TRUE VIDEO ON DEMAND VS. NEAR VIDEO ON DEMAND, STATISTICAL MODELING, COST, & PERFORMANCE TRADE-OFFS.

By
Winston Hodge, Hodge Computer Research
and
Chuck Milligan, Storage Technology Corporation

Abstract

Video On Demand (VOD) has the potential of giving individual television viewers nearly instant access to a wide range of recorded movies, video programs, games, information and other services. It is distinguished from more conventional TV viewing by a high degree of interactivity between the viewer and the material being viewed.

A perception exists in this industry today that each person interacting with their TV demands instantaneous response. This is called True Video On Demand (TVOD). As this paper will show, TVOD is extremely expensive when it provides for all services possible.

The alternative to TVOD is Near Video On Demand (NVOD). This paper will demonstrate that while NVOD is significantly less expensive to implement, an NVOD system can be designed so that its delays are not objectionable to the user for many applications. Procedures and strategies for concealing customer latency time will be described, along with the cost differential attendant to eliminating it.

Access to recorded material with zero access time is not physically possible. Fractional second access is possible, but would be very expensive for an unlimited menu of choices by an unlimited number of subscribers.

Clearly, the quantification of cost to provide service versus the latency time is of serious importance. But there is more to the implementation decision than cost. The psychological effects of waiting come into play. For example, is one second too long to wait? How about two seconds? How about two minutes? All things being equal, (which they are not), the shorter the service time the better.

This paper will provide a clear view of physically possible service times and the cost to provide those services *using advanced technology hierarchical storage.*

A model will be described which demonstrates how the system cost varies with viewer latency. This model will be applied separately and collectively to the video server, disk storage complex, *large terabyte robotic tape farms,* VOD selector switch, communications channel and viewer selection mechanism.

Block diagrams used in the systems analysis and simulation will be included, along with charts and graphs which will clarify the results of the analysis. The paper will conclude with recommendations for an economically viable system design.

Definition & Requirements

Video On Demand (VOD) trial systems in one form or another are currently being implemented. An understanding of the cost factors related to response time (i.e. viewer selection latency) will provide insight into the overall system costs.

Interactivity is much more than channel selection. It may be the simple ability of the viewer to decide **what** program he wants to watch, and **when** he wants to watch it. It might allow him to select from among several different endings to a movie thriller. It may allow him to take a simulated walk down a supermarket aisle **he** selects, ordering products from among those displayed. It could allow him to engage in a simulated trip through the solar system or a Mayan temple, making decisions about which planets to explore or which corridor to turn down, through the wonder of virtual reality. It could even allow him to engage in a simulated dog fight with another viewer through

an interactive video game which could be offered.

The foregoing scenarios require progressively increasing levels of interactivity. The response times required of the system also vary widely between the applications. For example, when home shopping, the response time from advertisement to order placement is not critical but the navigation response from product to product is more significant.

The viewer may be more concerned about the time between selection and delivery of a new movie, but whether this time interval is fractions of seconds, seconds or even minutes may not be consequential.

A video arcade game or a virtual reality session requires much more rapid response--far beyond the capabilities of even a very large mainframe computer to service a large number of clients. For these applications, the interactivity will be supplied by downloading a program to a set-top box for execution. Given this fact, once more the initial response between ordering the game and its actual delivery falls into the same degree of urgency as the ordering of a movie.

Selection time is subject to the laws of physics. These laws place limits on what it is physically possible to achieve. By knowing where the limits are, and by understanding the cost of approaching these limits, one is in a position to make objective decisions on implementation approaches. This paper will enlighten the reader with the options currently available.

Strictly speaking, True Video On Demand (TVOD) requires instantaneous response, probably less than a second from the time a program request is made until the time the program is delivered. This has significant cost ramifications not only for the video server and video disk drives, but for the communications channel and other system elements not addressed in this paper.

Near Video On Demand (NVOD) requires only a reasonable and convenient response time from program selection to program delivery. This interval could range from seconds to a few minutes or in some cases even a few hours. During the interval, stock material (such as seen in theaters) or interactive advertising for food or other products to be delivered to homes, or music video interludes may be presented.

The system to be discussed will even allow a viewer to see new movies at reduced prices by selectively permitting advertising inserts in the subscriber's now less expensive pay per view movie. This scheme could allow several price levels, depending on the total number of minutes of commercials the viewer is willing to tolerate. This, in turn, would allow the service provider to offset the reduced customer billing with advertising revenues so earned.

The bottom line for the service provider should be: Which operating procedure, NVOD or TVOD, produces the largest revenue stream at what cost, ultimately providing the greatest return on investment? This paper will summarize these issues.

System Possibilities

In order to analyze TVOD vs. NVOD costs, it is necessary to understand the three prominent hardware implementation philosophies illustrated in Figures 1 through 3. The differences between approaches depends on a vendor's reliance on his installed hardware architectures, as well as his philosophy on whether a general or "tuned" solution is preferable.

In all the examples to be presented, it is assumed that the transmission system employs Asynchronous Transmission Method (ATM). This protocol utilizes data packets consisting of a five byte header and a 48-byte data field. The header describes the destination and the content of the information portion of the packet. It is further assumed that the appropriate storage solution is a 3 level hierarchy of disk and robotically managed tape libraries. The general solution uses standard operating system functions and software, and the more "tuned" solutions employ significantly more specialized software and firmware to manage the hierarchy.

For applications where the volume is not adequate to justify a custom or tuned design philosophy (such as for a small number of tests sites, or for concept validation where reduced non-recurring costs are important), the generalized solution as shown in Figure 1 may be preferable to a tuned solution. It is less expensive because it relies mostly on the procurement of off-the-shelf hardware and possibly off-the-shelf software. The generalized system can produce both TVOD and NVOD, but the cost of delivery is high.

