Evaluating Video Servers

By Al Kovalick

Video Servers are quickly replacing tape machines in a variety of applications. Enabled by A/V compression, networking and storage, servers have advantages over traditional tape transports. This paper will describe how to evaluate the various server architectures that are in use today. Servers are designed based on four distinct architectures; each method has advantages and tradeoffs. The evaluation also includes a discussion on reliable storage, and scalability issues.

Alexander Calder (1898-1976) is considered the artistic father of the mobile. A mobile hangs in mid-air, suspended from a single point. It is composed of individual elements that are in perfect balance with the whole. By way of illustration, let's use the mobile to understand how to evaluate broadcast products, particularly video servers.

Figure 1 shows an example of a mobile. The mobile has two distinct sides, both in balance. All of the elements represent either product features (right side) or the manufacturer's values (left side). The left side represents the extrinsic (outside the product) value of a product and the right side the intrinsic (inside the product) value. The combination of the two is the total value of a product. Ideally, this total value should equate to the price of the product.

Let's view the left side of the mobile. These are aspects of a product that are external to its hardware or software components. Certainly a vendor's vision, product track record, financial stability, and service/support are key to making a purchase decision. On the other hand, the right side displays aspects that are very specific to the functionality and benefits of the physical server. This article will focus on the intrinsic values and leave the extrinsic vendor analysis to your better judgment.

Presented at the 34th SMPTE Advanced Motion Imaging Conference (paper no. 34-14), in San Francisco, CA, February 3-5, 2000. Al Kovalick is with Pinnacle Systems, Mountain View, CA 94043.

The Intrinsic Elements

There is little debate that video servers are central in the operational environment of any modern broadcast facility. Day by day, servers are replacing videotape machines. Regarding server evaluation, what are the salient aspects for which one needs awareness? We will consider five main categories. They are noted on the right side of the mobile and in Fig. 2:

- Planes
- Architectures
- Storage subsystems
- Scalability
- Reliability

The Three Planes

In the world of Telecom, information technology (IT), and internet equipment, devices are often designed using the 3-plane* model. This model is ideal for describing the data, control, and management aspects of a device; the names are self-explanatory. Each plane offers specific functionality as shown in Fig. 2. Until recently, most broadcast equipment was not designed using this model, but industry awareness is changing this.

There are operational advantages to keeping the three planes separate. Each is composed of layers: physical, data-structures, protocol/framing and application. An SDI (SMPTE 259M) interface on a server may be considered a data plane component; it has physical, framing, and format layers. Today, the control plane is mainly proprietary command sets (Sony protocol, Louth protocol, etc.) over RS-422 links. This too is changing.

SMPTE is standardizing various dialects for machine control, and machine control over standard LANs is becoming a reality. Some servers offer only simple control for recording/playing. Others offer a rich functionality, including keying, output wipes, trimming, and low-resolution proxy viewing.

Industrywide, the management plane is the least mature of the three. Currently, in most cases, it is completely absent from broadcast-related devices, so what is its purpose? It is a portal into a device to configure and monitor all aspects of its operations. Again, the IT industry has taken the lead in this area. Broadcast equipment suppliers are just starting to include simple network management protocol (SNMP) and management information base (MIB) support in their products. SNMP and MIBs form the basis of the management plane.

When assessing a product, inquire about the specifics of each plane. As the industry moves forward, these planes will become fully standardized. At this time, the data plane is the most mature, followed by the control plane then, finally, the management plane.

Before leaving the topic of planes, it is worth mentioning the need for file exchange interoperability. Most servers compress the incoming A/V and store the content as either MPEG or DV files. Many scenarios require that files be transferred between servers over LANs and WANs. For this to work smoothly, the file types and associated metadata must be standardized. When evaluating a server, inquire whether the exchange format is a recognized standard.

*A plane in this context is a region of functional operation The regions are separate and offer different values to the user. For example, the data plane (region) has features that are completely different from the features of the control plane or managment plane.

