LARGE-SCALE ON-DEMAND DELIVERY ARCHITECTURES – TOWARDS AN EVERYTHING-ON-DEMAND FUTURE

Cliff Mercer, Ph.D., Director of Technology
Satish Menon, Ph.D., Chief Technical Officer
Kasenna, Inc.

Abstract

Pioneering technologists and business leaders have been driving our industry toward new ways of watching and interacting with our televisions, that is, toward an "everything-on-demand" future. Accomplishing this lofty goal requires the cost-effective implementation of "extreme narrowcasting" techniques, which in turn demand coordinated use of network technologies, powerful server architectures, and intelligent software design.

The early Video-On-Demand (VOD) solutions of the past with their proprietary interconnection technologies and custom hardware designs fail to scale cost-effectively with the new demands for flexible integration with new services, delivery performance and storage capacity. Advances in commercial off-the-shelf computing and networking products along with smart software design meet these current and future challenges cost-effectively and without the obsolescence problem experienced with historical solutions.

INTRODUCTION

Since the early 1990s, technologists and visionary business leaders from a group of pioneering companies in the content, computer, networking, consumer electronic, broadcast and cable TV businesses have been working to change the way we watch and interact with our television. They have been working towards an *everything-on-demand (EOD)* future:

- *News-on-demand*: news you want to watch, when you want to watch it – without being limited to rigid broadcast times at 6am, 6pm, and 10pm.
- *Music-on-demand*: any tune ever recorded from Hillary Duff's latest to Rachmaninoff.
- *Education-on-demand*: listen to Feynmann's physics lectures at your convenience.
- *Sports-on-demand*: re-live the memories of Pele's most famous goal in the 1962 World Cup.

Realizing such a vision requires commercially viable implementation of delivery techniques commonly referred to as "extreme narrowcasting" – the ability to deliver a unique stream of rich media (such as audio or video) from the content source (typically video servers) to an end user's TV, PC, mobile phone, or PDA. Extreme narrowcasting requires novel system designs, such as distributed network topologies, powerful server architectures, and intelligent software design, that can take advantage of "localities of reference," predict future usage patterns, manage a variety of bandwidth requirements and resource needs, etc. More importantly, these designs must make for a cost-effective system, allowing operators to deploy these services without breaking the bank and allowing them to realize returns on their investment quickly.

In the past, proprietary servers, proprietary interconnection technologies, and custom designs were employed to solve these problems of extreme scale in an initial application area: delivering movies over cable plants to TVs. However, the advances in mainstream server technologies and intelligent software designs can solve these problems much more cost-effectively, while avoiding obsolescence associated with proprietary technologies. Furthermore, these new technologies enable the delivery of a plethora of additional video and rich media applications and services to a broader range of end-user devices as the full vision of EOD comes within reach.

In this paper, we highlight the technical and economical challenges involved in the deployment of an extreme narrowcasting network. We begin with a discussion of modern trends in high-performance computing, which leverage commercial off-the-shelf (COTS) hardware combined with intelligent software management to scale to very high performance levels. Next, we relate those techniques for high-performance computing to the particular application area of concern: VOD and other video and rich media delivery applications. We discuss typical growth of a narrowcasting video network from a centralized system to a potentially decentralized system as the number of subscribers grows and the amount of content in the system increases. We introduce the concept of "hierarchical storage" – an implementation of storage systems (from RAMs to disks) to address the requirements inherent in scaling up the number of subscribers and amount of storage. Intelligent software techniques help to solve these problems of extreme scale within the entire video network. Finally, we offer quantitative comparison between common architectures for VOD and EOD deployment.

Modern Techniques for High Performance

COTS cluster systems have become the most cost-effective way to satisfy high-performance computing requirements. Real-time media streaming, VOD streaming, and on-demand delivery of rich media require the best and most cost-effective high performance computing techniques available. According to Thomas Sterling of the California Institute of Technology, "Cluster systems are exhibiting the greatest rate in growth of any class of parallel computer and may dominate high performance computing in the near future." [Sterling 2001]

We define a COTS cluster as a collection of server-class computers built completely from commercial off-the-shelf components and which themselves are built using commodity off-the-shelf chips and components. The interconnecting network technology must use COTS components as well, the most cost-effective technology right now being Gigabit Ethernet.

