A QUANTITATIVE APPROACH TO MULTI-TIERED STORAGE AND STREAMING FOR VOD SERVERS

Robert Duzett
C-COR Inc.

Abstract

The uneven popularity of titles within a VOD content library, as expressed by its content demand profile, opens the door to improved efficiencies from multi-tiered storage architectures. Quantitative methods are shown for addressing the design and provisioning of such architectures. A single constant 'r' can describe the content demand profile. Analysis using 'r', along with storage media pricing ratios 'p' and system streaming capacity 'S' results in a model for optimizing and balancing the mix of storage across multiple storage tiers. This analysis is applied to both 2-tier and 3-tier architectures, and is extended to consider the effects of future trends in content profiles, streaming demand, and storage media.

VOD Demand Profile

Data gathered from VOD deployments show that there is a steep curve representing the distribution of content to streaming demand ("hot" content feeds a lot more of the streams). One very useful way to represent this Content Demand Profile is to plot a curve mapping cumulative library content hours (sorted in order of popularity) on the X-axis to cumulative fraction of total streams on the Y-axis. Thus, for a selected number of the most popular content hours you can look up what fraction of the streams are driven by that content (see Figure 1).

We have observed repeatedly from actual VOD profile data that the curve representing this relationship can be mapped closely to an exponential formula: 1-e^(-rc), where 'c' represents the number of content hours along the sorted content list and r is a ratio characterizing the steepness of the curve. In this way, a single constant, 'r', can represent very closely the content demand profile for a given time window of VOD usage for a given content library ('r' characterizes the library and the demand it generates). A higher value of 'r' means a steeper curve. As will be shown hereafter, we can use 'r' to help calculate other useful relationships based on the demand profile. The graph of Figure 1 shows a demand profile from an actual deployment as well as a curve fitted closely to it using the formula 1-e^(-rc). For this data, r = 0.007.

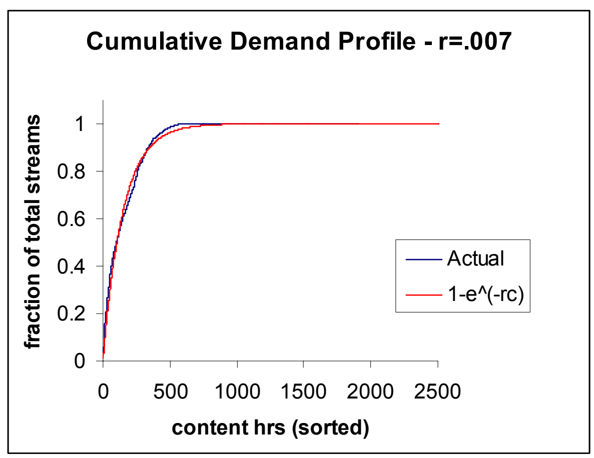

Figure 1

Tiered Caching Principles

The nonlinear nature of the VOD stream-from-content demand profile, much like other resource demand profiles in general computing, suggests the opportunity for improved efficiencies via caching. Placing a small amount of "hot" content into a small cache of expensive but fast memory could potentially make the whole system more efficient.

Today's leading edge VOD servers, for example the C-COR n5 server, make use of this caching opportunity to create two or more "tiers" of storage based on different technologies or different performance/density points – e.g. RAM, Flash, fast disk, slow disk, etc. The basic assumption is that for each higher storage tier, storage gets cheaper while bandwidth gets more expensive. Thus the lowest tier is likely the fastest but the most-expensive per byte (& the least dense).

At each storage tier, we are trading off additional STORAGE (cached, copied) at this tier to replace STREAMING from the next higher tier.

The ideal model for making this tradeoff allows all tiers unrestricted scaling for streaming or storage. In practice, however, there are physical and architectural limitations. Also, the requirements of a VOD deployment, as mapped to a given storage tier or technology, will be unlikely to yield a perfect balance of Bandwidth and Storage. The storage will be either "content-limited" or "streaming-limited".

"Content-limited" means that the bandwidth available from the required content exceeds the bandwidth required for streaming – in other words, storage is being added for content, not streaming. "Streaming-limited" means that the storage capacity from the required streaming storage exceeds the storage required for the current content library – in other words, storage is being added for streaming, not content.

This tension between streaming and content requirements can lead to inefficiencies. For example, consider a content-limited situation in which a centrally-located storage system could provide all necessary streaming bandwidth but for limited transport bandwidth to the edge. In this case, the content must be pushed out to the edge and duplicated at various headends. On the other hand, consider a streaming-limited case in which every headend has more than sufficient content storage because of streaming bandwidth requirements placed on the storage. In this case, an excessive amount of storage is paid for but a portion goes unused.

