QAM Resource Allocation in Mixed-Format VoD Systems*

Jiong Gong[1], David Reed[1], Terry Shaw[1], Daniel Vivanco[1] and Jim Martin[2]

[1] Cable Television Laboratories, Inc.
858 Coal Creek Circle
Louisville, CO 80027
j.gong@cablelabs.com, d.reed@cablelabs.com, t.shaw@cablelabs.com, dvivanco@engr.colostate.edu&

[2]Department of Computer Science
Clemson University, USA
jim.martin@cs.clemson.edu

***Abstract*— A Quadrature Amplitude Modulation (*QAM*) resource allocation algorithm for Video on Demand (*VoD*) traffic is presented. Based on stream encoding rates and available system capacity, the proposed on-line *non-mixing algorithm* will select the *QAM* modulator that maximizes system efficiency in a *VoD* service group. A discrete event simulator has been developed to compare the performance of the proposed algorithm against two commonly deployed algorithms known as the *least-loaded* and *most-loaded* algorithms. Our analysis suggests that the *non-mixing* algorithm performs better than the two algorithms under a range of assumptions that take into account peak concurrent usage rate and traffic mixture composition. We show that the *non-mixing* algorithm leads to an average of 4.39% higher allowed peak usage rate than the *least-loaded* and *most-loaded* algorithms.**

***Index Terms*—VoD, Broadband access, Capacity planning, Congestion control, Traffic modeling, Resource allocation, Network simulation.**

I. INTRODUCTION

Video on Demand (VoD) systems operating in cable networks will see significant change in usage patterns and demand over the next five years. VoD peak usage is likely to significantly increase from the current 5% to approximately 30% as a result of larger deployments and the introduction of new, innovative applications [5]. Further, the percentage of high definition (HD) VoD stream requests is likely to increase drastically from zero to approximately 10%1. Cable operators will require a detailed understanding of the impact of these changes in the provisioning process.

When a cable subscriber purchases a VoD selection, the video stream is assigned to a QAM modulator over a specified 6 MHz RF channel. The encoding bit-rate of the stream along with the specific *QAM* configuration determines the maximum aggregate number of streams that can be assigned to the channel. Forecasting the optimal number of *QAM* modulators needed in a system traditionally has relied only on the number of concurrent requests received by the system during a busy period. A widely used rule-of-thumb provisioned *VoD* systems to support 5% of the peak subscriber load. The planning process was trivial as only one encoding rate (referred to as standard definition or *SD* if MPEG2 stream formats are used) was used on the delivered streams. For instance, over a typical 256 *QAM* modulated channel (which provides 37.5 Mbps of capacity), if all content is in *SD* format that is encoded at a constant bit rate of 3.75 Mbps, 10 streams can be assigned to the same channel fully utilizing the channel bandwidth. However on current and future *VoD* systems, different stream formats (i.e., combinations of *SD* and *HD* streams) must be considered. A mix of *SD* and *HD* competing for *QAM* resources in a service group may leave some amount of the channel bandwidth 'stranded'. This occurs when a channel is not fully utilized but there is insufficient capacity to satisfy a further stream request. The situation is exacerbated as the difference between encoding rates of different stream formats grows. In this paper, we assume an *SD* stream consumes a constant 3.75Mbps and an *HD* stream consumes a constant 12.5Mbps.

The worst case percentage of stranded bandwidth, (B_s), in a mixed bit-rate *VoD* system is $B_s = \left(\frac{r_h - r_s}{Q}\right)$,

---

* This paper is an expanded version of our paper "*VOD QAM Resource Allocation Algorithms*" published in the conference proceedings of IFIP Networking 2006.

& Daniel Vivanco, was working at Cable Television Laboratories, Inc at the time this research was conducted.

[1] From a commercial North American cable operator's market forecast. One cable operator is lately seeing close to 10% peak usage rate after the introduction of sVoD service.

where r_s and r_h denote the streaming bit rate for *SD* and *HD* streams respectively, and *Q* is the channel capacity. In the worst case, each *QAM* modulator has under r_h bandwidth stranded. This could occur if a series of *HD* stream requests arrive that almost fills the *QAM* modulator (i.e., to the point where one more *HD* request would completely fill the modulator), but then an *SD* stream request arrives and gets allocated. For the 256 *QAM* scenario described above, up to 23.3% of the channel bandwidth could be stranded.

The two most widely used *QAM* allocation algorithms are the *least-loaded* and the *most-loaded* algorithms. The former allocates incoming streams starting from the lightest-loaded *QAM* modulator. The most-loaded algorithm starts from the busiest-loaded *QAM* modulator. *Most-loaded* algorithm performs better than *least-loaded* algorithm when there is the presence of *HD VoD* streams, a fact that is confirmed in our analysis. We propose and evaluate a new *QAM* resource scheduling algorithm called the *non-mixing* algorithm. The performance of the proposed algorithm is validated using a discrete-event simulator. Our results suggest that *non-mixing* algorithm can allow peak usage rates 4.39% higher than *most-loaded* algorithm. Further contributions of this paper are the development of a *VoD* usage model and the development of a realistic *VoD* simulation model.

This paper is organized as follows. Related work is presented in section II. The *VoD* usage model and the proposed *non-mixing algorithm* are presented on section III and IV, respectively. In sections V and VI we present our analysis methodology and simulation-based results, respectively. Finally section VII presents the conclusion and identifies future work items.