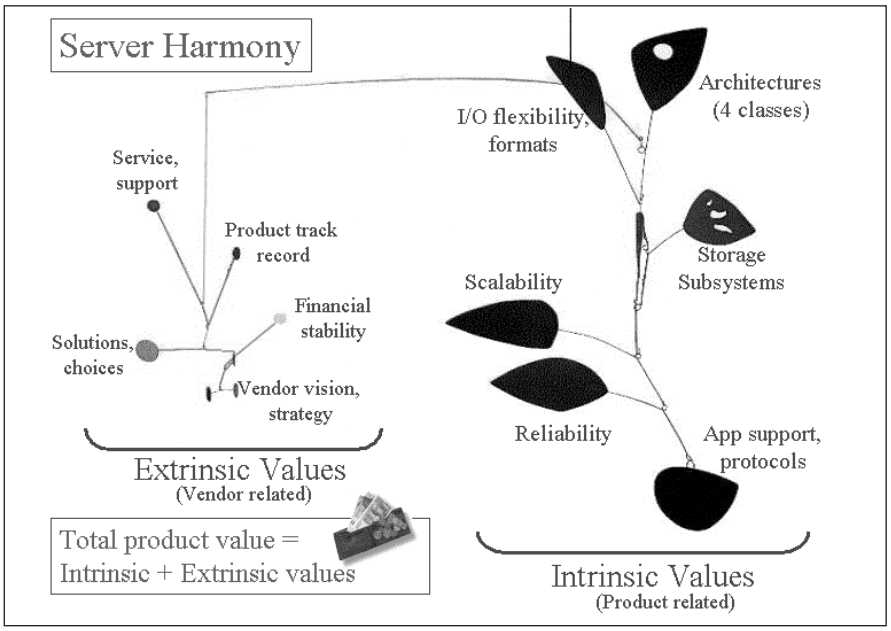


Figure 1. A mobile design illustrating the balance between vendor-related and product-related values of a video server.

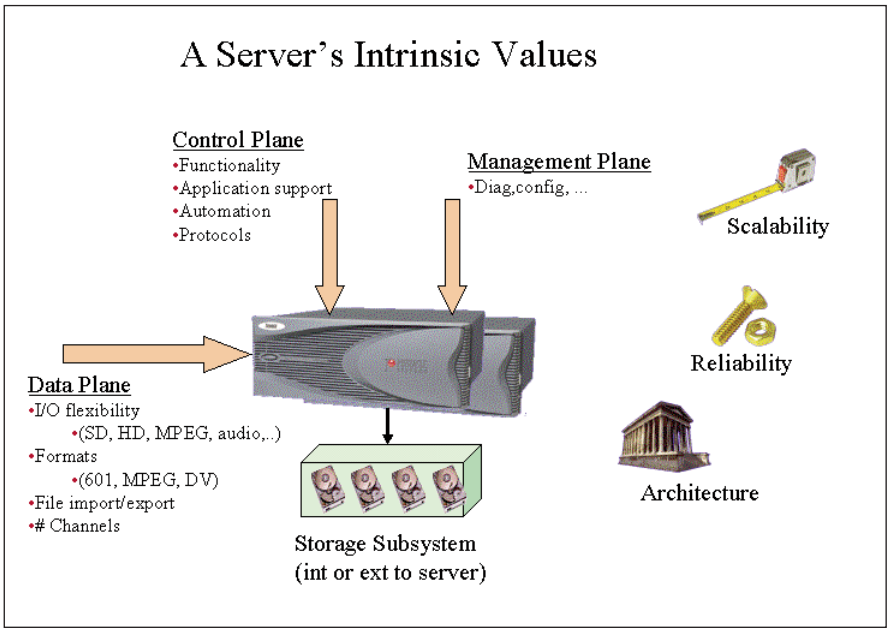


Figure 2. The intrinsic elements of a video server.

The Four Architectures

There are many manufacturers of video servers and each seems to claim some special advantage with their architecture. Are all servers really that different, or, are there some common themes by which all servers may be categorized? The following will outline the four fundamental architectures by which all A/V servers may be classified. Even these four have one common recurring theme.

Fundamental Server Class #1

Figure 3 depicts the simplest of all the classes, the computer-like class. It looks and acts like a computer but has specialized I/O, response time, and storage requirements. Typically, the central connecting bus and CPU horsepower limit the performance of this class. Some systems use a distributed bus/switch structure to move beyond a single bus's performance.

In addition, some systems use multiway processors to improve throughput. Of course, one can build a large server of this nature, but then the issues of cost, scalability, and reliability come into play. In general, servers of this class usually support less than 15 high-bandwidth (20 Mbit/sec) video channels.

In a traditional computer, most internal data traffic passes through the CPU. If the I/O cards are designed correctly, most storage related data transfers could bypass the CPU by moving directly from/to storage and I/O ports. This can increase bus performance by up to 2X.