Tiered storage can ameliorate these kinds of imbalances and make overall operations more efficient.

<u>A 2-tier Caching Model: Disk vs. DRAM</u>

The Hard disk drive is a commodity high-DENSITY storage. DRAM is a commodity high-BANDWIDTH storage. The ideal storage would have disk density and DRAM bandwidth. Based on today's pricing:

Density-per-$ ratio of disk:DRAM = 60:1
Bandwidth-per-$ ratio of disk:DRAM = 1:30

We can somewhat balance these two ratios by storing the content library on hard disks, while caching hot content in DRAM.

Until recently, RAM caching was not economical ($cost/density was too high). However, because the RAM density growth trend (~40% per year) is much steeper than the disk performance growth trend (12-15% per year), RAM caching will prove more and more cost-effective as time goes on.

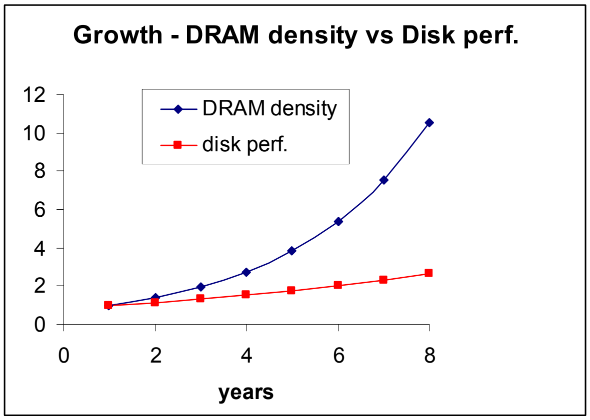

Figure 2

If we're stream-limited, RAM cache provides an opportunity to remove disks or increase server performance (to the limit of the platform). If we're content-limited or transport-limited, RAM cache provides an opportunity to radically reduce content duplication at the edge servers while centralizing the complete content library on high-density disks.

How Much Cache is Cost-Effective?

To provision the storage subsystems of a large VOD server for optimum cost-performance, one can expect the hottest titles will be stored in, and streamed from, cache while the rest of the content will be streamed from higher storage tiers. But, what is the optimum balance of cache content and higher-tier streams? That is, what is the optimal cache size?

If one were to attempt to get ALL streaming from RAM, then ALL the content would have to be stored in RAM, which would obviously be too large and expensive for even moderate content libraries. So, one must attempt to achieve a reasonable portion of streaming, as cost-effectively as possible, from RAM.

The demand profile curve indicates how much CONTENT must be cached to achieve a given HIT RATIO. As content is added incrementally to the cache, the hit ratio rises and incrementally more streams can be fed from the cache. So, the hit ratio, and therefore the CONTENT SIZE, of the cache, NOT the BANDWIDTH CAPACITY of the cache, determines how many streams it can feed (assuming of course sufficient bandwidth capacity from the cache).

We can create a cumulative hours vs cumulative streams graph by multiplying the demand profile curve (cumulative hours vs cumulative *fraction* of streams) by the total number of streams for the system (for example, 8000 streams). See Figure 3. This graph shows how many streams will be sourced by the cache for any given size of cache. In effect, for a given cache size, the streams underneath the curve come from the cache while the streams above the curve come from disk.

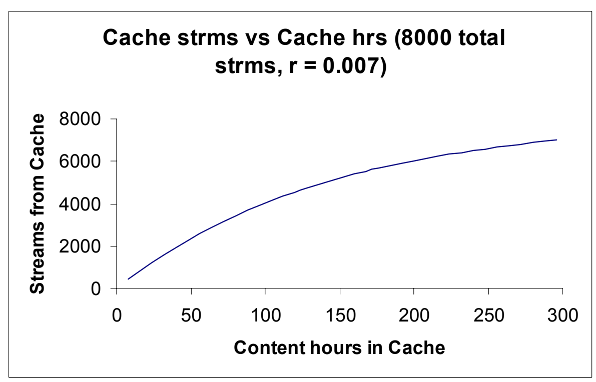

Figure 3

Considering this graph, adding cache is cost-effective as long as the incremental cost for cache content is LESS than the corresponding decremental cost from streams displaced from the next higher storage tier. Therefore, adding cache is cost-effective as long as the slope of the curve is GREATER than the pricing ratio (p):

p = \$-per-hour for cache / \$-per-stream for higher-tier

So, cache is cost-effective while the slope of the streams vs hours curve is greater than p. The slope is the derivative of the curve, so we have:

Sre^(-rc) > p (S=total streams, c=cache content hrs)

Solving for c, we can determine the maximum cache size that is cost-effective for any system of 'S' streams and a content library with demand profile 'r', given a pricing ratio 'p' between two storage tiers.