The latest COTS components have such increased price performance that using high-performance COTS servers connected by GigE interfaces yields a very high performance computing system. For example, a COTS cluster might consist of 10 commonly available dual-Xeon (> 2.2 GHz) servers, each with 2 GB of DRAM, up to 16 disk drives at 146 GB each, and up to three Gigabit Ethernet NICs. Such a cluster would have:

- Processing capacity of 20 Xeon processors,
- 20 GB of DRAM,
- 23 TB of disk storage and
- 30 Gbps of networking capacity.

Moore's law has continually increased the price performance of COTS computing components for the past 40 years. Intel and others spend billions of dollars on research and development to drive this technology and the price performance forward, and COTS clusters benefit directly in terms of price performance as a result.

The Fall of Proprietary Hardware

Hardware designed for specialized applications use the best currently available technology at design time, but by the time the hardware design goes through prototyping, bug-fixing, revision and finally manufacture, the design is already outdated. Even before the product ships, it is behind the technology curve. The specialized hardware design itself is costly and negatively impacts the solution's price performance. Furthermore, maintaining a hardware design for a specialized market is extremely costly as well. Bug fixes and upgrades required for any hardware design must be done by a small team employed by the single vendor. In contrast, COTS computing systems leverage the design expertise of many competing companies and benefit from the significant investment these companies make for the broader computer equipment market.

Over time the hardware design ages quickly. Within a year or two after initial product shipments, the hardware technology is outdated, and the price performance is significantly behind current technology and products. Moore's Law cannot be exploited effectively to drive down costs over time for a proprietary hardware design targeted toward a specialized application.

COTS Clusters for Price Performance

Vendors that bring COTS servers to market are under constant pressure to leverage improvements of new components in their server product lines. Vendors of motherboards, NICs, and complete servers spend billions of dollars in research and development to incorporate the latest underlying hardware technologies and provide standards-based products to the market as fast as possible. COTS NIC vendors are under similar constant pressure to improve the price performance of their products. As a result, COTS servers and network products track Moore's law very closely.

Commodity cluster techniques dominate specialized hardware designs in terms of price/performance. In fact, a cluster of COTS servers connected by fast Ethernet won the Gordon Bell Prize for Price/Performance awarded in conjunction with the Supercomputing Conference in 1997 [Karp 1998]. COTS clusters enable a "trickle up" effect, whereby technological improvements and better price performance characteristics developed and perfected for standalone applications in the mainstream computer industry are leveraged to provide significant advantages in the area of high-performance computing applications such as real-time, high-bandwidth video, and audio streaming. Figure 1 shows the COTS value chain, which starts with processors, chipsets, and other semiconductor products at the bottom. From those parts, components are developed and those move up into COTS servers, which can then be combined, to form COTS clusters.

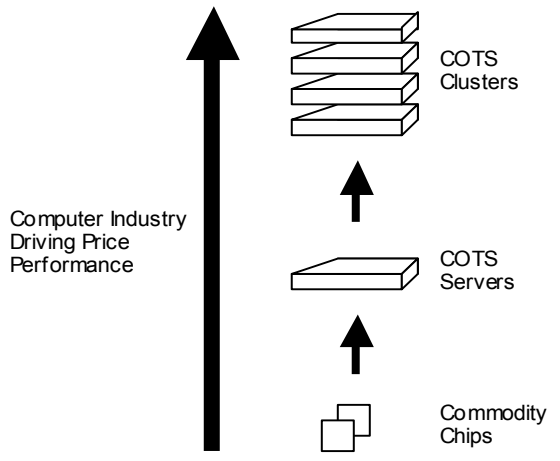

Figure 1 COTS Price Performance Chain

Since COTS clusters are built from COTS servers, the component servers are flexible in terms of configuration. They can be configured with more or less processing power, DRAM for main memory, and buffering and storage for content. Since the cluster itself is composed of smaller, high-performance building blocks (the individual servers), the clusters can scale from small capacity to large capacity requirements.