In Figure 1, the term "mainframe" is intended to mean a general purpose processor running a "standard" operating system (e.g., a RS6000 running UNIX). Such mainframe system solutions are often more expensive than tuned solutions in production because a great deal of system hardware and software must be provided which is unnecessary for the specific application. Further, the mainframe data flow is designed for data processing, not data movement. Video applications require a great deal of data movement, with very little data processing.

The image processing (such as image compression and decompression) is usually performed by specialized hardware units. This is because affordable mainframes cannot handle the computational load required to deliver multiple video programs in real time.

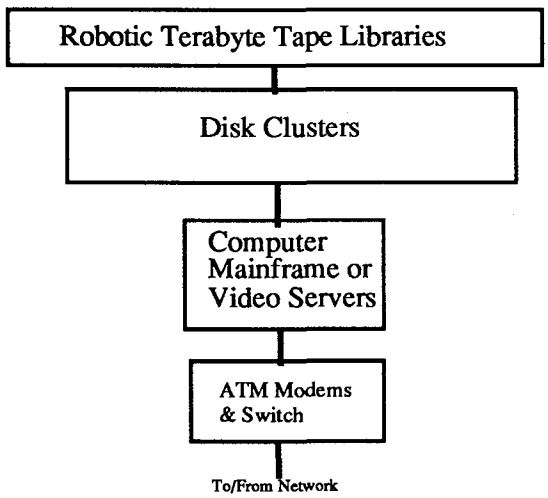


Figure 1, A Generalized Video On Demand System

When the opportunity exists to construct thousands of units for a specific application, the tuned solution is preferable because of lower cost, higher performance, superior function and just a better fit to the problem being solved.

There are various degrees of tuned systems. Some systems are very good at creating databases of still images or moving video which use general purpose operating systems, database managers, networking facilities and the like. These systems rely on small amounts of customization. They can do a good job of delivering a small number of selected videos on demand to a small customer base. As in the previous systems, they can produce either NVOD or TVOD, but the program selection is limited and the size of the client base is severely restricted when operating in TVOD mode.

These systems may be cascaded to accommodate more videos and more clients. An example of such a cascaded system is illustrated in Figure 2.

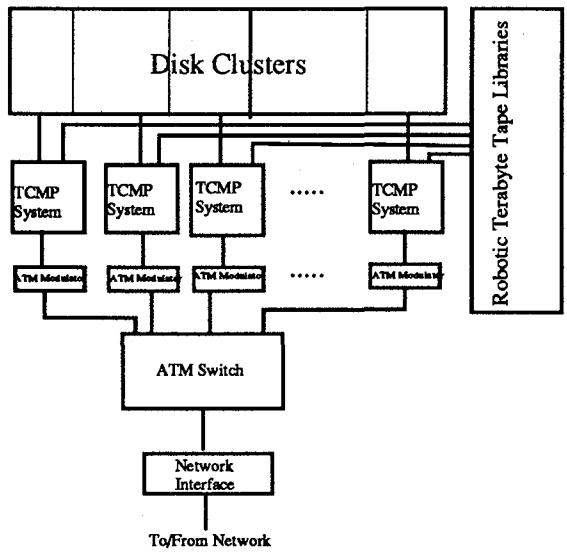


Figure 2, Cascaded Tightly Coupled Multiprocessing (TCMP) Video Server

The ultimate tuning of a video server exists when special paths are provided for moving digital video information. An architecture can be created which relaxes the throughput requirements on the computer performing the server function.

Once the server has interpreted the customer's video request, validated billing and program availability, confirmed that the requester at the customer premises is not restricted (child requesting X Rated movie), and arranged for the short term scheduling (seconds or minutes), the server computer submits the program material request and the electronic customer address to the Server Saver/ATM Switching system. Then for the balance of that transaction, the server has nothing more to do until the program is complete (for a typical movie this would be between 90 and 110 minutes). This system is shown below in figure 3.

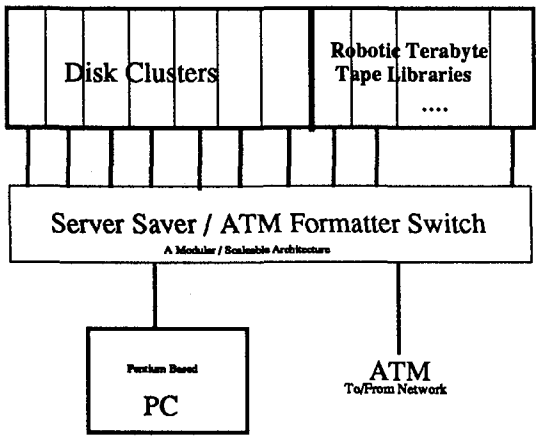


Figure 3. Composite Server Saver System / ATM Switching System

The Server Saver Sub-System permits the use of inexpensive components and simplifies data routing and manipulation while simplifying computational requirements to such an extent that a single high performance PC such as a Pentium [1] or a Power PC [2] can assume responsibility for a 500 program 10,000 subscriber system.

If a larger system is required, these systems can be cascaded to produce greater program selection for more patrons. The Server System can produce TVOD for a small number of subscribers or NVOD for a large number of subscribers, or some combination of both TVOD and NVOD. This capability is similar to that of the above systems, but **at very low relative cost.**

The Server Saver system is a simple device both architecturally and physically. It connects to a "storage farm" through multiple SCSI data paths, to the PC via one or more SCSI data paths, and to the CATV or other network through the ATM Formatter/Switch.

The Server Saver has only three types of interfaces:

(1) to/from the PC computer
(2) to/from the Storage Farm
(3) to/from the ATM network.

The Server Saver provides storage control, flow control, packet switching and an interface to the ATM network.

Costs

Each of the preceding systems can supply TVOD, NVOD, or some combination of NVOD and TVOD, but at substantially different costs. The cost of each of these systems varies as a function of program capacity, subscriber capacity, and the degree of responsiveness to customer requests.