Most servers of this class, use replicated components to achieve reliability. By doubling up on power supplies, fans, controllers, and storage drives, these systems achieve the status of high availability. However, even the most redundant systems can fail.

Fundamental Server Class #2

Figure 4 depicts fundamental Class 2. This is a cluster of Class 1 servers. Usually, the cluster is formed using a Fibre Channel loop or switch fabric. This class has several advantages. First of all, each node is an independent server: independence improves the fault tolerance of the entire cluster. A/V content may be encoded into any server and migrated to another under automation control. Most automation vendors can load balance the content across this class of server. Load balancing is a mature technology used everyday in broadcast facilities worldwide. This server class is ideal for the following applications:

- Multichannel, satellite feed recording.
- NVOD server farm.
- On-air and satellite playout of short and long form material.
- True bulletproof fault tolerant system.

One application space that is not ideal for this architecture is collaborative editing (news, sports, etc.). This often requires that all content be available to many editors simultaneously, as will be shown in Class 3, which is ideal for this application.

For applications that require from a few to hundreds of A/V channels, this class shines. As we will see, it scales beautifully and can be made to be

truly fault tolerant. For example, DirecTV's Los Angeles Broadcast Center uses a Class 2 MediaStream Server with over 175 A/V channels configured in a fault-tolerant architecture.

Fundamental Server Class #3

Figure 5 depicts server Class 3. This is fundamentally a storage-switched architecture.* Each I/O node connects to a common storage pool using a switched network. (Note the dotted box in the figure.) The contents within this box should look familiar; they describe a Class 1 architecture. One of the hallmarks of this configuration is the notion of a distributed file system.

In contrast to the Class 2 design where each node on the ring is an independent server, each one in a Class 3 system is dependent, especially regarding the file system. When a node imports or encodes a new video file, all other nodes must be immediately aware of the new file's presence. This is done without the assistance of traditional automation logic. Depending on file access permissions, any node can access any stored content; this is the strength of Class 3. It is most applicable when many users need access to the storage pool simultaneously.

The concept of a common file system "shared" by all nodes is nontrivial. There are several ways to implement such a file system. One popular way is to use a so-called metadata controller for storing file attributes and disk data-block locators. All nodes have access to the file metadata so all nodes share the same file system.

The metadata concept may be implemented in several different ways. One method uses a separate controller (computer) to store the metadata. For high availability systems, this controller should be configured as fault tolerant. Although the metadata controller is not shown in Fig. 5, it should be assumed to be present. Other methods place the equivalent of the metadata controller within a master node or distributed among the nodes.

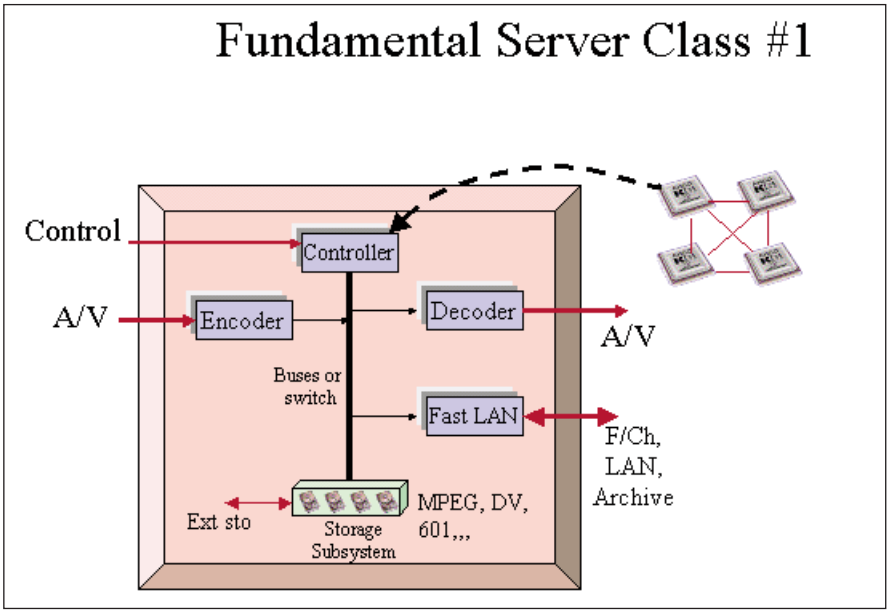


Figure 3. Class 1 server (computer-like).