II. Related Work

Large amount of prior research has addressed the scalability of large-scale *VoD* systems. Techniques have been identified that reduce the resources that are required per session. Batching requires users to wait in a group for the same content for a predetermined amount of time and then serves them in a batch using a single multicast channel [4][5][1]. Periodic broadcasting schedules the transmission of content over multiple channels in periodic intervals allowing arriving users to join the next cycle [3][13][8]. Patching attempts to merge users who are on separate channels to an existing multicast channel [12][14]. Piggybacking merges users on separate channels by slightly changing playback rates of users in an effort to have everyone get to the same point in the stream at which time the separate channels would be exchanged for a single multicast channel [9][15]. While these ideas are likely to be relevant in future cable *VoD* systems, most current deployments are relatively small in scale. Provisioning the optimal number of *QAM* modulators in a *VoD* service set is generally based on the rule of thumb that says about 5% of the total subscriber population will use *VoD* during peak periods. There has been industry discussion on *QAM* allocation algorithms [10][11]. However, academic evaluation of *QAM* resource allocation algorithms is in its early stage [7].

The *QAM* allocation problem is essentially a bin packing problem. The classic bin packing algorithm packs a list of items $L = (a_1, a_2, ..., a_n), a_i \in (0,1]$ for all *i*, into the minimum number of bins each with a capacity of 1. The *least-loaded QAM* allocation algorithm is a form of best fit packing and the *most-loaded* allocation algorithm is a form of worst fit packing [2]. In brief, a best fit packing algorithm selects the bin that has the most free space and the worst fit algorithm selects the bin that has the least free space. The standard metric that is used to evaluate bin packing algorithms is a measure of the number of bins that are required to pack various input lists. The ratio of the number of bins required by the algorithm under study to the number of bins required by an optimal algorithm (i.e., an off-line algorithm) is known as the *R* value. It has been shown that both the best fit and worst fit algorithms have an *R* value of 2 [2]. In the *QAM* allocation problem domain, the number of bins is fixed. Items in bins may leave after an amount of time (i.e., when the subscriber finishes watching the movie the stream is removed from the *QA*M). Rather than use the *R* metric, we are interested in the probability that a stream's request is denied due to insufficient capacity. We use the blocking rate to characterize allocation algorithm performance.

III. VoD Traffic and Usage Model

We have developed a model of *VoD* usage based on empirical data. The data used in this study were collected from 200 service groups of a large cable operator in North America. The average size of each service group is approximately 500 set-top boxes.

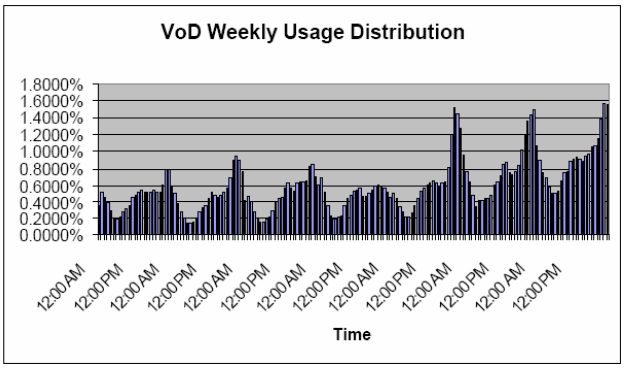

Figure 1. Weekly VoD Usage (Sunday 12:00am through Sat 12:00am)

Figure 1 illustrates diurnal average usage patterns over the course of one week for all requests, starting at 0:00 a.m. on Sunday. Note that the maximum 2% *VoD* usage rate shown in Figure 1 was the averaged over all 200 service groups analyzed, while some service groups exhibited peak usage rates close to 8%. Results of Figure 1 show higher usage rate values for Thursday, Friday and

Saturday evening from 10pm until midnight. Viewed on a daily basis, Saturday, Friday and Thursday represent 22%, 19%, and 15% respectively of the total weekly stream volume, with the rest almost evenly distributed around 11% each among the other days of the week. As a result, *VoD* usage behavior over a typical week may be categorized as either heavy usage or light usage.

Figures 2 and 3 show the daily usage patterns for Saturday and Sunday. After 7 p.m. stream usage starts to increase, peaks in the 10–12 p.m. window, and then rapidly drops after midnight. Note that Figures 2 and 3 show a *VoD* hourly usage distribution over a daily stream volume, while Figure 1 does the same over weekly stream volume.

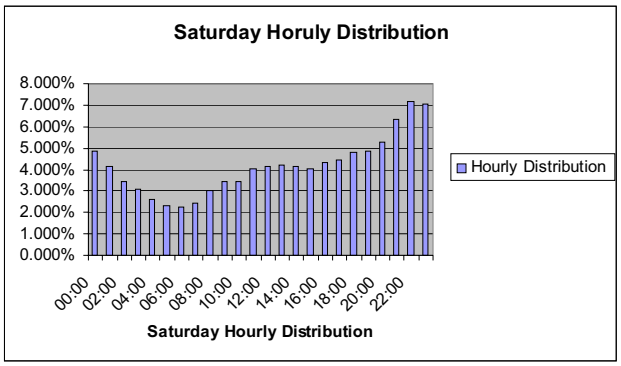

Figure 2. VoD Usage Behavior on Saturday.

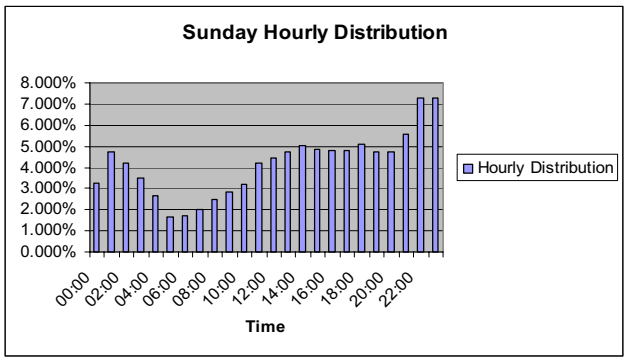

Figure 3. VoD Usage Behavior on Sunday.