It is obvious that video program capacity is a function of storage and that storage, in turn, makes up a major portion of system cost.

Each 90 to 110 minute program can require from 1 to 9 gigabytes of storage depending upon resolution requirements. Each gigabyte of disk storage will cost from $750 to $1300 at the system level, while data in robotically controlled tape systems (e.g. StorageTek Nearline offerings) will cost $7 to $10 per gigabyte of storage. There are also performance differences between disk and such tape systems. These will be discussed later in some detail.

Obviously, the more programs desired, the more storage is required, which in turn increases hardware costs.

The generalized video server systems typically cost $250,000 and upwards. Tightly coupled multiprocessing video server

systems currently cost between $65,000 and $100,000 per module, each of which is capable of producing up to 25 programs concurrently.

For example, 500 channels of programming could cost (500/25 X $65,000) or $1,300,000 per video server complex, not including ATM formatting, switching or interfacing.

Each of these systems has limited capacity, requiring additional system hardware replication to yield more capacity and more responsiveness. Again, added system hardware increases system costs.

The purpose of this article is to determine for the various generic hardware approaches the costs to produce the continuum between TVOD and NVOD and how much responsiveness can an interactive TV system cost-effectively produce.

This paper will generate some approximate best case and worst case pricing for each of this trio of approaches, determine reasonable pricing intervals, and the subsequent cost relationships for TVOD and NVOD. This will facilitate the qualitative judgment as to whether, for instance, it is worth an additional $500,000 or more to give the customer a program selection response time of 1 second/minute instead of 30 seconds/ minutes.

Further, after the analysis, procedures for camouflaging program latency will be discussed.

The following spread sheet represents estimates of significant costs for each of the three prominent system architectural philosophies shown in Figures 1 to 3. While these numbers may be challenged as being tomorrow's prices, guesses or inaccurate, they do represent working approximations derived from potential vendors in this industry. It is interesting to observe that using any set of different reasonable numbers does not change the comparative relationship, i.e. - NVOD is much less expensive than TVOD.

This paper has alluded to video programs and threads. A thread is defined as a continuous stream of video representing one complete program, using one of the available broadcast channels. Since both tape drives and Video Friendly disks can produce data transmission rates greater than required for a single channel, it is possible to store the data in such a fashion that it can be read out multiple times in real time.

If a device is able to sustain a data rate 10 times greater than is required for normal video rates, 10 video streams or "threads" could be produced if only short duration device read interruptions occur (e.g. for turnaround at end of tape track or for head or next cylinder seeks). An alternative is for additional buffering to be used to mask longer duration read interruptions. It is possible and therefore desirable to interleave the programming material such that each thread is displaced in time.

For example, a 90 minute (1 gigabyte) video program can be structured to allow 10 threads, and would have each thread offset by 9 minutes. This can be accommodated by appropriate data structures using only one gigabyte in either tape or disk storage.

Because TVOD requires the ability to instantaneously access the first and then subsequent video frames of the program at random and arbitrary intervals, it would require that the storage device be capable of rapidly switching from one random spot to another to support even two threads, let alone a number as large as 10 or 12.

Although tape can support that many threads of NVOD, multiple thread TVOD is not feasible with tape devices, because they require seconds to move from one random spot to another.

TVOD is feasible but more expensive with disk because buffers must be included in front of each device for each thread, which substantially increases the cost per thread. More importantly, the random seeks reduce the sustainable rate of the device so that it is less efficient, and even with external

buffering, can sustain significantly fewer total threads.

For example: Assume a particular disk can sustain 3 MB/sec with a maximum (because the video stream must be guaranteed) random seek time of 33 ms. If the disk is rotated at 5400 rpm, it will have approximately 33KB on a track that will spin by the head in 11 ms. (These of course are budgetary numbers, but may be adjusted for any particular device).

If a random seek is allowed at the end of each track transfer in order to switch to another thread, then the sustainable rate is:

$$3^{MB}/_{sec} \times \{(11^{ms}/_{T})/(44ms/_{(T+sk)}\}$$
$$= .75 \text{ MB/sec}$$

If a video stream requires 1.5Mb/sec (~.2 MB/sec), then the NVOD approach allows 15 threads without buffering, while the TVOD approach allows only 3 threads, with buffering. The buffer size, however, need not be very large, i.e.:
Since:

.2 MB/sec=>.2KB/ms

Then let buffer size for each thread $=B_T$

B_T=.2KB/ms x (3 seeks of 30ms + 2 transfers of 11ms)

B_T=.2 x (90+22) KB

B_T=25KB

Since each track must be buffered, it would adjust to 33 KB/thread => 100 KB buffer total.

This of course assumes the video friendly type of device that has no other non transfer activities to mask.

Therefore, it is clear that TVOD threads with only one (or serendipitously, a few) customers per thread will require many more disks than a NVOD with schedulable threads which allows a significantly greater number of customers per thread and a significantly greater number of threads per disk.

Furthermore, where as the TVOD approach limits tape storage to 1 thread per device (as opposed to the 2 or 3 for the disk), the NVOD approach works as well from tape as it does from disk. The systems configured below are intended to support 1,000 to 10,000 program titles and use the tape as the primary storage media. The disks are used as a buffer for the currently active programs primarily to reduce the number of passes against each tape volume for reliability purposes rather than for performance. As a matter of fact, tape performance in some instances will exceed that of disk devices in terms of the number of simultaneous threads that can be sustained. With NVOD threads scheduled in greater than 30 second increments,(e.g. 5 to 15 minutes) the delay would completely mask the initial few seconds of startup to mount the tape.