Cache_max = ln(rS/p)/r (in hours)
Hit-ratio_max = 1-p/(rS)

For example, graphing hit-ratio_max vs total streams for p=47 and r=0.007 shows that caching is cost-effective for VOD servers larger than about 7000 streams; and a cost-effective hit-ratio of about 50% is reached with a server size of about 13000 streams. That 50% hit ratio corresponds to about 100 hours of content in cache (160 GB). See figures 4 and 5.

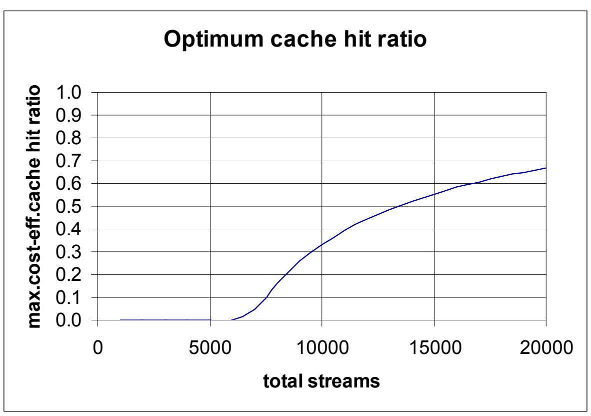

Figure 4

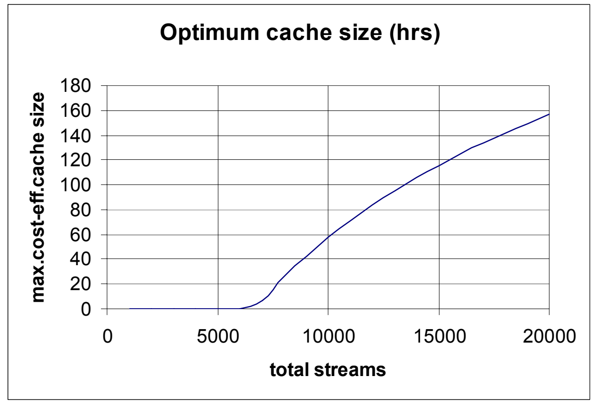

Figure 5

So, we have determined four major factors that determine the appropriateness and size of a given caching storage tier:

1) the demand PROFILE of the content library, represented by 'r'.

2) the PRICING ratio 'p'.

3) the total #STREAMS of the system

4) the total CONTENT library size (this dictates the size of the highest storage tier and thus the minimum bandwidth that may be streamed from it, ie whether we are content-limited)

We have also determined that RAM is not a cost-effective way of achieving streaming bandwidth for systems smaller than 7000 streams, because disk bandwidth is cheaper for those systems; and that even above 7000 streams cost-effectiveness places limits on the amount of RAM that is desirable to displace disks for streaming bandwidth.

For this reason, VOD servers should be designed such that the disk vs RAM tradeoff can be made in a balanced and flexible way based on the size of system to be deployed. The server architecture should not place unreasonable restrictions on storage tier provisioning. For example, this is why the C-COR n5 VOD server was specifically designed to make both disk and RAM independently scalable, and thus balance a wide range of potential needs from both disk and RAM.

A 3-tier Caching Model

This 2-tier storage model (tier-1 is the cache, tier-2 is the disk array) can be extended to a 3-tier storage architecture. For analysis purposes we will consider here a 3-tiered global architecture in which all the storage on all tiers is globally accessible by all streams of the system. Tier 1 is the fastest storage; tier 3 is the densest storage. Tiers 1 and 2 cache content from the global library at tier3.

