SIMULATION OF VIDEO-ON-DEMAND TRAFFIC[1]

Mark Cronshaw, Ph.D. & Victoria Okeson
AT&T Broadband Labs, Westminster, CO

Abstract

This paper describes a discrete-event simulation model of VOD traffic in a system with a server and hierarchical network. The model predicts the magnitude and location of blocking as a function of the demand for VOD sessions and their durations. We found that an Erlang-B model may overestimate the blocking probability, since it is based on a statistical equilibrium that may not be reached with time-varying VOD traffic.

The model provides an inexpensive way to explore various network configurations and modulation schemes. As such, it can be very useful for VOD capacity planning.

INTRODUCTION

Interactive cable traffic such as video-on-demand (VOD) content presents fascinating capacity planning issues. For example, VOD systems must be sized so that there is only a small probability of a subscriber being unable to initiate a session. Demand for sessions varies by time of day and day of week. Also, the duration of a VOD session depends on the content selected, the time spent selecting the content, and on the use of VCR-like features such as pause, rewind and fast-forward.

In some respects network traffic associated with video on demand is similar to telephony traffic. Requests for new sessions arrive randomly over time, and the duration of each session is random. Furthermore, the average arrival rate of new requests is not constant over time, but rather there is a peak period during any given day, and this peak will vary on different days. For example, the peak demand for VOD will probably be on a cold and rainy Friday night. However, there is an important difference in the nature of the traffic: the average length of a VOD session is long relative to the duration of the peak period. In contrast, the duration of a typical phone call is short relative to the duration of the peak calling period.

Figure 1 is a stylized representation of VOD traffic. Suppose that the peak period lasts for 3 hours, and that movies are only requested during the peak period, at a uniform rate of 10 per hour. Suppose also, that each VOD session lasts for exactly 2 hours. The solid curve in the figure shows the arrival rate of requests. The dashed curve shows the rate at which movies end: it is the same as the arrival rate curve but delayed by 2 hours.

[1] The authors would like to acknowledge many indispensable conversations with Jim Want and Doug Ike, and to thank Rodger Woock for his review of this paper.

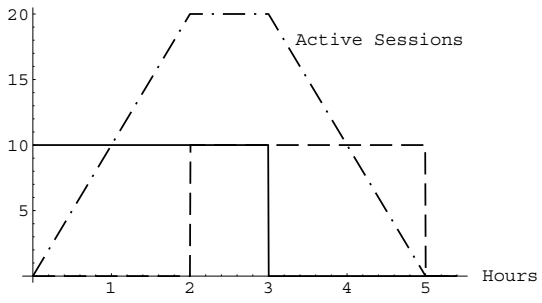

Figure 1 Stylized Video on Demand Traffic

The dash-dotted curve shows the number of sessions that are active, assuming that there are no capacity constraints. During the first two hours the number of active sessions grows at a rate of 10 per hour, reaching a maximum of 20 sessions. Beginning at hour 2, movies finish that had started earlier. Simultaneously, new requests for movies arrive, so there is no change in the number of active sessions between hour 2 and hour 3. The peak period ends at hour 3, after which there is no demand for new sessions. From hour 3 to 5 movies finish which had started previously, and the number of active sessions falls to zero.

By way of contrast, Figure 2 shows the number of active sessions for stylized telephony traffic assuming that the peak period runs for one hour with a uniform arrival rate of 400 calls per hour, and that each phone call lasts for exactly 3 minutes.

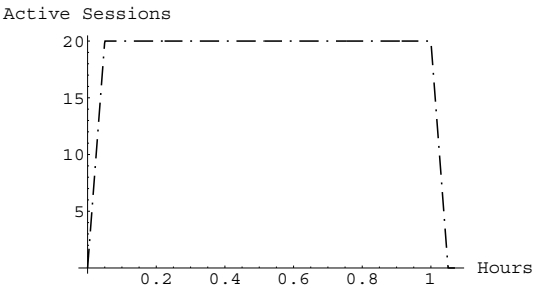

Figure 2 Stylized Telephony Traffic

The number of active sessions (phone calls) over time has a similar shape to that for the VOD traffic, but there is an important qualitative difference. The number of active phone calls is at a sustained maximum level for almost all of the peak period. In contrast, the maximum number of active VOD sessions is at its maximum level for only a small part of the duration of the peak period. This is a direct consequence of the relatively long length of a VOD session compared to the duration of peak demand for VOD.