Using tape directly, or using disk as a buffer in front of the tape for most of the active programs (assuming the disk described above and that each will hold 3 to 5 gigabytes) it would be possible to have each tape or disk provide as many as 15 threads (channels) of broadcast. This could be all for one program, or split among the number of programs that could be stored on that one device (e.g. three 90 minute movies would require 3 to 4 1/2 gigabytes of storage).

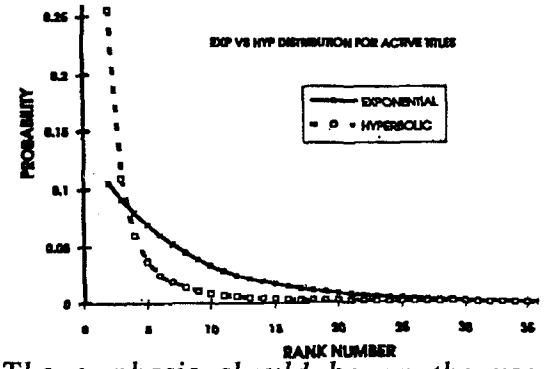
To support 200 channels of NVOD would require a minimum of 14 devices, and 500 channels would require a minimum of 34 devices.

The experience in this industry is that in any particular week there is a very small subset of programming that accounts for most of the demand. One specific example is for video rental where 97% of revenue comes from less than 25 titles. With this tight a skew, out of a population of 1,000 to 10,000 titles between 33 and 68 titles account for 99% of the demand and between 39 and 129 titles account for 99.5% of the demand. See the inset below for the details on this set of calculations.

Customer Demand vs. System Performance Limits Analysis

The given task is to identify the number of program titles necessary to satisfy a "large" proportion of the customer requests. Obviously the greater the percentage of requests one desires to satisfy, the larger the population. Also the distribution of the requests across the inventory of titles significantly affects the number requested. If the total number of titles is significantly greater than what can be simultaneously broadcast (e.g. more than an order of magnitude such as 200 channels for 2000 titles) then the true answer will generally lie between an exponential and a hyperbolic distribution. Experience has shown that the number will quite often track hyperbolic through some significant portion of the range (e.g. 95% to 99% depending on the tightness of the skew) and then drift to the exponential and then terminate at some finite number far short of where either distribution would predict.

Without knowing the actual distribution of requests to the most popular titles, it is difficult to calculate the exact number of titles that must be broadcast with any confidence. However using what little is known about the reference patterns of the video rental base (e.g. one company reports that 97% of revenues come from 20 to 25 titles) one can calculate a range and bound the problem using distributions that historically tend to fit skew problems of this sort; i.e. Binomial (to give an easy but very gross and optimistic first approximation), Exponential, and , and "Hyperbolic" (or "Pareto") probability distributions.



The emphasis should be on the use of hyperbolic distributions (with the probability density form $p(n)=A/n^k$ for $n\geq 1$). It is a convention to use the word "3-sigma" to mean the value of the tail beyond $z=\pm 3\delta$ limits for the case of a "normal" or gaussian distribution (even though the actual distribution is not normal and may not even have a "sigma". Framing the given problem between EXP and HYP limits gives the approximate value calculated here. One caveat is that historical skew distributions tend to deviate from perfect hyperbolic shapes at the high end tails (i.e. they drop faster than $1/n^k$ and this is formally called "droop"). This shortens the real use tail so that the actual expected answer should be below that calculated at the 3-sigma limit for the hyperbolic distribution.

1) **Most elementary approach (Binomial)**

p = probability of selecting "choice" movie $\approx \frac{25}{10^4} = .0025 = \frac{C}{N}$

$\mu = Np = 25$

σ binomial $\approx \sqrt{Np} = 5$, $3\sigma = 15$

$\mu + 3\sigma = 40$ Titles for 99.74% of demand

Same for Poisson (some N large, p small), $\mu = \lambda, \sigma = \sqrt{\lambda} = \sqrt{\mu}$

$\mu + 3\sigma = 40 \leftarrow$

2) **Exponential approach** $\int_0^{25} \alpha e^{-\alpha x} dx = 97\% = 1 - e^{-25\alpha}$

so $\alpha = .14$

find n s tail is .26% (Gaussian tail interpetation, 2 sides at 3σ)

so CUM $= 1 - e^{-14n} = .9974$

$e^{-14n} = .0026$

$n = 42.5$

% Demand Satisfied	25 Titles at 97% Skew	25 Titles at 95% Skew	20 Titles at 97% Skew	20 Titles at 95% Skew
99.74%	43	50	34	40
99.50%	39	44	30	35
99.00%	33	38	26	31

Scale Invariant Distributions

3) **Hyperbolic Distribution** Assume $f(x) = \frac{A}{x^k}$, k unknown > 1

normalize (find A): $\int_1^N f(x)dx = -\frac{A}{k-1}\left(\frac{1}{N^{k-1}} - 1\right) = 100\%$

our N very large $\Rightarrow$ A~(k-1).

Now CUM = 97% = $\int_1^{25} \frac{k-1}{x^k} dx = \left(1 - \frac{1}{25^{k-1}}\right) \Rightarrow k - 1 = 1.09$ $k \approx 2.089$ (Note: $k \approx 1.931$ for 95% skew)

Then for 99.749.(3σ),

$$(k-1)\ln(n) \underset{n \approx 236}{=} -\ln(.26\%) = \ln(\text{"tail"})$$

e.g. $f(x) = \frac{1.089}{x^{2.089}}$

% Demand Satisfied	25 Titles at 97% Skew	25 Titles at 95% Skew	20 Titles at 97% Skew	20 Titles at 95% Skew
99.74%	236	599	161	385
99.50%	129	296	93	200
99.00%	68	107	51	100

Max values: reality less due to "droop".