We modeled the inter-arrival times of *VoD* request streams and their duration. Our results indicate that inter-arrival times follow an exponential distribution, although each of the main genres of content, including mainstream movies, adult content and video browsing have different fittings. Video browsing refers to short-lived streams mainly generated by *sVoD* subscribers that browse the available *VoD* channels available in his/her subscription package.

Figure 4 shows the fit of inter-arrival *VoD* request times associated with 7800 instances of stream arrivals observed in the data set. The x-axis represents 10-second windows over a period of 24 hours. The solid line is a fitted exponential distribution curve with a λ of 0.091. We have also modeled the viewing time distribution of mainstream movies. Our results suggest a mixed probability distribution. As illustrated in Figure 5, there was a significant mode at a viewing time of 2 hours which is the average length of mainstream movies. The data suggests a large number of early exits, which can be associated with video browsing generated by *sVoD* users.

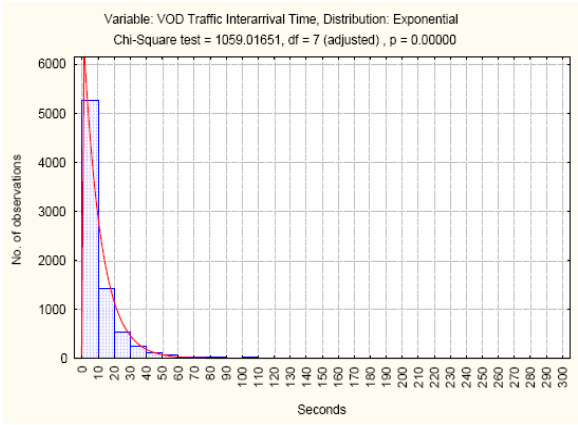

Figure 4. Histogram and Fitted Exponential Distribution of *VoD* Request Inter-arrival Times

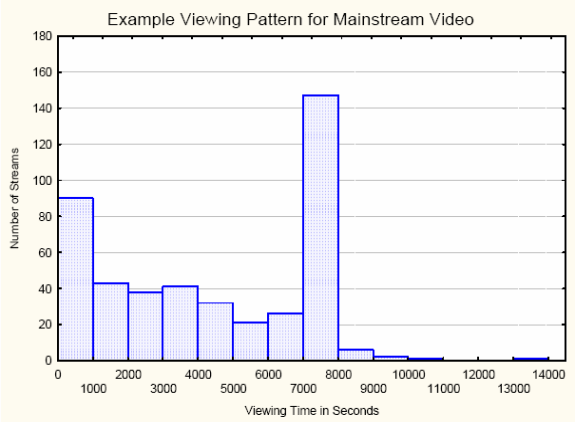

Figure 5. Distribution of Stream Length for Mainstream Video Titles

IV. Non-Mixing Algorithm

In this section we describe a new *QAM* resource allocation algorithm, named *non-mixing* algorithm. Suppose a collection of *n QAM* modulators is deployed to serve a *VoD* service group. Let q_i, $i = 1, 2, ..., n$, denote the used capacity of each *QAM* modulator *i*. Total capacity, *Q*, which is 37.5 Mbps for a 256 *QAM* modulator, is assumed to be the same for all modulators. Therefore, the remaining capacity that can be used for new stream requests on that *QAM* modulator is then

$Q - q_i$. Let r_s and r_h denote the streaming bit rate, respectively, for *SD* and *HD* streams. The two types of streams may arrive at a collection of *QAM* resources according to two distinct random processes, but exit the system based on the same holding time distribution. We call the current state of any given *QAM* (q_i) at a particular time as an allocation. We define a *VoD* system as inefficient, if,

$$Q - q_i < r_h, \ \forall i, \text{ and } \sum Q - q_i \geq r_h \qquad (1)$$

In other words, none of the *QAM* modulators individually has the capacity, even though the sum of all available resources on each *QAM* modulator is able to support one or more *HD* stream requests. A better scheduling algorithm would generate fewer cases of inefficient allocations. Note that while each type of stream is assumed to be in itself modulus in its own bit-rate, they jointly are not when they are mixed together in the same *QAM* modulator. As a result, inefficiency tends to arise when different encoding bit-rate streams are mixed together. Both *most-loaded* and *least-loaded algorithms* don't avoid mixed allocations at the *QAM* modulators.

We introduce the *non-mixing* algorithm by defining four possible states for any *QAM* modulator on the system at any given time, depending on its current allocation. The states are:

- No streams have been allocated.
- A mixture of *SD* and *HD* streams are occupying it.
- Only *SD* streams are occupying it.
- Only *HD* streams are occupying it.