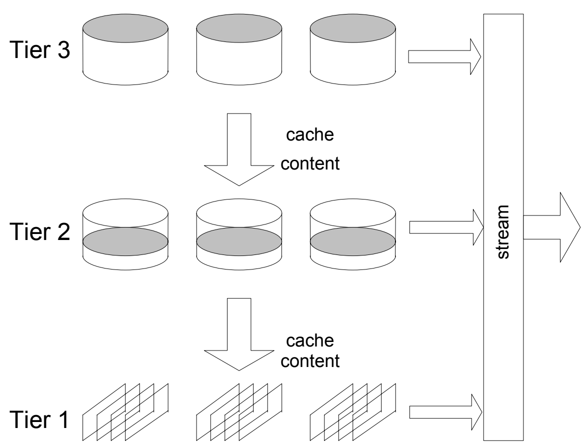

Figure 6

We consider all practical/reasonable storage technologies – e.g. RAM, SCSI, SATA – and various devices from each. We consider costs for each device ($-per-GB, $-per-Mbps) and pricing ratios (p) between tier candidates, and then choose 3 reasonably-priced devices that reflect increasing streaming costs and decreasing storage costs. See Table 1.

Tier#:	**1**	**2**	**3**
	DRAM	**SCSI**	**SATA**
	DDR266	15K73	7.2K320
p=$/hr_this / $/strm_nxt	48.5	1.8	
$/hr_this / $/hr_nxt	22.3	8.8	
$/strm_nxt / $/strm_this	18.2	1.2	
Raw unit capacity (GB)	2.1	73.0	320.0
net hrs/unit	1.3	29.2	128.1
net strms/unit	1120	64	27

Table 1

For a given content library size and profile (‘r’), we analyze the cost-effective tier boundaries for various stream counts, keeping in consideration both content-limited and streaming-limited effects.

This analysis allows us to determine the optimum storage balance and hit ratios for the 3 storage tiers, for systems of any total stream count. We can also determine the optimum storage cost, and compare this with various 2-tier and 1-tier storage technologies.

Figure 7 shows the optimum storage mix for a 3-tiered 5000-hour system, across a wide range of system sizes (characterized by maximum stream counts, on the x axis), using a content demand profile of r=.007. Note that tier-3 is used to archive the entire library and is designed to be content-limited, while tiers 2 and 1 are used to provide the necessary streaming bandwidth. Tier-2 storage is the most cost-effective streaming storage up to 7000 streams and ramps up to that point, beyond which tier-1 takes over. Figure 8 shows the corresponding hit ratios for the three tiers.

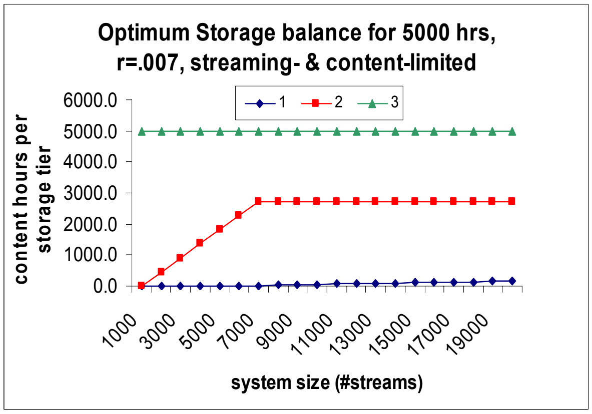

Figure 7

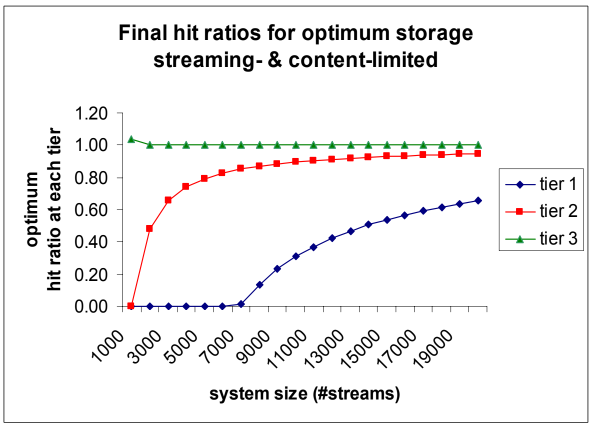

Figure 8

Figure 9 shows the total per-stream storage cost for this 3-tier, 5000 hour architecture. For comparison, it also shows the 1-tier equivalent, which applies the most cost-effective disk drive to meet content and streaming requirements for each stream-count point on the x-axis. Note the cost-savings achieved by architecting a 3-tier storage hierarchy. This difference widens further when you consider the real-world physical constraints generally faced by a single-tier storage system. While this analysis assumed unlimited scalability, the number of drives that can realistically be supported often forces the use of faster, more-expensive drives to meet bandwidth requirements within the allotted density.