If the 25 titles were placed on shared disks at 12 threads each, and the rest of the programming spread with a few on tape drives, the 200 channels could be supported by 12 disk drives and 12 tape drives. The 500 channels would require about 24 disk drives and 16 tape drives. The following cost analysis is on the basis of about 20 disks at 10 threads per disk plus 16 tape drives.

Individual disks and RAID (Redundant Array of Independent Disks) systems have different performance characteristics, so the numbers derived for individual disks and RAID systems is different. However, even when using the more expensive RAID technology, only a few TVOD threads can be produced.

		1 thread/disk	10 thread/disk	ATM Encoder/	1 Thread/disk	10 Thread/disk
COSTS for 25 Thread Video System		Disk Cost	Disk Cost	ATM Switch	system cost	system cost
MP Server Only	$65,000	$25,000	$2,500	$6,250	$96,250	$73,750
Server Saver	$30,000					
Server Saver+Pentium	$40,000	$25,000	$2,500	$6,250	$71,250	$48,750
Mainframe	$250,000	$25,000	$2,500	$6,250	$281,250	$258,750
COSTS for100 Thread Video System						
MP Server Only	$260,000	$100,000	$10,000	$25,000	$385,000	$295,000
Server Saver	$120,000					
Server Saver+Pentium	$130,000	$100,000	$10,000	$25,000	$255,000	$165,000
Mainframe	$500,000	$100,000	$10,000	$25,000	$625,000	$535,000
Cost for 250 Thread Video System						
MP Server Only	$650,000	$250,000	$25,000	$62,500	$962,500	$737,500
Server Saver	$232,500					
Server Saver+Pentium	$242,500	$250,000	$25,000	$62,500	$555,000	$330,000
Mainframe	$1,250,000	$250,000	$25,000	$62,500	$1,562,500	$1,337,500
Cost for 500 Thread Video System						
MP Server Only	$1,300,000	$500,000	$50,000	$125,000	$1,925,000	$1,475,000
Server Saver	$480,000					
Server Saver+Pentium	$490,000	$500,000	$50,000	$125,000	$1,115,000	$665,000
Mainframe	$2,500,000	$500,000	$50,000	$125,000	$3,125,000	$2,675,000
Cost for 1000 Thread Video System						
MP Server Only	$3,250,000	$1,000,000	$100,000	$250,000	$4,500,000	$3,600,000
Server Saver	$930,000					
Server Saver+Pentium	$940,000	$1,000,000	$100,000	$250,000	$2,190,000	$1,290,000
Mainframe	$5,000,000	$1,000,000	$100,000	$250,000	$6,250,000	$5,350,000

Figure 4 - Assumptions Comparing the Three VOD Video Server Architectures

Total# Thread System	MP Server Only System	Svr+Srvr Saver	Mainframe
25	$96,250	$71,250	$281,250
100	$385,000	$255,000	$825,000
250	$962,500	$555,000	$1,562,500
500	$1,925,000	$1,115,000	$3,125,000
1000	$4,500,000	$2,190,000	$6,250,000
	10 threads / disk		
25	$73,750	$48,750	$258,750
100	$295,000	$165,000	$535,000
250	$737,500	$330,000	$1,337,500
500	$1,475,000	$665,000	$2,675,000
1000	$3,600,000	$1,290,000	$5,350,000
	Cost comparison of 1 Thread vs 10 thread systems		
25	1.305084746	1.461538462	1.08695652
100	1.305084746	1.545454545	1.1682243
250	1.305084746	1.681818182	1.1682243
500	1.305084746	1.676691729	1.1682243
1000	1.25	1.697674419	1.1682243

Figure 5 - Table representing completed video server costs for Multiprocessor System, Server Saver System, and Mainframe System. The last 5 rows of numbers represents cost improvement multipliers per thread.

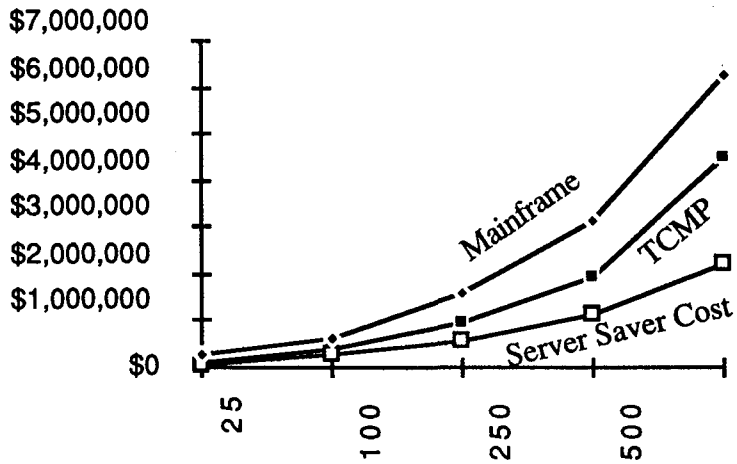


Figure 6 - Chart depicting relative system costs for each of the 3 candidate TVOD video server system implementations.

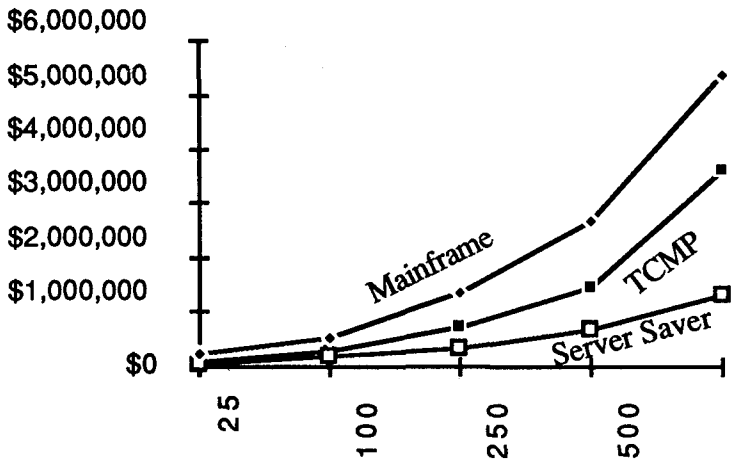


Figure 7 - This chart represents NVOD cost per program thread for each of the 3 candidate systems assuming 10 threads are available from each storage device simultaneously. Depending upon desired video quality and device performance, these numbers can change, but their relationships remain the same.