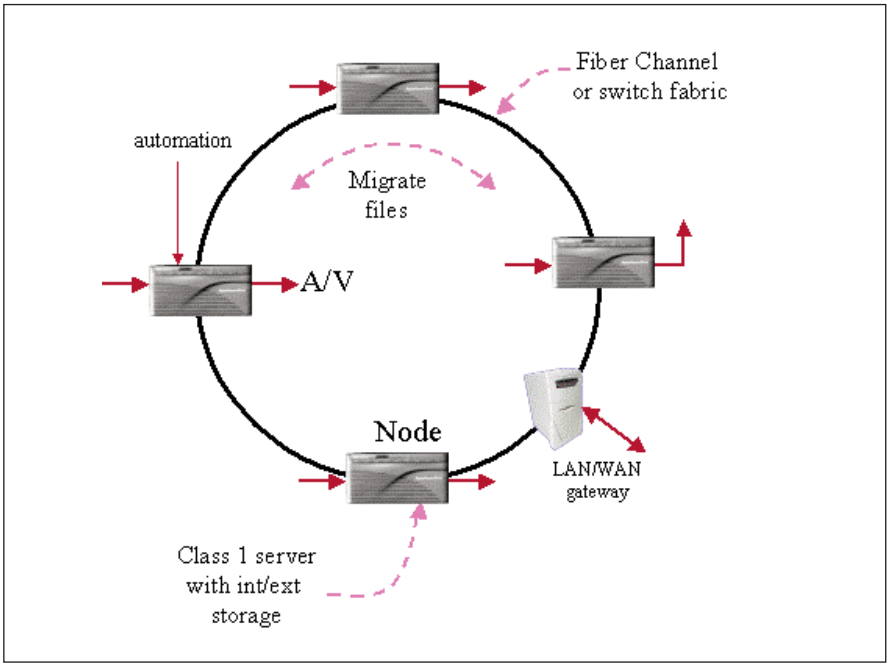


Figure 4. Class 2 server (clustered nodes).

The actual storage switching function may be accomplished via a variety of methods:

- Fibre Channel loop switching.
- Fibre Channel switch fabric.
- Ethernet switch.
- Direct point-to-point access (each node has direct access to any storage node).

For correct load balancing, all the A/V content is usually striped across all the discs. To a large extent, the performance of this architecture depends on the nature of the switch. By evenly distributing all stored files, the available access bandwidth is increased thereby supporting more nodes. Even then, without disc access regulation rules, it's possible for data queuing problems to arise.

Striping creates another form of dependence. If the storage subsystem

*In general, a storage-switched architecture is also called a Storage Area Network (SAN). Another variation on this theme is called Network Attached Storage (NAS). With NAS, the storage subsystem is replaced by a file server with its own storage subsystem. The distinction is relevant but not important for this paper. Potentially, a NAS-based server could support hundreds of channels.

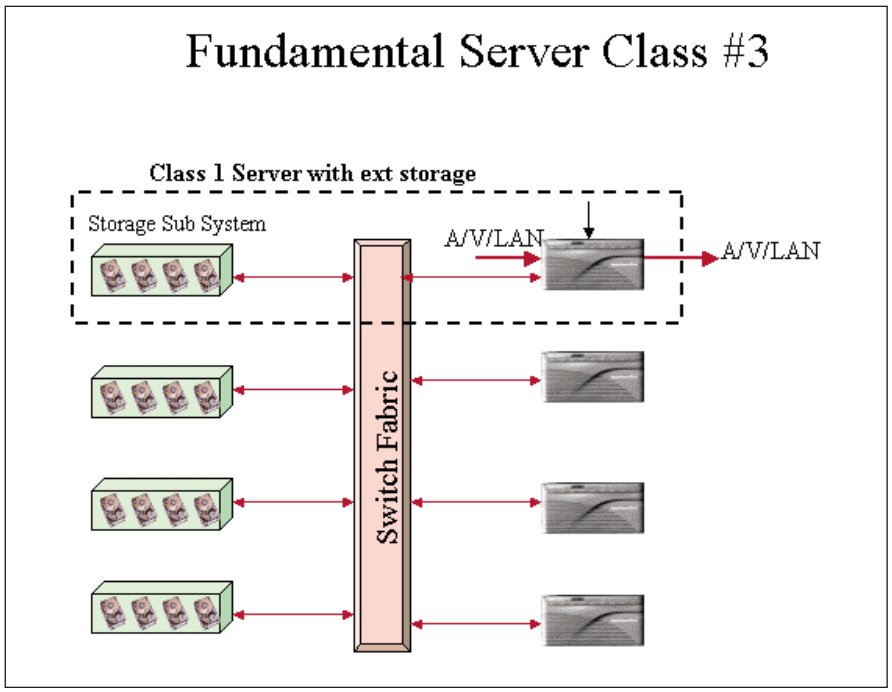


Figure 5. Class 3 server (storage switched).