In addition to scaling conveniently to appropriate requirements for initial installation and deployment of a particular application, the clusters are flexible in that additional capacity requirements that come in after the system has been in service for some period of time can easily be accommodated by the addition of new servers and storage capacity into the existing cluster.

Clusters are becoming so popular that some vendors are beginning to provide not only the individual servers and network devices needed to construct a cluster, but also products that are themselves pre-packaged clusters with all the components priced together as a single bundle.

As an extension of the cluster concept, blade servers are the next, high-density incarnation of the commodity cluster. Blade servers do not yet have the same level of standardization that yields the benefits of interoperability and competition for performance. Therefore, the price at this point is higher for blade server solutions.

With the benefits provided by COTS products and their application to high-performance computing applications, the trend in the VOD market is toward open hardware systems rather than expensive, difficult-to-maintain proprietary hardware systems.

Applying COTS Clusters to VOD

VOD applications and more generally EOD applications have a structure that is well-suited to parallel processing and to COTS cluster techniques in particular

Mapping the VOD Application to the Cluster

For example, if we first consider a centralized network architecture using COTS clusters for VOD, the cluster must satisfy certain requirements including the maximum simultaneous subscriber sessions, each with an associated bitrate that may be 3.75 Mbps for standard definition (SD) content or much larger for high definition (HD) content. The cluster must also satisfy requirements for maximum content storage capacity. This application breaks down into the following primary compute or communication bandwidth intensive parts:

- Business logic that allows a subscriber to request content and pay for it.
- Session resource management that takes requests and identifies resources needed to fulfill them, including the content itself, processing capacity, and network bandwidth available to stream the rich media.
- Computing capacity to actually stream the high-bandwidth content.
- Low latency communication channels from subscribers back to the server to control the streams (with pause/fast-forward/rewind).

The business logic and processing can be achieved using modern, flexible integration technologies that power the web's still-burgeoning e-commerce activities.

Session resource management is a very important piece where intelligence and accurate information about the current system load and currently available resources are very important. This activity, however, does not take a lot of processor resources or bandwidth - merely timely information.

The bulk of the work, in terms of processor cycles and network bandwidth consumed that is required to provide rich media on demand, is the job of streaming out the media. Streaming can be anywhere from 30 seconds to 3 hours in duration, for a quick video commercial or a long-running movie, respectively.

The latency required for responsive trick mode transitions places more requirements on network bandwidth and availability between the subscriber and the server site without putting a significant additional computation burden on the VOD server (unless fast-forward/rewind content are not computed in advance and must be computed on the fly).

The upshot is that most of the computational load associated with VOD is in the streaming of independent content to independent viewers. Thus, computations do not interact with each other, so as a high-performance computing task, relatively little network bandwidth is consumed in coordinating within a cluster on the execution of these tasks. The fact that tasks are independent makes them particularly well suited to the COTS cluster technique for high-performance computing.

Smart Cluster Management Software

To ensure that cluster resources are used most effectively, the cluster must look at incoming requests and determine the best way to service the request. The cluster management software must decide which node in the cluster is best suited to stream a request for a particular piece of content. This software handles the load balancing, availability and failover requirements for the cluster. This software uses distributed algorithms such that each server is capable of performing this management function, and therefore no single point of failure exists in the cluster management.

In order to make resource management decisions, the cluster management software must collect information from each node in the cluster describing the content available on the node, the performance characteristics of the node, and the current load on the node. With this information, the cluster manager has information about available capacity to service new requests for specific content on various nodes in the cluster. It can then make the appropriate load balancing decisions and assign the request to a particular node.

The cluster management software also keeps track of server or node availability in the cluster. For example, if a particular node is taken out of service or suffers from some

type of hardware failure, the cluster manager detects this condition and makes the appropriate adjustments to its resource management policies for the period when the server is not available. Likewise, when a new server is added to a cluster, it automatically participates in the cluster and coordinates with the other nodes in the cluster to make its performance capabilities and availability known.