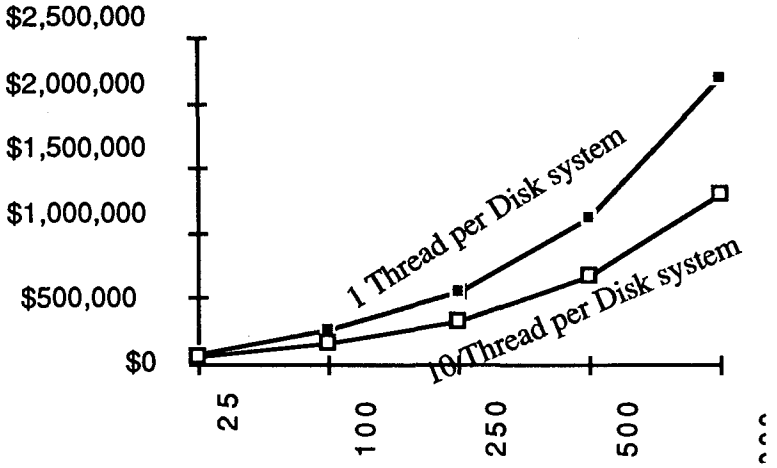


Figure 8 - This chart illustrates the cost of the Server Saver application. The upper curve represents system cost when only 1 thread per storage device (TVOD) is provided and the second curve represents system cost for a 10 thread per disk system is implemented.

The chart depicted in Figure 9 illustrates the cost savings as a percentage savings using the Server Saver System Architecture for 1 thread per storage device giving TVOD and 10 threads per storage device rendering Unlimited Capacity NVOD with a response time of 10 minutes.

When the system program capacity is 20 units, NVOD can be produced for about 68% of the cost of TVOD while systems above 250 programs flatten out such that NVOD costs less than 60% of TVOD systems as depicted in Figure 9.

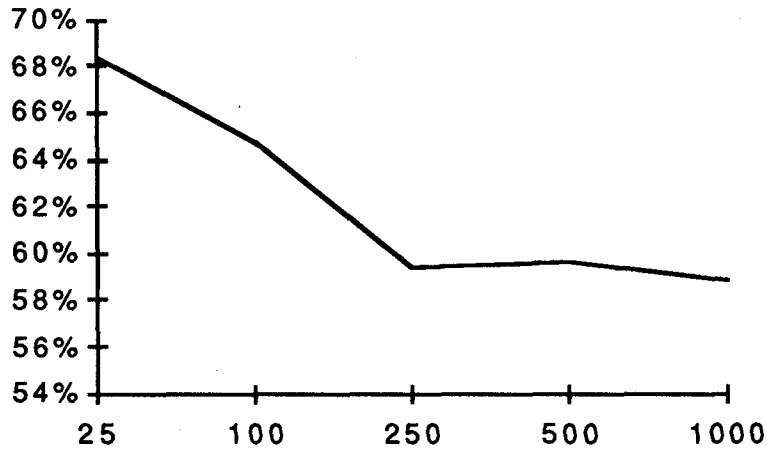


Figure 9 - The above chart depicts cost savings of NVOD system over TVOD system for Server Saver style architectures.

Figure 9 demonstrates how the cost per thread is reduced as the number of threads is increased. The vertical axis represents the cost relationship between the server saver system with one thread per storage device (TVOD) and the same server system with 10 threads per storage device NVOD). Ten threads per storage device implies that for a 90 minute movie, 10 equal space start times can exist providing a new start time for the movie every 90/10 or 9 minutes.

The horizontal axis represents the number of threads (channels) available to subscribers. The multiple thread system assumes that the disk storage system is video friendly.

Unlike standard drives, Video Friendly drives are designed to provide a worst case data rate that will assure highly predictable delivery of data so that discontinuities in the audio/video data stream will not exist.

Figure 10 illustrates the savings that multiple thread disks (NVOD) can have on each of the candidate architectures versus single thread (TVOD). NVOD produces more programming at less cost per program than TVOD.

Furthermore, since NVOD has an interval of time during which subscribers can request a program, NVOD can accommodate unlimited subscribers without requiring a subscriber to tune in late. Therefore, NVOD can produce substantially more revenue.

This paper is also intended to determine the cost consequences of employing tuned solutions to the TVOD

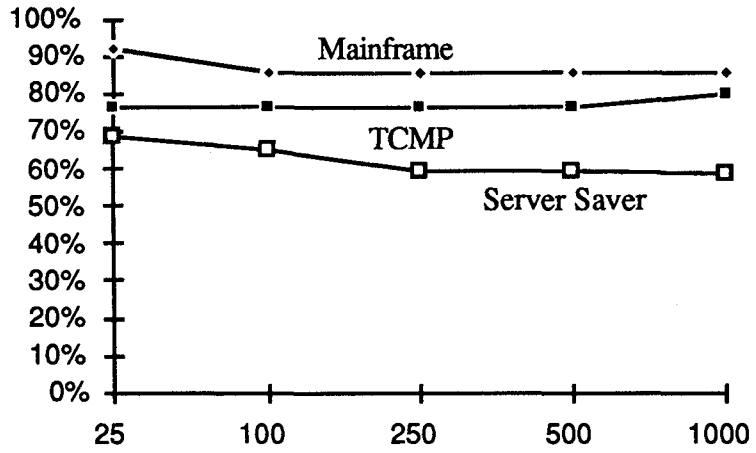


Figure 10 - This chart depicts the same server saver information as figure 9, but it includes the related information for the Tightly Coupled Multiprocessor Application and Mainframe Application

application versus general purpose solutions, or partially tuned solutions. Figure 11 illustrates that the tuned solution (i.e. the Server Saver architecture) with 200 or more threads will cost about 50% as much as the partially tuned solution (Tightly Coupled Multi-Processor) and about 25% as much as the general purpose (mainframe approach)!