This qualitative difference has important implications for modeling VOD traffic. The Erlang-B model that predicts blocking of telephony traffic is based on a stochastic equilibrium, which corresponds roughly to the long sustained period of maximum sessions in Figure 2. Under typical conditions, VOD traffic may not reach such a stochastic equilibrium, and hence the Erlang-B model [Wall] may not provide a good approximation of blocking. In fact, if the network is initially empty, then no blocking will occur at the beginning of the peak period. Blocking will only occur after the network has filled to a level where a capacity constraint is reached.

This paper describes a numerical simulation model of VOD traffic that predicts blocking. It shows that

- blocking increases rapidly when the demand for VOD sessions exceeds a threshold,
- blocking is sensitive to the average duration of a VOD session, and
- the Erlang-B model may overestimate blocking.

SIMULATION FRAMEWORK

Network configuration

We modeled a network architecture with a central VOD server that supports several hubs. Video streams are carried from the VOD server to each hub on fiber. Each hub delivers VOD traffic modulated onto RF channels to several nodes. The RF channels are shared among groups of three nodes, called supernodes. Thus there are three potential locations for blocking: at the server, in the fiber between the server and the hub, or in the RF.

The results in this paper are for a network with 1 server, 12 hubs, 217 supernodes and 4 RF channels per supernode. The number of supernodes per hub ranged from 6 to 27. The modulation of each RF channel was 64 QAM, providing 8 digital streams at 3.375 Mbps.

Characteristics of VOD sessions

For the purposes of traffic modeling the relevant characteristic of a VOD session is its duration.[2] This is determined by

- the length of the content,
- the time used while selecting the content,
- the proportion of sessions which are terminated before the end of the content, and
- the time used for VCR-like features such as pause, fast-forward and rewind.

We modeled two distinct types of sessions: long and short, corresponding to feature length films and shorter content such as children's programs. We assumed that the durations of each type of session are normally distributed with means of 120 and 55 minutes respectively, and standard deviations of 25 and 5 minutes respectively.

Demand for VOD sessions

The demand for each type of session varies by day and by time of day. Our simulations cover a single day, which for capacity planning purposes should be the busiest day anticipated. A natural unit for demand is the number of buys per subscriber per month. However, for the purposes of simulation, it is necessary to specify how these buys are distributed over time. We assumed that session requests are distributed according to a Poisson distribution, and that the demand is the same for each supernode. Demand is specified in terms of an average arrival rate of session requests. Different demand can be specified for each type of session.

Demand for each type varies by time of day. For each type the day was divided into four phases. The average arrival rate in each phase was specified as a percentage of the rate during the peak phase, as shown in Table 1. The table shows time based on a 24 hour clock. We assumed that the peak rate was the same for both long and short sessions (although the peak for each occurs at different times, as shown on Table 1.)

[2] We only considered traffic associated with VOD content. There is also out-of-band traffic associated with control of a VOD session, such as ordering, pause, fast-forward and rewind.

	Early		Pre-peak		Peak		Late	
Type	Time	Demand	Time	Demand	Time	Demand	Time	Demand
Long	0300-1700	5%	1700-1900	30%	1900-2200	100%	2200-0300	5%
Short	0300-1400	5%	1400-1600	30%	1600-1900	100%	1900-0300	5%

Table 1 Average arrival rates (as a percentage of peak rate)

We considered peak average arrival rates requests (for each type) between 2 and 5 per hour per node. Based on the time varying demand, a peak rate of 2 requests per hour per node is equivalent to 18.2 requests per node per day. For a node with 500 homes passed and 11% digital penetration, this means that the average number of session requests per digital home in that day is 0.33. If the total demand over all of the other days in the week were ½ of that on this busy day, then the monthly buy rate would be 4*1.5*0.33 = 2 buys per digital household, based on a four week month. The buy rate scales linearly with the peak rate. So a peak rate of 4 requests per hour per node corresponds to 4 buys per month given these assumptions.