These states are defined as $S_i(q_i)$, and mathematically expressed in equation (2).

$$S_i(q_i) = \begin{cases} 1, \ if \ q_i = 0 \\ 2, \ if \ q_i = x_i r_s + y_i r_h, \ x_i \neq 0, \ y_i \neq 0 \\ 3, \ if \ q_i = x_i r_s, \ x_i \neq 0 \\ 4, \ \ if \ \ q_i = y_i r_h, \ y_i \neq 0 \end{cases} \qquad (2)$$

The variables x_i and y_i are positive integers representing the number of *SD* and *HD* streams, respectively, occupying *QAM* modulator *i*. In the above four states, we call a *QAM* modulator in state 1 an empty modulator. We call a *QAM* modulator in state 2 ($S_i(q_i) = 2$), a mixed modulator. *QAM* modulators in state 3 and 4 are called non-mixing *SD* and *HD* *QAM* modulators, respectively. The algorithm selects a *QAM* using the following prioritized rules:

1. Select a non-mixing *QAM* modulator of the same stream type.
2. Select an empty *QAM* modulator.
3. Select a mixing *QAM* modulator.
4. The last resort is to create another mixing *QAM* modulator by selecting an existing *QAM* that currently has only *SD* or only *HD* streams.

If there are multiple *QAM* modulators available within the same state class, priority is given to those *QAM* modulators that have higher likelihood of becoming a non-mixing *QAM* modulator or an empty *QAM* modulator once some streams start to drop. This implies the following rules:

- If multiple non-mixing *QAM* modulators are available to a stream request of the same stream type, priority should be given to the busiest non-mixing *QAM* modulator because other mixing *QAM* modulators have a higher likelihood of being non-mixing or empty.
- If multiple mixing *QAM* modulators are available to a *SD* or *HD* stream request, priority is given to the busiest mixing *QAM* modulator, because other mixing *QAM* modulators have a higher likelihood of being non-mixing or empty.
- If multiple non-mixing *QAM* modulators are available to a stream request of a different type, that is if a stream request will have to create a new mixing *QAM* modulator, priority is given to the least busy *QAM* modulator, because it has the highest likelihood of becoming non-mixing again.

The algorithm can be expressed mathematically as follows:

1. Identify a set of I, s.t. $Q - q_i \geq r_S$ for $\forall i, i \in I$
   1.1. If *I* is empty, reject the stream request;
2. Identify a subset of J, $J \subseteq I$, s.t. $S_j(q_j) = 3, \ j \in J$;
   2.1. If *J* is empty, go to the next step;
   2.2. If J has multiple elements, select $j^* = \arg \min_{j \in J} Q\text{-}q_j$;
   2.3. If there are multiple j^*, select randomly among j^*;
3. Identify a subset of J, $J \subseteq I$, s.t. $S_j(q_j) = 1, \ j \in J$;
   3.1. If *J* is empty, go to the next step;
   3.2. If J has multiple elements, select j^* randomly;
4. Identify a subset of J, $J \subseteq I$, s.t. $S_j(q_j) = 2, \ j \in J$;

4.1. If J is empty, go to the next step;
4.2. If J has multiple elements, select $j^* = \arg \underset{j \in J}{Min}\ Q\text{-}q_j$
4.3. If there are multiple j^*, select randomly among j^*;

5. Identify a subset of J, $J \subseteq I$, s.t. $S_j(q_j) = 4,\ j \in J$;
5.1. If J has multiple elements, select $j^* = \arg \underset{j \in J}{Max}\ Q\text{-}q_j$;
5.2. If there are multiple j^*, select randomly among j^*;

V. VoD System Modeling

We developed a discrete-event simulation model with which we evaluate the performance of a set of *QAM* allocation algorithms. The model mimics a pool of 256 *QAM* modulators in a *VoD* service group. Session requests are either *SD* or *HD* streams. *SD* and *HD* stream requests have been modeled as independent *Poisson* processes with inter-arrival times exponentially distributed [16]. The aggregate stream request is the combination of the *SD* and *HD* streams, which also follows a *Poisson* process with inter-arrival times exponentially distributed. Equation (3) shows the relationship used to calculate the aggregate *VoD* request inter-arrival rate, λ, based on the number of users in a service group and the aggregated concurrent usage rate during the peak hour.

$$\lambda = (Number\_user) * (Peak\_usage\_rate / 3600) \quad (3)$$

Since the peak-usage rate is defined as the maximum number of stream requests during the peak one hour time period, this parameter was converted from hours into seconds. Equations (4) and (5) represent the *SD* and *HD* mean inter-arrival rates, respectively.

$$\lambda_{SD} = (Percentage\_SD\_streams) * \lambda \quad (4)$$

$$\lambda_{HD} = (Percentage\_HD\_streams) * \lambda \quad (5)$$

Arrival requests have been already classified in section 3 in three genres; mainstream movies, adult content and video browsing, and each of them is characterized by their own unique average duration time. Stream durations for each of these genres have been modeled as independent random variables distributed exponentially. This conclusion has been found from the empirical data shown in Figure 5 Note that this figure shows the aggregate stream duration distribution, thus this is the aggregation of three exponential distributions with different average duration times. Equation (6) shows the aggregate stream duration, μ, based on the weighted average based on the proportions of the genres that make up the streams, where *m* represents the number of stream types (in this case *m*=3).

$$\mu = \sum_{j=1}^{m} (Percentage\_movie\_type_j) * (\phi_j) \quad (6)$$

,where ϕ_j is the average duration of the movie type *j*.

The developed *VoD* discrete-event simulation replicates a real-word stream processing experience as follows:

- Accepted streams are released from the *QAM* modulator when their duration expires.
- Incoming stream requests are compared with the available *QAM* capacity.
  - If the available capacity is insufficient to handle the request, it is denied and the number of sessions rejected count is incremented by one.
  - If the request is accepted, it is placed in one of the available *QAM* modulators. The modulator selection is determined by the stream allocation algorithm that has been configured.

