resources should benefit the industry by allowing on a single edge QAM to support multiple services thereby extending its usefulness.

Abbreviations and Acronyms

AA	Addressable Advertising
CBR	Constant Bit Rate
CMTS	Cable Modem Termination System
CPE	Customer Premises Equipment
DOCSIS	Data over Cable Service Interface Specifications
ED	Edge Device
EQAM	Edge QAM
ERM	Edge Resource Manager
GSRM	Global Session Resource Manager
HFC	Hybrid Fiber Coax
HSD	High-Speed Data
IP	Internet Protocol
MPEG	Moving Picture Experts Group
MPTS	Multiple Program Transport Stream
PCMM	PacketCable Multimedia
QAM	Quadrature Amplitude Modulation
RM	Resource Manager
SDV	Switched Digital Video
SM	Session Manager
SPTS	Single Program Transport Stream
SRM	Session Resource Manager
STB	Set-Top Box
UEQ	Universal Edge QAM
VBR	Variable Bit Rate
VOD	Video On Demand