Network performance

There are two aspects of blocking that are of interest

- a customer perspective: the percentage of session requests which are not granted, i.e., blocked, because of capacity constraints, and
- a network perspective: the location of blocks in the network, which can be at the server, in the fiber between the server and the hub, or in the RF.

Since demand varies over time, the blocking probability also varies over time. The simulations generated blocking probabilities for each hour. The blocking probabilities reported in this paper are for the hour with the greatest blocking. For most of the runs this was from 9 to 10 pm. For some of the runs it was between 8 and 9 pm. Based on Table 1, the highest average arrival rates occurred from 5 to 7 pm. The highest blocking occurs later than the period of peak demand, since the network is relatively empty at the beginning of the peak request period, but fills as session requests are granted. As shown on Figure 1, it takes some time for the number of active sessions to grow to a point where a capacity limit is reached.

MODELING APPROACH

The model was implemented using a discrete event Monte Carlo simulation written in Microsoft Excel and Visual Basic for Applications. The user has the option to enter up to five network configurations and three demand scenarios. Each unique network configuration – demand scenario pair is called a case, and the user must select the specific cases to study. The model treats each case independently – this mechanism simply allows for more efficient processing of multiple cases.

The model is designed to simulate a peak day in the system. Movie demands are defined for a 24-hour day, starting at 3 am. The model considers 3 am to 7 am

to be a warm-up period to initialize the network, so no statistics are collected during this time. At 7 am, the model begins collecting statistics for the next 20 hours. Two types of statistics are collected: blocked requests and active sessions.

Because this is a Monte Carlo simulation, multiple trials must be run for each case. At the beginning of each trial, movie requests are generated according to a Poisson process for each type of movie at each node in the network. These requests are then sorted by arrival time and processed in turn. Processing a request has three steps: clearing out completed movies, assessing the network for capacity and reacting to the result of the assessment. First, the entire network is checked to see if any movies have completed since the last movie request was processed. If so, the capacity counters in each network element are updated to reflect the newly released capacity. Next, the supernode at which the movie arrives is identified. This supernode, the fiber that serves it, and the VOD server are all checked for available capacity. If sufficient capacity exists in all of these network elements, then the movie request is granted and the duration of the movie is obtained according to a predefined statistical distribution. The movie is recorded in the network and capacity counters are updated. If there is not enough capacity, then the location of the block is recorded. This process repeats until all of the requests are handled. The active sessions data is recorded by evaluating the state of the network at user-specified intervals throughout the course of the trial.

Once all of the trials are completed, summary reports, which include the averages over all of the trials as well as individual trial results, are output for later analysis.[3]

Validation

Validating the model is essential to insure that it accurately represents the system under consideration. For this model, an analytic benchmark is available for validation. A VOD system resembles an M/M/m/m queue, which has multiple servers that block requests if no server is idle when the requests arrive. Applying this to the VOD simulation model is something of a generalization, since a VOD system is a hierarchical network. However, for the network we modeled, the vast majority of the blocking occurs at the supernode level, so it is a reasonable approximation. For an M/M/m/m queue, the blocking probability (i.e. likelihood the request is lost) is simply an Erlang-B function [Tanner]. We compared the analytic benchmark to the VOD simulation model for a system with only one movie type. Table 2 illustrates the comparison between the analytical blocking and the VOD simulation model blocking for three different cases, varying only the movie request arrival rates. The arrival rates were kept constant over time in the simulations to satisfy the stochastic equilibrium assumption implicit in the Erlang-B model. Note that our simulation model shows excellent agreement with the analytic results.