Three allocation algorithms are implemented in the simulation model: *least-loaded*, *most-loaded*, and the proposed *non-mixing* algorithm. In the *most-loaded* algorithm, incoming stream request are assigned on the *QAM* modulator that has the smallest remaining capacity enough to handle the request. In the *least-loaded* algorithm the reverse occurs, arriving requests are assigned to the modulator that has the largest remaining capacity enough to handle the request. In the *non-mixing* algorithm, the available *QAM* capacities are grouped in virtual clusters. Incoming stream is assigned to a *QAM* modulator according to the rules described in section 4.

The model developed accommodates a great number of scenarios depending on the load demand, mixture of *SD* and *HD* streams and other factors. *SD* and *HD* encoding bit-rate have been selected on the model to be 3.75 Mbps and 12.5 Mbps, respectively. Number of subscribers on a service group has been chosen to be 500. Table 1 shows the stream characteristic assumptions used, these values were obtained from current deployments and therefore reflect actual *VoD* patterns.

Stream Characteristics Assumptions			
	% Movies type in SD streams.	% Movies type in SD streams.	Average Duration
Mainstream Movies	40%	57%	2 hours
Adult Movies	30%	-	20 minutes
Browsing stream.	30%	43%	15 minutes

Table 1. Stream Characteristics Assumptions.

VI. RESULTS

The performance of the *VoD QAM* stream allocation algorithms was measured in two ways; calculating the average blocking probability and measuring the efficiency of the allocation algorithm. Algorithm efficiency is calculated using equation (1). Analysis varies the peak usage rate, *SD* and *HD* stream composition percentages.

Figures 6.a, 6.b and 6.c show the blocking probability for the three mentioned algorithms against a range of peak-usage rates for systems with 4, 8 and 12 *QAM* modulators, respectively, for the case where the traffic consists of 90% *SD* streams and 10% *HD* streams. Figures 7.a, 7.b and 7.c show similar results for the case where the traffic consists of 70% *SD* streams and 30% *HD* streams. From these results, it can be seen that *non-mixing* allocation algorithm leads to a lower blocking probability than the other two algorithms at all usage levels. Filling a *QAM* modulator with only one type of stream can guarantee maximum capacity utilization given the modular nature of the streaming bit rates. On the other hand, a mixing *QAM* modulator is likely to have stranded bandwidth that is not sufficient to accommodate an incoming *HD* stream, see equation (1). Figures 6 and 7. also indicate the poor ability of the *least-loaded* algorithm to efficiently allocate streams on a congested mix traffic system. Figure 6.a illustrates that in a *VoD* system consisting of 10% *HD* content with 4 *QAM* modulators, and under 6% peak usage level, the *most-loaded* and the *non-mixing* algorithm lead to a blocking probability close to 0%, while the *least-loaded* algorithm results in a blocking probability close to 4%.

In Figures 8 and 9 system capacity planning and algorithm performance is measured in scenarios where 12 and 16 *QAM* systems are needed. Thus, these systems are prepared to hold up to 450 and 600 Mbps of *VoD* traffic, respectively. It is assumed that these situations are likely to happen when demand of *HD VoD* movies reaches peak values.

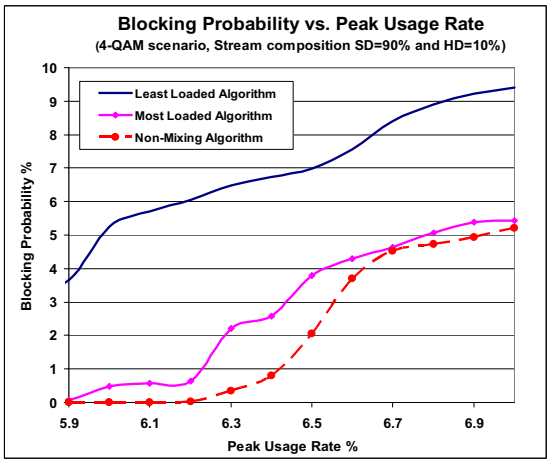

Figure 6.a

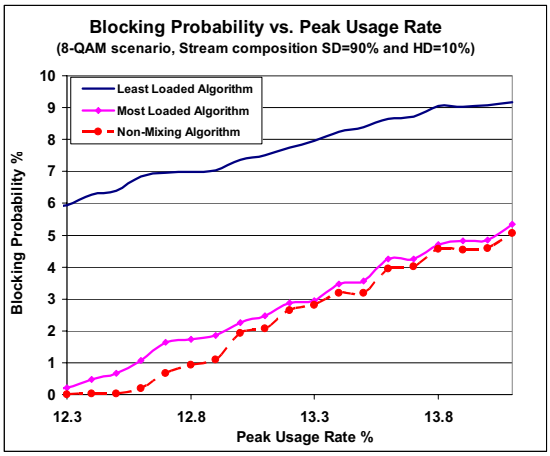

Figure 6.b

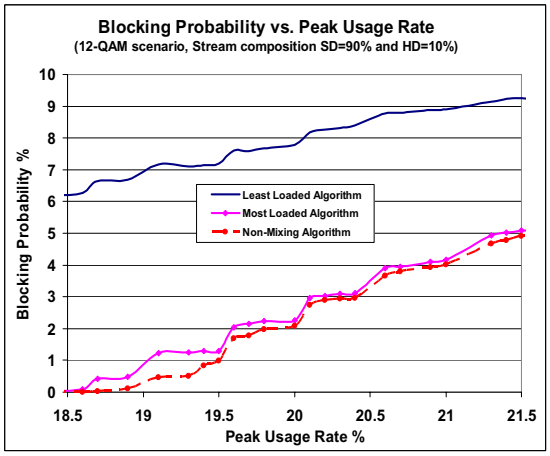

Figure 6.c

Figure 6. Blocking Probability vs. Peak-Usage Rate for Least-Loaded, Most-Loaded and Non-Mixing QAM Allocation Algorithms for 90% SD and 10% HD Streams, (a) 4 QAM scenario, (b) 8 QAM scenario, (c) 12 QAM scenario