Driving Price Performance for VOD

The performance advantages of using COTS clusters for high-performance computing described above drive price performance for VOD as a specific application. Scalability for the COTS cluster translates directly into scalability of streaming capacity for a VOD cluster. VOD clusters can scale smoothly from small numbers of streams (a few hundred SD streams using single dual Xeon class servers) to much larger numbers of streams (tens of thousands of SD streams when these single server building blocks are racked up in quantity in larger COTS clusters).

Additionally, COTS clusters for VOD can accommodate different storage configurations that may be required for different types of VOD storage requirements, performance requirements, and usage patterns. For example, a VOD deployment that must support large numbers simultaneous users (greater than 10,000) may benefit from a shared storage model that enables each of the nodes in the cluster to share access to content on a single large SAN-type storage unit. A flexible software solution for COTS clustering can accommodate appropriate COTS hardware configurations and reap the benefits of lowering storage costs in cases where large amounts of storage accessible by each individual node is needed.

Content Management Within the Cluster

Part of the job of the cluster management software is to make sure the streaming resources for particular content expand as the demand for that content expands. One technique is to divide the storage available on each node into a unique content partition and a cached content partition. The cache then temporarily holds content that has become popular.

To utilize the cache partition of storage, one technique is to identify content that is becoming popular by monitoring the request rate in real-time. Once a piece of content is becoming popular, the cluster manager can copy the content to another node to increase the capacity of the cluster and create streams for that content.

Another technique is to bring new content directly into the cache partition of storage on the assumption that fresh, newly released content is more likely to be popular. The caching policy then keeps the content for as long as it is popular and replaces content in the cache based on policies that take into consideration usage characteristics. On an ongoing basis, the number of content copies that are stored in the cluster can be adjusted up and down based on usage characteristics of the content over time.

The content management UI for the cluster must allow the cluster to be managed as a single entity with the information from individual nodes aggregated into whole-cluster totals. This includes information about the current number of streams being served, a content list as well as number of copies of each piece of content, etc.

EXPANDING THE VIDEO NETWORK

High performance in the COTS cluster and effective cluster management provide the essential foundation for a complete, scalable VOD rich media delivery system, but additional techniques concerned with the distribution of clusters, the effective communication of information and content between them, and the efficient use of network bandwidth are also very important for large deployments.

Transport Network Architecture for VOD

More sophisticated techniques for content placement and management extend the idea of caching to a general storage hierarchy that overlays the network topology hierarchy as depicted in Figure 2.