[3] Since demand is random, the actual highest level of blocking may be higher or lower than the average. There are other criteria besides average blocking that might be used for network design, such as a 5% chance that more than 1% of session requests are blocked.

Average session request arrival rate per supernode per hour	Analytical blocking	VOD simulation model blocking	% of blocking occurring at supernodes
15.9	4.8%	5.0%	100%
17.7	8.8%	9.2%	100%
19.2	12.9%	13.4%	99.8%

Table 2 Model Validation

MODELING RESULTS

Table 3 shows how the blocking probability varies with the demand for VOD sessions. With time varying demand, there is no blocking if the peak average arrival rate is 2 requests per type per node per hour (or lower). On average, 3% of requests are blocked if the peak rate is 4 requests per type per node per hour.[4] But if the peak rate rises to 5, then blocking reaches an unacceptable level of 12.7% of session requests.

Peak avg. arrival rate per type per node per hour	Demand	
	Time varying	Steady
2	0.0%	0.05%
3	0.15%	4.2%
4	3.0%	18.7%
5	12.7%	35.2%

Table 3 Blocking probabilities (highest hour)

The table also shows the blocking probabilities if demand is constant at the peak rate over time, rather than time-varying. The blocking is significantly higher with steady demand. This shows that the Erlang-B model can overestimate blocking if the demand for VOD sessions varies over time.

Table 4 shows the location of blocking in the network, expressed as a percentage of all blocks. All blocking occurred in the RF, except for a small amount of blocking at the server when the peak average arrival rate was 5.[5]

Peak avg. arrival rate per type per node per hour	Location		
	Server	Fiber	RF
2	none		
3	0%	0%	100%
4	0%	0%	100%
5	0.8%	0%	99.2%

Table 4 Blocking locations

The blocking probability is sensitive to the average duration of a VOD session. If the average duration of the long

[4] Blocked requests were not queued in the simulations. They were simply not granted. In reality a blocked request would probably lead a subscriber to submit another request immediately. This behavior was not modeled.

[5] It is possible for there to be insufficient capacity simultaneously in several parts of the network. We attributed blocking to the lowest level of the network where it occurred. For example, if both the RF and the server were full when a new request arrived, then the block would be attributed to the RF and not to the server.

sessions is reduced from 120 to 110 minutes, and that of the short sessions from 55 to 50 minutes, then blocking is approximately halved. With time varying demand and peak rates of 4 and 5, the blocking probabilities are 1.3% and 7.4% respectively, compared to the values of 3.0% and 12.7% in Table 3. A drop in blocking would be expected due to the reduction in offered load. Yet the magnitude of the drop is large. This sensitivity demonstrates the potential benefit from reducing the average VOD session length, perhaps by taking steps to reduce session time involved in movie selections.

We also did simulations to explore a lower standard deviation of the session duration. Such reduction had only a small impact on the blocking probability.

CONCLUSIONS

Simulation is a useful tool to assist in capacity sizing. It is flexible, which makes it possible to evaluate the anticipated performance of many possible network configurations, before making significant investments.

The discipline of building a simulation model brought to light many crucial operational issues, such as

- the importance of the average duration of a VOD session, and
- the time-varying nature of demand and blocking.

It would be highly desirable to collect network data to calibrate the model. The model predictions do match the analytical Erlang-B model. But the Erlang-B model is not valid for time-varying traffic on an actual hierarchical network. Calibration with actual VOD session traffic would increase confidence in the predictions of the simulation model. After such calibration, the model can generate significant money-saving insights on how to achieve a balance between the costs of adding VOD capacity and the costs of having subscribers frustrated by blocked VOD requests.

REFERENCES

Tanner, Mike. *Practical Queueing Analysis*. McGraw-Hill, Berkshire, 1995.

Wall, Bill. "Traffic Engineering for Video-On-Demand Systems", Scientific Atlanta, 2000.

Contact info
Mark Cronshaw
mbcronshaw@broadband.att.com

Victoria Okeson
vokeson@broadband.att.com

10355 Westmoor Dr., Suite 100
Westminster, CO 80021