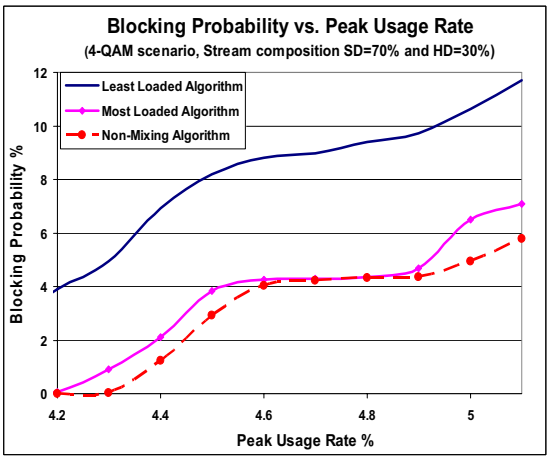

Figure 7.a

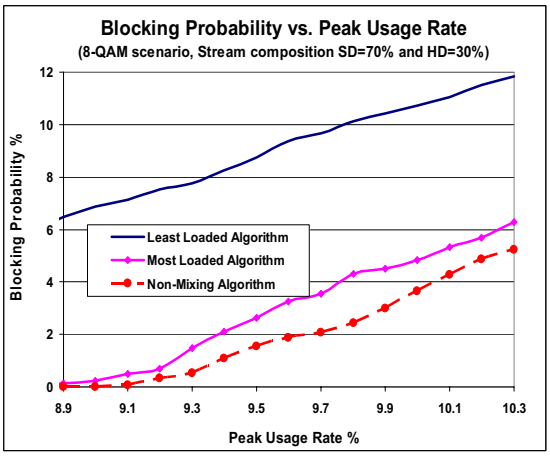

Figure 7.b

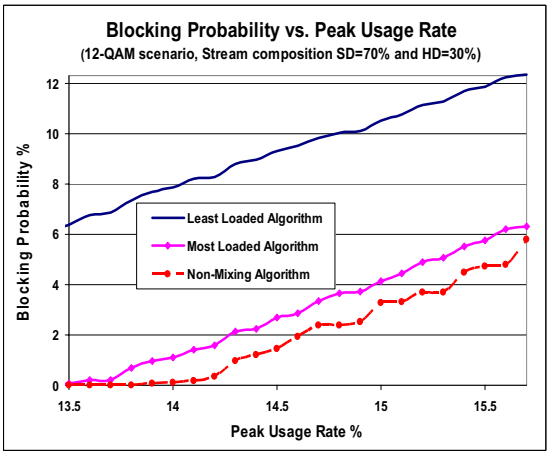

Figure 7.c

Figure 7. Blocking Probability vs. Peak-Usage Rate for Least-Loaded, Most-Loaded and Non-Mixing QAM Allocation Algorithms for 70% SD and 30% HD Streams, (a) 4 QAM scenario, (b) 8 QAM scenario, (c) 12 QAM scenario

Figures 8 and 9 show the maximum peak-usage for 12 and 16 *QAM* systems respectively when subject to blocking rate objectives of 0.3% and 1%[2]. These figures suggest that the maximum peak-usage rate that can be supported to achieve blocking probability objectives decays as the percentage of *HD* streams increases.

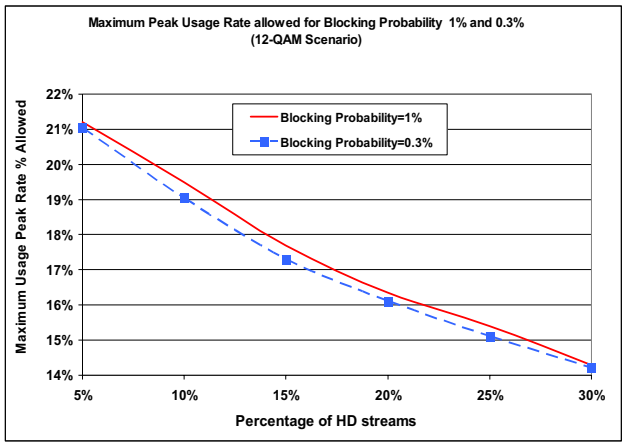

Figure 8. Maximum Peak-Usage Rate Allowed vs. Percentage of HD streams using No-Mixing Algorithm for a 12-QAM system.

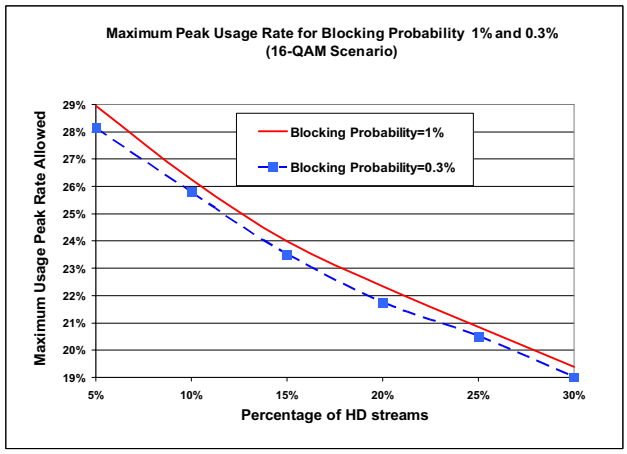

Figure 9. Maximum Peak-Usage Rate Allowed vs. Percentage of HD streams using No-Mixing Algorithm for a 16-QAM system