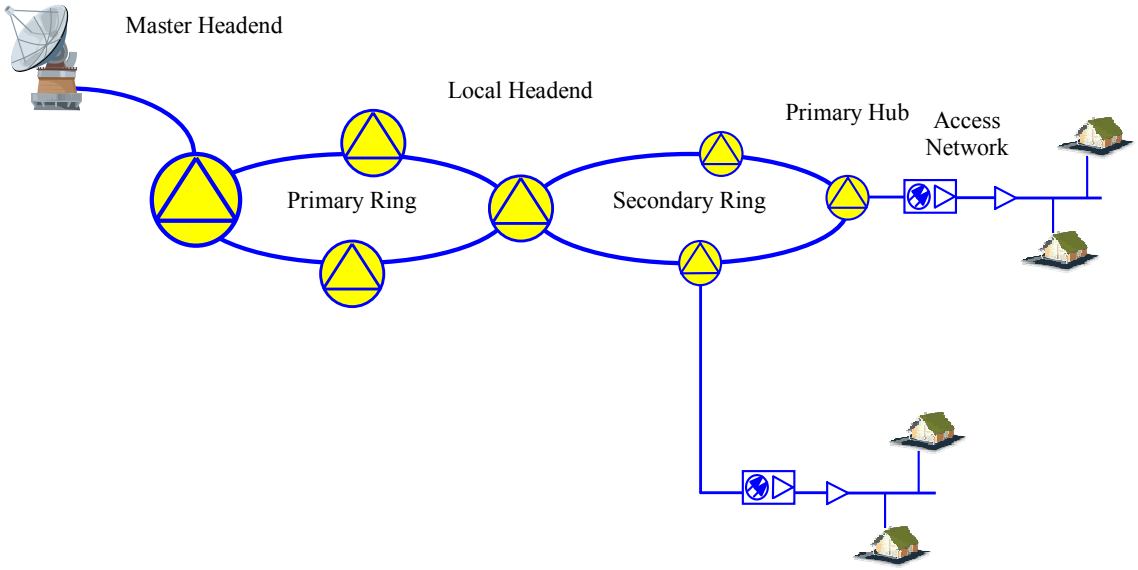

Figure 2: Hierarchical Network Architecture

A typical network topology includes:

- A master headend, which is a central site where many of the entire network's operational services such as transaction processing and billing are housed,
- The headend that acts as a point of origin for various video, data and voice services,
- Primary hubs that are closer to the subscriber and allow for effective aggregation of transport functionality and services for the subscriber, and
- Nodes that reside closer still to the subscriber.

Network architectures vary in terms of how much bandwidth is available through the backbone between regional headends and how much bandwidth is available from headends down to primary hubs. These bandwidth-provisioning factors impact the degree to which streaming can be centralized and how much motivation there is to decentralize streaming.

Storage Hierarchy Mapping to Network

The storage hierarchy can be designed to mirror the network hierarchy for purposes of cost-effective use of bandwidth in the backbone and in the access network and for effective us of network equipment at appropriate locations in the network. illustrates how this storage hierarchy would map onto a typical network hierarchy.

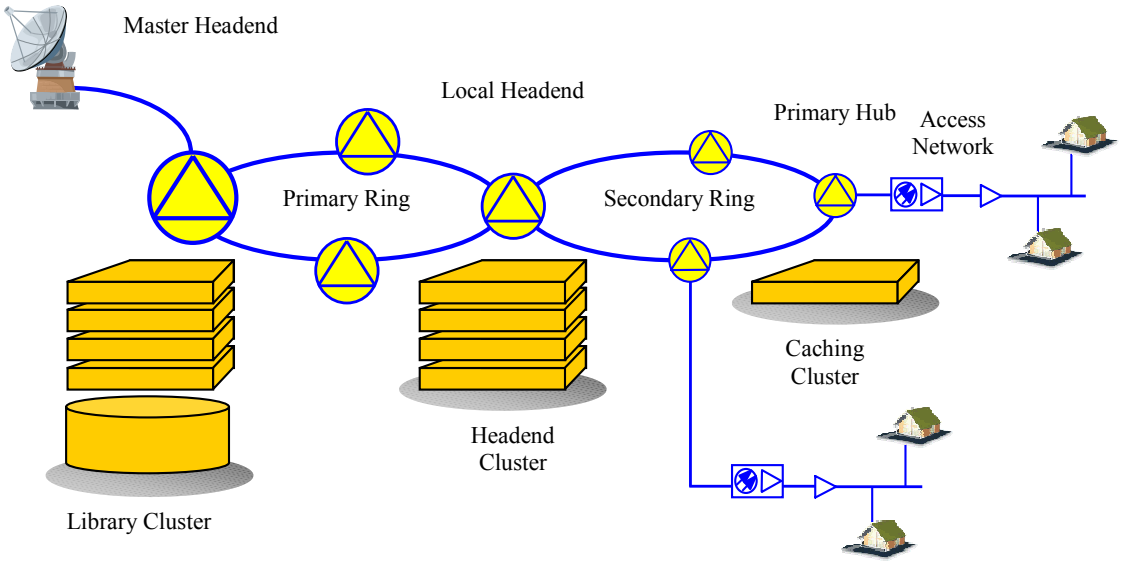

Figure 3: Storage Hierarchy

This storage hierarchy consists of levels that include:

- Library servers, which act as large content repositories for a wide range of different content, located at the master headend.
- Centrally located popular content repository and source for content flowing toward the subscriber on a day-to-day basis, typically from the headends.
- Localized "caching" servers, located perhaps in the primary hubs.
- Smaller hubs and nodes that would not typically contain active streaming on-demand servers but might house elements of the transport network associated with rich media delivery.