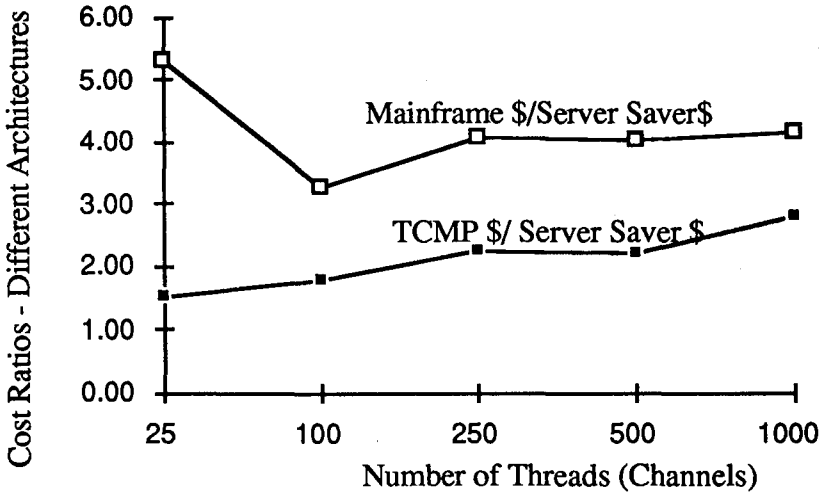


Figure 11 - This chart depicts NVOD based system costs normalized to the Server Saver Architectural approach.

The crucial element to facilitate both TVOD and NVOD is smooth, high bandwidth, uninterrupted device transfer capability which we will refer to as "Video Friendly". Interruptions in data will produce interrupted video unless extensive and costly video buffering is provided. Interrupted video of course is unacceptable.

Figure 12 shows Transfer time vs. wall clock time for four representative vendor disk drives.

The horizontal line at 33 ms represents the threshold of intolerable disk access times.

It should be observed that only one vendor drive achieves this requirement

(Micropolis) and one approaches this requirement.

Why use Video Friendly Devices

Video friendly devices are able to cost-effectively produce multiple threads of smooth program video without the requirement for external video buffering which requires extensive amounts of video RAM. Therefore, video friendly drives are a significant component in reducing system cost.

Why use Server Saver Style Architecture

The Server Saver architecture represents a highly tuned VOD application, not a generalized solution. It is the most cost effective tool to solve the VOD problem. It provides significant advantages, including cost / performance, system flexibility, simplicity, high uptime and low maintenance, as discussed in the author's previous referenced VOD articles (1). It can produce TVOD, NVOD and combinations of TVOD and NVOD.

Also, NVOD systems can produce unlimited customer showings per movie (unlike TVOD systems).

Why use NVOD instead of TVOD

NVOD systems can require approximately 1/2 the hardware cost to produce 10 times the video flow as do TVOD systems. Therefore, NVOD systems are the highly preferred economic approach. NVOD has been shown to cost substantially less to implement than TVOD and has the ability to support unlimited clients. NVOD can be tuned by the system operator to produce waiting intervals other than discussed in this paper.

Perhaps an average 3 minute wait for the program is too long, even though that time is used for information on upcoming attractions, to sell food to be delivered to the home, to sell other services, or to merely provide a music-video interlude, or some combination of these.

The system operator can reduce the NVOD interval by 50% while increasing his hardware costs substantially less than 50%, thus moving closer to the TVOD model. This procedure can be repeated as often as desired to further reduce viewer latency time.

Studies in one TVOD vs. NVOD trial by a hotel pay per view TV operator indicated no increased revenue stream for the TVOD application, only added cost to provide the function to the same number of clients.

One could make a career of looking at numerous other variations of data in the spread sheet and graphing and plotting them. It seems obvious to the authors if an operator is decided on a TVOD system, he can use the Server Saver technology and video friendly disks. If he desires the economies of NVOD, he can also cost effectively employ the Server Saver technology.

If the operator is unsure of whether he wants TVOD or NVOD, he can use the Server Saver technology and provide both styles of programming to his clients. Statistics collected from the real world will probably tell the real story.

What is the Impact of VOD on CATV delivered ATM

The basic non-cascaded Server Saver supports a 500 thread (or channel) system. The industry seems to support the idea of employing 50 MHz to 500 MHz for conventional analog TV and 500Hz up to 1000 MHz for digital interactive TV, while leaving 5MHz to 50 MHz for reverse channel communications.

If this is the case, then it is expected that as many as 500 streams or channels of digital interactive video could be placed in the upper CATV frequency band. If it were desired to support more program sources or threads, a different delivery system (such as fiber optic cables) might need to be in place.

Since fiber optic cable would only go to a city section, block or curb, costly ATM switches would be required to move the

proper packets from one transmission facility to another. This leads to the hotly debated question: Does a city require more than 500 channels of interactive TV and if so, how much more will it cost to provide them?

NVOD will not require as many channels for transmission as TVOD to support the same number of viewers; hence provides a great deal of relief from the expenses required to provide the infrastructure to support the greater number required by TVOD.

Filling in the Viewer Latency Time

The following strategy is proposed as a means of preventing the viewer from becoming frustrated at the delay between the time he makes his selection and the time it is actually delivered.

Assume a maximum viewer latency time of 10 minutes. A number of pre-packaged "mini-programs" may be prepared. They could be binary divisions of the 10 minute maximum time to be filled if a viewer requested a program only 1 second after the previous start time.