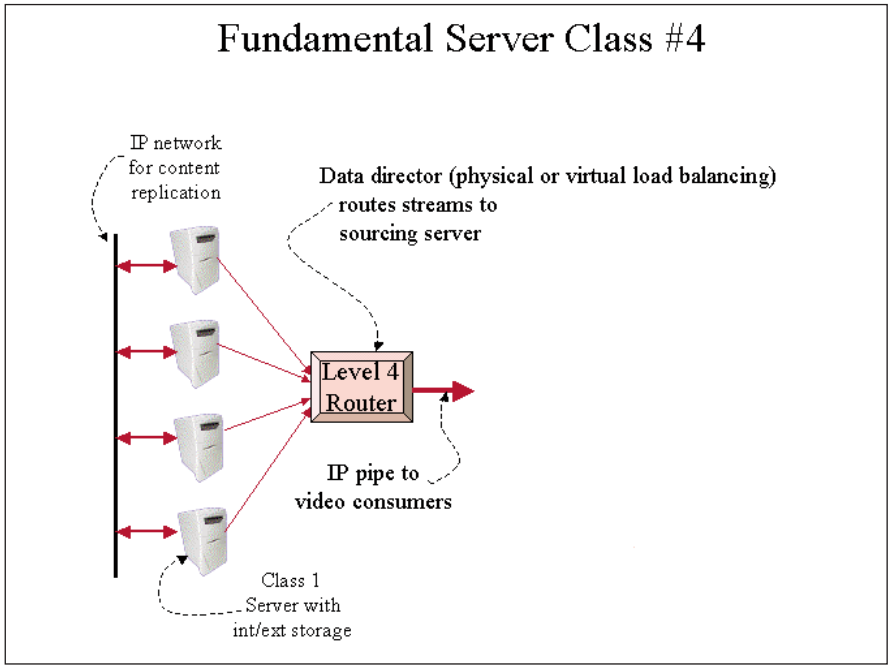


Figure 6. Class 4 server (clustered web video server).

ever fails in a catastrophic way, then all nodes lose access to all data, thus killing the entire server. There are strategies to reduce the effect of single point of failure components, but this increases the complexity of the solution.

Building large, reliable Class 3 systems is nontrivial. More nodes usually require additional common storage and more access bandwidth through the switch fabric. There are other exotic variations on this theme, but most practical systems (<100 I/O channels) will adhere to the principles outlined here.

Fundamental Server Class #4

Figure 6 demonstrates a Class 4 server. This class finds application in internet video streaming, but most broadcasters will not have one of these in their facility. However, as our industry embraces the internet as the "new antenna," it will rely on video service providers, who will use this class to serve an unlimited number of simultaneous streams.

This class is composed of a cluster of Class 1 servers. Viewers are directed to a selected server node by a so-called level 4 router. A level 3 router routes at the IP level. A level 4 router routes at the application level (TCP for this discussion) and uses various strategies to load balance user requests.

It should be noted that a physical router is not the only way to load balance the nodes. Another popular strategy is to use features of a Domain Name Server (DNS) in assigning IP addresses for the destination node servers. Regardless of the method used, the intention of load balancing is to populate each server node with the same number of users, on average. The nature of IP and of the internet in general, allows for this class to be physically distributed over a wide geographic area. It is not uncommon to have the nodes in different locations or even in different countries!

You may have noted a problem with this method of streaming video. In a worst case, every node must have identical stored content, but is this practical. Web video files (usually Microsoft, Real Networks, or QuickTime formats) are small in size and duplication of content is not a major burden. Several companies offer application software to automatically load balance each node's stored content.

Storage Subsystems

Another component of the balanced server is the storage subsystem. In general terms, there are three types of disk storage systems. The first is the single disk drive. As of July 2000, 72 GByte drives are available and 144 GBbytes, or greater, are due in 2001. A good figure to remember relates to 10 Mbit/sec encoded (say, with MPEG-2 compression) A/V content. It requires 4.5 GBytes/hr to store 10 Mbit/sec content. So a 72-GByte drive can, excluding various overhead factors, store 16 hr of compressed

video/audio.