Content Propagation in a Video Network

The video network architecture described provides for a great deal of flexibility in terms of network configuration. A typical usage scenario for the video network would have content coming into the network through a combination of content aggregators that distribute via satellite and real-time

content capture and ingest. This new content is stored initially in the library server cluster in the master headend, perhaps using a shared storage hardware configuration to accommodate very large amounts of content such as those generated through real-time ingest applications.

The library cluster provides archival storage for content that has the appropriate licensing for long-term storage. The content for which a licensing window applies is also stored during the appropriate window and the copy of such content in the library clusters serves as the master copy of that content for the entire system.

The library cluster makes the content available, not for streaming directly to subscribers, but for transfer to streaming servers in other parts of the hierarchical video network.

The closest set of streaming servers belongs to the VOD clusters located at headend sites. These clusters are expected to store only content that is expected to be popular or that has been dynamically discovered to be popular. The headend cluster will pull content from the library cluster in cases where a request comes in for the content but that content is not currently resident in the headend cluster. Latency associated with copying the content is minimized through the use of streaming mechanisms that allow for content streams to be played out from the beginning while the later parts of the content are still being transferred in to the headend cluster.

Mechanisms for monitoring content usage and making copies of popular content in clusters closer to the subscriber, are based on the algorithms for dynamically creating copies of content within a single cluster as described above.

Moving down one level in the hierarchy, the VOD clusters located in the primary hubs economize on network bandwidth in the higher trunks of the hierarchy by allowing for very popular content to be cached in the closest possible location to the subscriber. Assuming an “80-20” rule where 80% of the content requests can be serviced by 20% of the actual content available, these caching clusters in the primary hubs serve a relatively large percentage of streams with a relatively small percentage of available content stored locally. Thus, the bandwidth required to service all of those streams is only consumed from the primary hub down to the individual subscribers rather than the same amount of bandwidth being required (and the amount compounded by adjacent primary hubs) in the higher-level trunks for the distribution network.

Streaming Bandwidth in the Video Network

The streaming bandwidth savings possible in the trunk links of the network when VOD clusters are distributed deeper into the network qualitatively justify the content propagation techniques as described above. However, a more quantitative analysis of those savings is even more convincing. The following analysis compares a centralized VOD cluster approach and the bandwidth required through the network to stream against a standard set of requirements to a distributed VOD cluster approach and the network bandwidth required in that case. The requirements of higher-bandwidth content such as HD video streams must be considered as part of the quantitative modeling of the bandwidth requirements looking toward changes we are likely to see in the near future.

QUANTITATIVE ANALYSIS

Let us consider the bandwidth requirements for a group of systems that require a total of 40 Gbps of simultaneous streams (equivalent to more than 10,000 SD streams at 3.75 Mbps) to be delivered to subscribers during peak load time, a total of 20 TB in archival storage and a total of 4 TB of unique content readily available to subscribers. Suppose that these subscribers are served by 5 systems which need to support 12 Gbps, 12 Gbps, 8 Gbps, 4 Gbps and 4Gbps for System A, B, C, D and E, respectively. To compare the aggregate network bandwidth required in the case of the centralized video network and the distributed video network, we will consider the typical network architecture described above and look at bandwidth required in each case for the 3 levels of the network: primary ring, secondary ring, and access network.

System	Number of Hubs	Bandwidth (Gbps)	Storage (TB)
A	4	12	20
B	4	12	20
C	3	8	20
D	2	4	20
E	2	4	20

Centralized Video Network

Bandwidth: Assuming that central site is different from any of the five headends: Requires full 40 Gbps across each of the three networks (Primary ring, Secondary ring, and Access network), which is 120 Gbps. If the central site is co-located with one of the headends, say one of the largest ones, which supports 12 Gbps streaming, the result is a savings of 12 Gbps since bandwidth for that headend does not need to go over the Primary ring. Therefore, total bandwidth required is 108 Gbps.