Thus, there could be one of several ten minute cartoons, five minutes of coming attractions, two and one half minutes of news headlines, one minute and 45 seconds worth of public service announcements, 50 seconds of helpful hints, 25 seconds of quotable quotes, 12 seconds of inspirational messages, and up to 12 seconds of a warning that the feature is about to begin. Using various combinations of the above, any amount of time up to the maximum latency time may be filled with entertainment. When it is determined what the delay will be, the viewer could be advised of the time remaining before the next feature

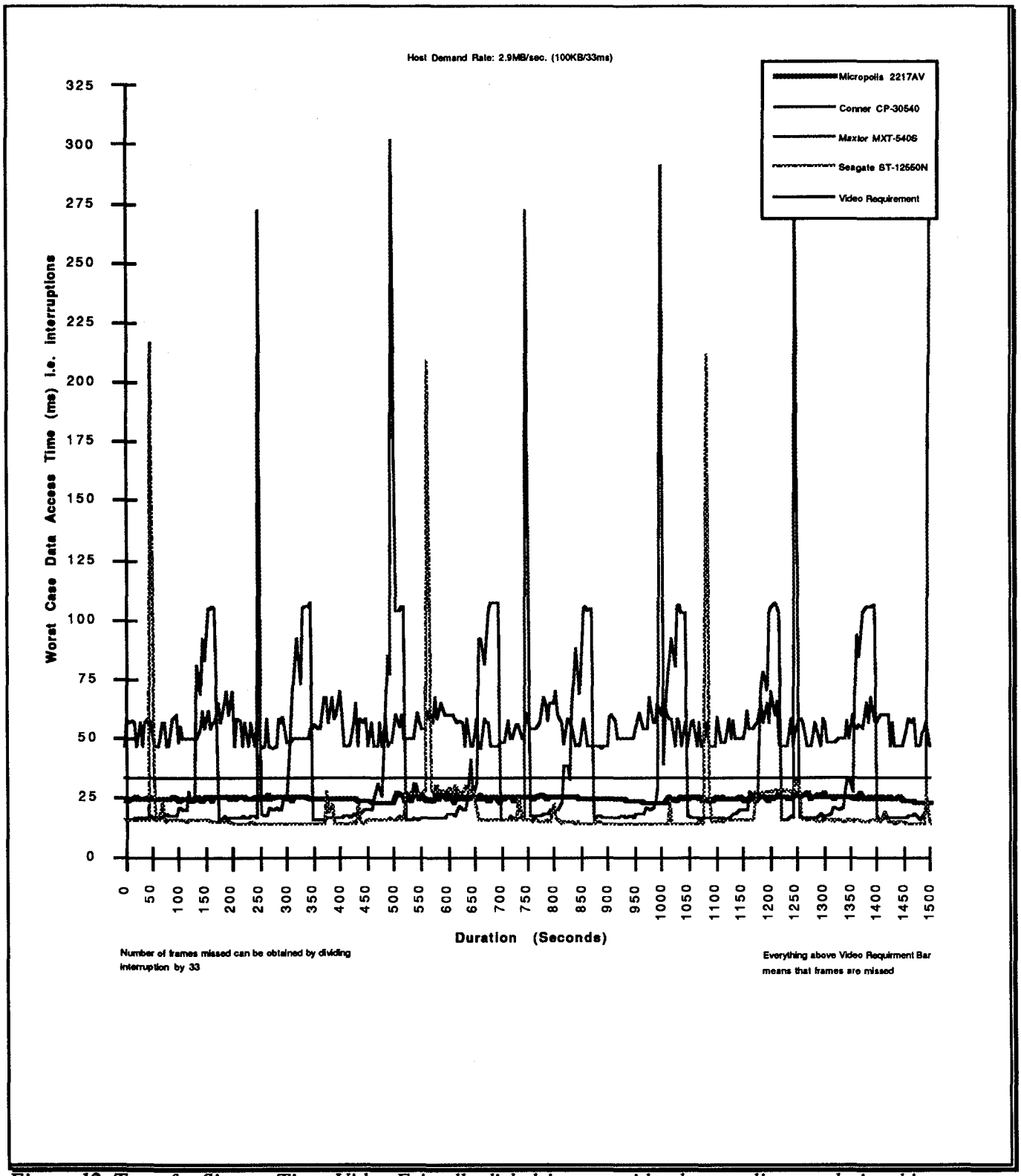


Figure 12. Transfer Size vs. Time. Video Friendly disk drives provide almost a linear relationship between transfer length and transfer size. The data in this graph includes command overhead, and is measured with the demand rate of 2.9 MB/s. This figure is presented courtesy of Micropolis Corporation.

starts, and given a menu from which he could select his own "fill in" entertainment.

Conclusions:

The ultimate goal of interactive TV is to provide the subscriber nearly instantaneous access to the programming of his choice. While this goal is attainable at very high cost, for a very limited number of subscribers, the authors do not believe it to be economically feasible to provide this type of service to the number of interactive TV subscribers projected over the next five years by leading industry market researchers.

NVOD offers a reasonable compromise between the ideal (zero viewer latency time)and an acceptable delay. This approach permits operators to obtain equipment which may be amortized by charges acceptable to subscribers.

---

[1]Pentium is a trademark of Intel Corporation

[2]Power PC is a trademark of Motorola Corporation

References:

(1) "Video On Demand: Architecture, Systems, and Applications", W.Hodge, S.Mabon, and J.T. Powers, Jr., SMPTE Journal, September 1993.

(2) "A Film Quality Digital Archiving & Editing System" W.Hodge, W.Harvey, and R.S.Block, SMPTE Conference on Advanced Television and Electronic Imaging for Film and Video, New York City, February 5-6 1993.