A step above the single disk is just a bunch of disks (JBOD) array. This is usually an array of 8 to 10 disks on a common Fibre Channel loop. The frame may have dual power supplies. This method yields a linear increase in storage, 160 hrs with ten 72-GByte disks, and a nearly linear increase in read/write (R/W) disk bandwidth. In addition, all the content is usually striped across all the disks in the array. Striping increases the available array bandwidth to ~N (number of disks in array) times the bandwidth of an individual disk.

A JBOD array is in jeopardy of losing its stored content if one or more disks fail. RAID (Redundant Array of Independent [or Inexpensive]) Disks comes to the rescue. There are many types of RAID, and a high confusion factor associated with it. RAID level 0 defines data striping—distributing the file as chunks across several disks. RAID level 1 defines duplicating the contents (mirror) of one drive on a second. RAID levels 2 to 5 use parity data to recreate missing data. Most storage subsystems that claim fault tolerant disks use some form of RAID. Let's demystify it.

If RAID is magic then the parity concept is the power behind the trick. For example, in Fig. 7, the top row has seven coins, two heads and five tails. The 8th position is a binary marker (parity) indicating whether there are an even or odd number of heads among the seven coins.

The second row shows the case where a coin's identification is unknown (bad bit). Using the parity bit and a small amount of logic, it is easy to deduce that the missing coin must have been tails since the parity bit indicates an even number of heads, it's that simple.

This specific example only works when there is one bad bit, including the parity marker. In practice, this method of data protection may be expanded and is sufficient to reconstruct bad bits, bytes, words, drives, and entire subsystems. Do not be fooled by marketing hyperbole: the exact type of RAID is not that important. What is important is that the server can transparently restore bad data in realtime even under the worst case operating conditions. A fault tolerant storage subsystem may be designed using dual RAID controllers.

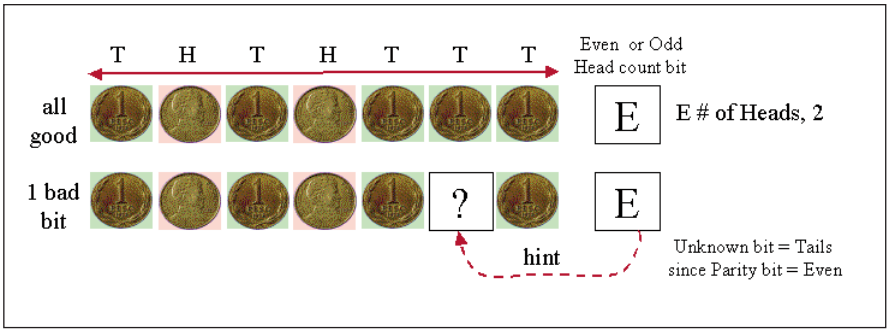


Figure 7. Storage systems: RAID demystified.

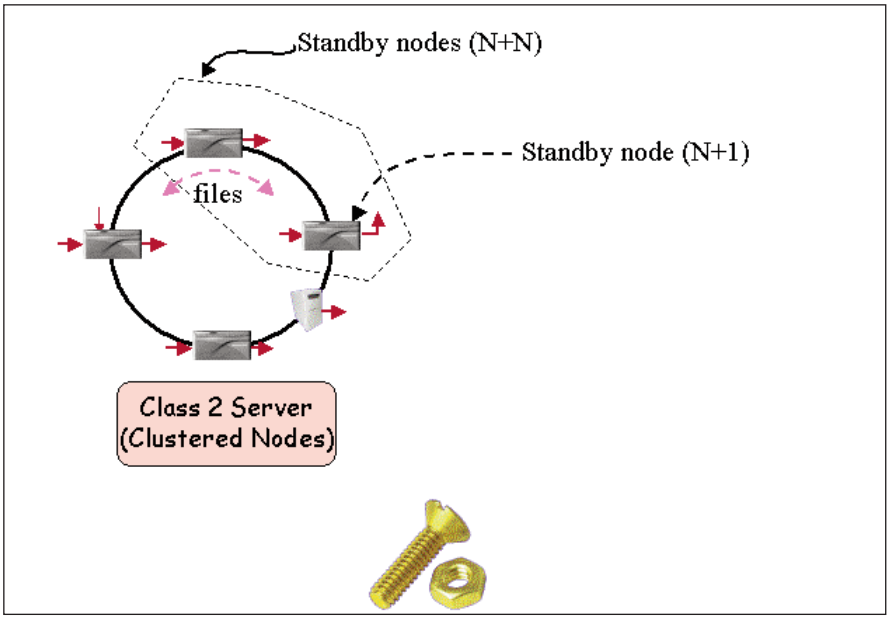


Figure 8. Achieving reliability using (N+1) and N+N sparing.