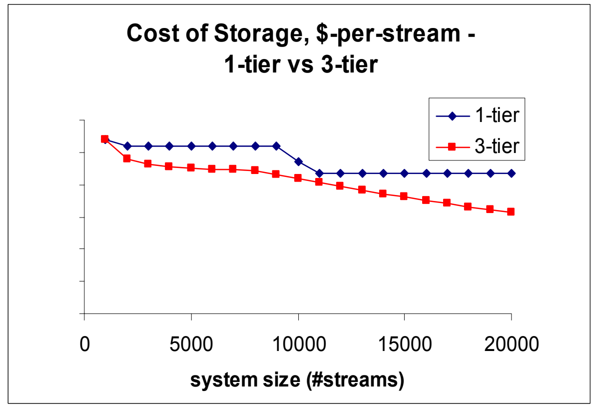

Figure 9

Other Tiered Architectures

Tiered architectures other than the global hierarchical architecture can also be analayzed. Among these are:

- central global library (t3) with distributed isolated local servers (t2 & t1 caches) (=edge servers);
- central global library (t3) with distributed switched local servers (t2 & t1 cache content distributed among servers);
- central global library (t3) + site global cache (t2) + distributed server local cache (t1); and
- content distributed optimally among 3 tiers – no caching, just shuttling among tiers (monolithic or switched servers).

as well as others. All of these are derivatives of the central tier-3 model.

Practical Considerations & Further Quantitative Analysis

All storage tiers, devices, and technologies have scaling limitations and overheads. These include mechanical and packaging limits, controller design tolerances, interconnect bandwidth &

latency limits, transport capacities, etc. In addition, the costs, bandwidths, and capacities of server platforms, storage systems, and other infrastructure can have a significant effect on the final cost-effectiveness of any tiered architecture, beyond the storage devices themselves. Many of these limitations, overheads, costs, and capacities can be built into a multi-tiered model such that their effects can be felt and accounted for in the architectural analysis.

For example, there exist server architectures today that narrowly limit the interconnect bandwidth coming from the disk array while maintaining highly-scalable bandwidth from RAM. This unfortunate bottleneck restricting disk-sourced streaming creates a severe imbalance in the architecture and a consequent cost premium, as shown in figure 10. Note that the optimum balance of disk and cache is broken by an architectural limit of 1000 streams from disk, which causes the storage costs of small systems with this disk bottleneck to be more than double those of systems that are well-balanced. A significant premium is paid even for large configurations.

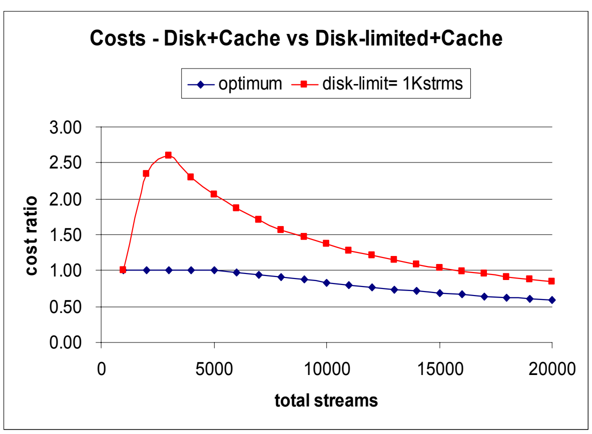

Figure 10

A successful tiered-storage server architecture will maximize the scalability and flexibility of each storage tier, and the storage system as a whole, within reasonably anticipated ranges, so that a wide variety of VOD deployments can be configured as close as possible to the optimum balance of tiered resources using the most appropriate storage media and technologies.

Other concerns, not directly affecting capital economics and difficult or impossible to include in a mathematical model, could alter architectural decisions. These include reliability, operational costs & considerations, interoperability, legacy, etc. and must be duly considered in all architectural design and development.

Tiered Caching Trends over the Next 3-5 yrs – Technology & Economics

It has been shown above that caching effectiveness hinges on content library characteristics (size and profile ‘r’); the price ratios (‘p’) between various storage types; and system stream counts (‘S’). It is very interesting to consider trends and expectations for these parameters over the next few years to see where caching and tiered storage may take us in the future. By extrapolating historical and predictive numbers for such things as device costs, bandwidths, and capacities; library sizes and content mixes; bit rates, take rates, penetration ratios, and HD ratios; effects of Moore’s law on platform capacities; etc. and then applying a practical best-case and worst-case range to each of these, one can build a model that looks at caching effectiveness and/or storage costs over several years as well as its sensitivity to particular parameters or groups of parameters. And this can be done for various tiered or non-tiered architectures. Below are graphs for a 2-tiered, disk+RAM, caching model (figures 11 and 12):