To demonstrate how results from Figures 8 and 9 might be used for provisioning, assume that a hypothetical *VoD* system will experience a peak-usage rate of 20%. For this load, none of the 4 *QAM* or 8 *QAM* systems for the 500 home service group can support this volume of traffic, regardless of the stream composition (see Figures 6 and 7). Figure 8 suggests that a 12 *QAM* system could handle this load as long as the traffic mix contains less than 7.5% *HD* streams. Figure 9 suggests

[2] Most cable providers would consider a blocking rate of 0.3% acceptable and a 1% rate marginally acceptable.

that a 16 *QAM* system could handle this load for traffic that includes up to 27% *HD* streams.

The blocking probability is determined by two factors: demand load and stream allocation efficiency[3]. The former can only be addressed by increasing system capacity. The latter can be addressed with an improved *QAM* allocation algorithm. Figure 10 shows the average efficiency of the three studied algorithms for a load ranging between 4% to 12% usage rate on a 4 *QAM* modulator system, similar results were obtained for larger *VoD* systems and thus they were omitted. From Figure 8 it can be seen that algorithm efficiency decays as mixture of *SD* and *HD* increases. This result is more apparent with the *least-loaded* algorithm. From here it also can be seen that the proposed *non-mixing* algorithm presents the highest efficiency among the other two algorithms, and its efficiency goes from 90% to 80%, when the percentage of *HD* goes from 5% to 50%, respectively. Under the same circumstances *most-loaded* algorithm efficiency goes from 80% to 70%, and *least-loaded* algorithm from 80% to 20%. These results can be interpreted as, on the worst-case scenario up to 80% of the streams rejected by the *least-loaded* algorithm were due to algorithm inefficiency. In a similar situation up to 30% of the streams rejected by the *most-loaded* algorithm were due to algorithm inefficiency, and up to 20% of the streams rejected by the *non-mixing* algorithm were due to algorithm inefficiency.

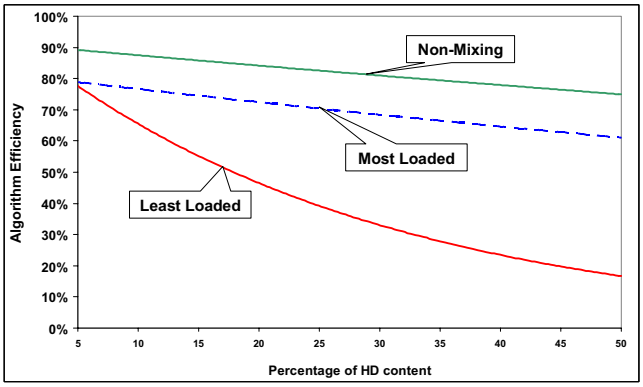

Figure 10. Maximum Peak-Usage Rate Allowed vs. Percentage of HD streams using No-Mixing Algorithm for a 16-QAM system.

The advantages of the proposed *non-mixing* algorithm over the other *least-loaded* and *most-loaded* algorithms are achieved primarily because the algorithm avoids placing *SD* and *HD* streams into the same *QAM* modulator. As a result occurrences of blocking are generally due to load demand.

[3] Inefficient stream rejection refers to a stream denied by a *VoD* system, which its state is represented by equation (1).

VII. Conclusions And Future Work

Our results highlight the importance that *QAM* allocation algorithms have on the efficiency of a *VoD* system in the presence of *SD* and *HD* mix traffic. The two commonly used algorithms, *least-loaded* and *most-loaded*, were designed for *VoD* systems that offer only *SD* streams. However, their respective performance deteriorates as the amount of *HD* streams increases in a *VoD* system. We have shown that *least-loaded* algorithm can result in more than a five fold increase in blocking probability compared to *most-loaded* algorithm when subject to varying levels of *SD* and *HD* stream requests. Our analysis shows that changing to the *non-mixing* algorithm on a 4 *QAM* modulators *VoD* system[4] can support up to 6.2% peak-hour concurrent usage[5] that contains 10% *HD* streams, which is difficult to be accommodated with *most-loaded* or *least-loaded* algorithms (see Figure 6.a). With more *HD* content, *non-mixing* algorithm can generate an average of 4.39% higher allowed peak usage rate over *most-loaded* algorithm. Blocking probability of *VoD* systems using *least-loaded* and *most-loaded* algorithms under mix traffic circumstances are mainly driven by algorithm efficiency rather than by system capacity.

The proposed n*on-mixing* algorithm performs better under all cases of usage level and under all cases of *HD* percentage assumptions used in our study. Its superiority over its predecessors depends primarily on its ability to avoid, to the extent possible, mixing *SD* and *HD* streams. This is driven by several factors, including *SD* and *HD* traffic composition, *SD* and *HD* streaming bit rates, traffic load, stream duration and many others. Results have shown that benefits of the proposed algorithm don't change significantly with the number of *QAM* modulators in the system, only with the stream bit-rate composition percentage. However, the number of *QAM* modulators that are needed to meet a blocking probability objective is highly dependent on the percentage of *HD* streams in the traffic mix.