Storage: One copy of the entire content collection is required at the single central location, so the total requirement is 20 TB.

Decentralized with Duplication

Bandwidth: Streaming for each headend starts at that headend and traverses the Secondary ring and the Access network. Therefore, we have the total of 40 Gbps crossing two networks for a total aggregate bandwidth requirement of 80 Gbps.

Storage: The content is fully duplicated at each of the five headends. With a content requirement of 20 TB at five systems, the total storage requirement is 100 TB.

Distributed Video Network

Bandwidth: Assuming minimal streaming traffic from the Library cluster across the Primary Ring, we focus on the traffic across the Secondary ring and Access network. Assuming that 80% of the streams can be addressed with 20% of the content from the Caching clusters in the primary hubs, the bandwidth required for that 80% is only in the access network and totals 80% of 40 Gbps or 32 Gbps. The other 20% of the streams are served from the headends; 20% of 40 Gbps is 8 Gbps and those streams must traverse both the Secondary rings and the Access network for a total of 16 Gbps of network bandwidth required for the 20% of streams. The grand total for all 100% of streams is therefore 32 Gbps + 16 Gbps = 48 Gbps.

Storage: Firstly, the Library cluster requires storage for the entire system at 20 TB. Assuming a working set of approximately 400 titles at the headend and assuming each title is approximately 1.5 hours, the headend clusters require 600 hours of content. At 2 GB/hour that would be 1.2 TB of storage required at each of the five

headends; grand total is 6 TB at the headends. In the caching clusters, we need the 20% most popular content of the 400 titles in the working set, so the requirement is for 80 titles or 120 hours (assuming 1.5 hours/title). At 2 GB/hour, we require 240 GB per primary hub. In the entire system, we have 15 primary hubs each requiring 240 GB of storage, so the total hub storage is 3.6 TB. The grand total for all storage in this scenario is therefore: 20TB + 6 TB + 3.6 TB = 29.6 TB.

SUMMARY

The summary of bandwidth and storage requirements for the three scenarios addressed in the analysis above appears in the following table.

Centralized	Bandwidth: 108 Gbps
	Storage: 20 TB
Decentralized with Duplication	Bandwidth: 80 Gbps
	Storage: 100 TB
Distributed	Bandwidth: 48 Gbps
	Storage: 29.6 TB

Clearly, the Distributed Video Network with Library cluster for archival storage and Caching cluster near the network edges yields substantial savings in both aggregate network bandwidth and in storage required. In comparison, the Centralized Video Network leans heavily on network bandwidth and the major expense is in that area. The Decentralized with Duplication architecture, with duplicate content at each headend, leans heavily on additional storage, which contributes greatly to the high cost of that solution.

CONCLUSION

We have summarized the requirements for the large-scale delivery architectures required for the fast-approaching EOD vision of coming services. Novel techniques for high-performance computing platforms and efficient video network architectures are necessary to realize this vision.

The COTS clustering approach to high-performance computing has proved to be successful in virtually all high-performance computing applications since the late 1990's and addresses the needs of VOD delivery very cost-effectively. This approach leverages all of the research and development investment in the computer industry, which amounts to billions of dollars, for the benefit of driving price performance in the VOD and EOD application arena.

Smart content management in the cluster management software as well as in the larger scale distributed video network context help to reduce costs elsewhere in the complete end-to-end system by using the network bandwidth and storage resources associated with the video network most effectively. A representative analysis of a typical multi-system VOD deployment quantifies the savings in bandwidth and storage that is possible with this approach. And as the requirements of the future scale toward extreme narrowcasting of everything-on-demand, the cost optimization available through these techniques becomes essential.

REFERENCES

1. Sterling, Thomas. "An Introduction to PC Clusters for High Performance Computing" The International Journal of High Performance Computing Applications, Volume 15, No. 2, Summer 2001, pp. 92-101.
2. Karp, Alan H., Lusk, Ewing and Bailey, David H. "1997 Gordon Bell Prize Winners." IEEE Computer, January 1998.