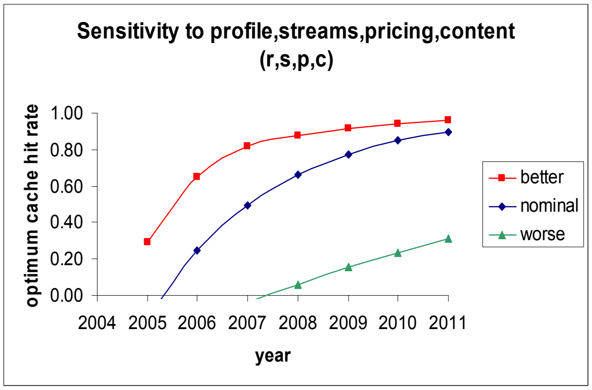

Figure 11

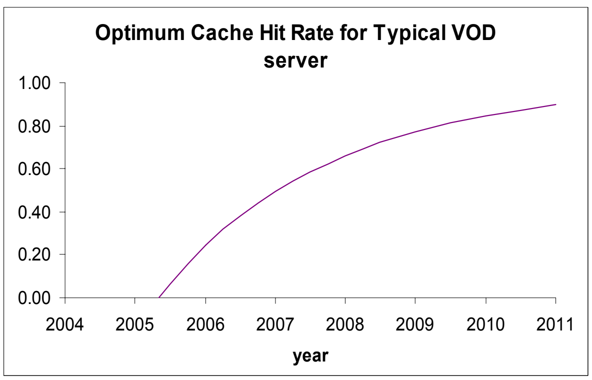

Figure 12

It is also important to consider the effect of new technologies on future tiered storage architectures. For example, NAND Flash has become denser and cheaper than RAM, as well as, in the right format, faster than disk. It therefore has the potential to become a cost-effective storage tier for video systems.

Measurement and Predictability in a Hierarchical Age

The addition of new storage tiers to the VOD architecture creates new complexities for the designer as well as the intergrator and system manager. The storage and streaming requirements of a VOD system or deployment now invoke multi-dimensional parameters. Storage is no longer determined by simple questions of "how much?" and "how fast?", but also by "which tier?" and "what hit-ratio?", etc.

This paper has offered some basic tools for identifying and talking about the key parameters that characterize a multi-tiered architecture. An integrator can anticipate and design for a required range of 'r', 'p', 'S', and 'C' values for a specific deployment or for a general architecture over many deployments. Architects and managers can characterize content demand profiles with a simple 'r' number so they can be discussed quantitatively.

As libraries grow and become more diverse, 'r' values will undoubtedly fall, though not linearly. At any given time, however, one may discuss how 'r' values change over the years, weeks, or months; for time of day; and across content management policies and marketing approaches. Variations in 'r' can also characterize the effect of the response times of cache algorithms being implemented or studied.

In other words, use 'r' as a measure not just of static content libraries but of server and cache efficiencies, and content management efficiencies and marketing efficiencies. Which titles are marketed and how they are marketed can make a big impact on the content demand profile and thus on storage and transport and streaming efficiencies. Measuring hit rates at various tier boundaries of a specific system tells you things about that system only, but then translating those hit rates into an 'r' value now describes the overall profile of access demand for the content library from the attached subscribers.

An integrator or system architect will specify and test a VOD deployment against an expected range of content profiles, as characterized by ‘r’ values; and against a range of expected pricing trends for various storage media, as measured by ‘p’ values; and against a range of system sizes, as specified by ‘S’ stream counts.

CONCLUSIONS

- A Content Demand Profile can be fairly characterized and quantified with a single number, which can then be used to drive a tiered caching model

- A tiered caching model can accurately model effects of both content- and streaming-limited cases.
- A caching model can be used to find an optimum cost-effective balance of 2 or 3 storage tiers.

- Cache effectiveness is determined by 1) the content profile, 2) total streams, 3) relative price-performance of storage devices, and 4) content library size.

- DRAM will be an increasingly cost-effective caching technology for VOD; current economics support it for medium-to-large systems.

- A successful tiered storage architecture will strive for balance, flexibility, and scalability across all tiers, so that the resulting VOD system can be cost-effectively and efficiently applied to a wide range of deployment opportunities.

Contact information:
Robert Duzett
rduzett@c-cor.com
503-690-6305