In future work, we plan to extend our study of *QAM* allocation algorithms to address Switched Digital Broadcast and the *M-CMTS architecture*. Switched Digital Broadcast (*SDB*) is a new method of distributing video programming. It differs from traditional broadcasting in the sense that only programs that are being actively watched by subscribers are broadcasted. Because only a fraction of the channels are ever watched at the same time, a video provider can offer more "virtual channels" than actual program channels and benefit from the statistical gains achieved through over-provisioning. For example, in a cable system with a 750 MHz plant, a traditional broadcast service can broadcast about 250 programs using a mix of analog and digital services. If all

[4] *VoD* systems with 4 *QAM* modulators in a service group is the default setting for many Cable Operators.

[5] 6.2% seems to be a reasonable peak-hour concurrent usage assumption in the near term for many VoD systems in North America that are currently experiencing peak-hour concurrent usage below 5%.

services are digital, the provider could offer 700 programs. *SDB* allows the provider to potentially offer more than 1000 programs. The current *SDB* architecture is evolving towards a common platform where both *SDB* and *VoD* services will be provided together. In the presence of *HD* channels in the *SDB* channel line-up and *HD VoD* streams, the *non-mixing* algorithm becomes important in achieving efficient use of bandwidth.

Another significant development in the cable industry recently is the emergence of the *M-CMTS* initiative as part of the DOCSIS 3.0 specifications [17]. In the *M-CMTS* architecture, the routing and MAC layer processing functions of a current generation *CMTS* is separated out from the *QAM* modulation and demodulation functions, where the former is addressed by the *CMTS*-core and the latter addressed by edge *QAM* modulators (*eQAM*). The edge *QAM* resources can then be shared between both DOCSIS data traffic and video traffic that might include both *SDB* and *VoD* traffic. Initial implementations will likely assign traffic flows to *eQAMs* statically. The better approach requires a dynamic approach where *eQAM* resources are allocated dynamically on a per-session basis. We are working on extensions to the non-mixing algorithm to support *eQAM* allocation in an *M-CMTS* environment where traffic flows consist of traditional DOCSIS, *VoD*, and *SBD* services.

REFERENCES

[1] C. Aggarwal, J. Wolf, P. Yu, *"On Optimal Batching Policies for Video-on-Demand Server"*, ACM International Conference on Multimedia Systems, pp. 253-258, June 1996.

[2] E. Coffman, M. Garey, D. Johnson, *"Approximation Algorithms for Bin Packing: A Survey"*, Approximation Algorithms for NP-hard Problems, pp 46-89, PWS Publishing Company, 1995.

[3] T. Chiueh, C. Lu, *"A Periodic Broadcasting Approach to Video-on-Demand Service"*, Proc. SPIE, vol 2615, pp. 162-169, 1996.

[4] A. Dan , D. Sitaram, P. Shahabuddin, *"Scheduling Policies for an On-demand Video Server with Batching"*, ACM International Conference on Multimedia, pp. 15-23 1994.

[5] A. Dan, D. Sitaram, P. Shahabuddin, *"Dynamic Batching Policies for an On-demand Video Server"*, ACM Multimedia Systems, vol 4, pp. 112-121, 1996.

[6] J. Flint, *"Marketers Should Learn to Stop Worrying and Love the PVR"*, The Wall Street Journal, Oct 2005.

[7] J. Gong, D. Reed, T. Shaw, D. Vivanco and J.Martin,*"VoD QAM Resource Allocation Algorithms"*. International Conferences on Networking 2006, Coimbra, Portugal. May 2006.

[8] L. Gao, J. Kurose, D. Towsley, *"Efficient Schemes for Broadcasting Popular Videos"*, NOSSDAV 98, July 1998.

[9] L. Golubchik, C. Lui, R. Muntz, *"Adaptive Piggybacking: A Novel Technique for Data Sharing in Video-on-Demand Storage Servers"*, ACM Multimedia Systems, vol. 4, no#0, pp 14-55, 1996.

[10] J. Gong, Y. Syed, *"Optimal QAM Assignment in the Presence of Mixed SD and HD Stream"*, NCTA National Show 2005.

[11] G. Hardin, *"Session Resource Management: How to Slice the Pie Allocating Bandwidth for Standard and High-Def VOD"*, Communications Technology Magazine, May 2005..

[12] K. Hua, Y. Cai, S. Sheu, *"Patching: A Multicast Technique for True Video-on-demand"*, IEEE Multimedia, vol. 4, pp. 51-62, 1997.

[13] L. Juhn, L. Tseng, *"Harmonic Broadcasting for Video-on-demand Service"*, IEEE Transactions on Broadcasting, vol 43 pp.268-271, Sept 1997.

[14] W. Liao, V. Li, *"The Split and Merge Protocol for Interactive Video-on-Demand"*, IEEE Multimedia, vol. 4, pp.51-62, 1997.

[15] S. Lau, J. Lui, L. Golubchik *"Merging Video Streams in a Multimedia Storage Server: Complexity and Heuristics"*, Multimedia Systems, vol. 6, no. 1, pp29-42, 1998.

[16] S. Ross, *"Introduction to Probability Models"*, Academic Press, 2003.

[17] DOCSIS® Specifications, Cable Television Laboratories, Inc.(http://www.cablemodem.com/primer/)

Jim Martin (M'88) received the B.S degree in Electrical Engineering from the University of Illinois, Champaign/Urbana, in 1983, and the M.S. degree from Arizona State University, Tempe, in 1989, and the Ph.D. degree from North Carolina State University, in 1999.He is an Assistant Professor with the Department of Computer Science, Clemson University.

He was previously a Senior Consultant for the Gartner Group where he provided consulting in the area of network design and performance management to service providers and companies. Prior to Gartner, he spent 10 years at IBM Network Systems Division where he worked on the research and development of network and network security products. His research interests are in communication networks, network performance management and Internet transport issues.