Scalability

So, you have decided that you need a small server for your facility. Seems that a 2 x 4 (2 in, 4 out) is ideal. Six months after the purchase, your boss asked if the server could be upgraded to a 2 x 8. Can it be? This is a scalability issue. What factors affect scalability? Which Class offers the best scalability? Let's see.

Let's start with small servers. For a small number of I/O ports, the Class 1 server is the most practical. If you purchase a Class 1 server, assure that it can become a node in a Class 2 or 3 system; this way, you can expand indefinitely. Class 2 scales into the hundreds of ports by adding small nodes to the ring or switching fabric (Fig 4). The ring/fabric may be of a fault tolerant nature.

Class 3 scales by adding nodes and separate storage. The SAN switch throughput and reliability will limit the number of nodes in practice. Expanding this class's storage may require a complete restriping of all content across all storage arrays. This is a nontrivial task. Scaling complexity increases considerably beyond about 100 channels but also depends on the data throughput per channel. Finally, Class 4 can scale to support terabytes of storage and millions of viewers.

Reliability Issues

Many factors affect the reliability of a server system:

- Server and automation software robustness.
- Software complexity and maturity.
- Storage protection strategy.
- Redundant components: fans, power supplies, codecs, etc.

• Redundant nodes, all classes can use this: (N+N) (true mirror) or (N+1) strategies.

In practice, software fails more often than hardware. A world class server is a complex device. It may include more than 50 man-years effort of specialized software, so how does one rate the reliability of this software?

Look for these basic things: the track record of the server, complexity of the design, and maturity of the solution including automation control. When in doubt concerning a manufacturer's claims, decide based on these fundamentals.

Hardware also fails, and there are really only two methods to achieve true fault tolerance. A Class 1 server cannot, in practice, be truly fault tolerant. There are designs that approach the ideal, but they are very expensive and do not scale to a small number of I/O. You need to look to Classes 2, 3, and 4 to achieve true high availability.

The two methods that work in practice are a true mirror (N+N) and the (N+1) approach. Here N is the number of live active nodes in a cluster. Figure 8 shows a Class 2 system with four nodes. With a mirror approach, imagine that two nodes are active (N=2) and two in standby. The standby nodes have identical stored content to the active nodes. If an active node fails for any reason, a standby node is called to resume its work schedule. Many automation vendors support this mirror approach. Replication by 2X is costly, but it buys piece of mind because of its simplicity and guarantee of performance.

A more efficient way is to add only one standby server node, hence, the (N+1) identification. Let's assume that one active node fails. If the standby node has copies of the stored A/V content of the other three, then it can take over the workload of any faulty active node. True, storage must be duplicated, but this is not as onerous as you might think. For example, a 4-server system in an (N+1) schema adds only 10 to 15% to the total system cost due to the extra storage on the spare node, and the cost of storage compared to the that of the other system components is falling rapidly. This is a very reliable configuration with virtually zero chance of a total system failure.

For a Class 3 system, the (N+1) approach also works well; however, remember that this type of architecture has dependent nodes compared to the independent nodes in Class 2. The storage subsystem and switch need bulletproof reliability as well.

Despite the cost effectiveness of the (N+1) approach, many facility designers still choose to use a true mirror (N+N) for either Class 2 or 3 systems, because it has a simple architecture and automation control during node failure and is very mature.

Classifying Contemporary Servers

As the themes were being developed in this paper, you may have tried to categorize various commercial servers into their classes. First of all, each class has its own strengths and weaknesses, and no one class is right for all applications. The purpose of this paper is to provide insight into server evaluation, not to malign a particular server family. In this spirit, what classes do the various commercial servers fall into? By way of example only, the following is representative of the current state of the art. For brevity, not every commercial server is included. Also, some manufactures offer servers in more than one class.

• Class 1: Pinnacle, GVG, Sony, Panasonic, and several others.

• Class 2: GVG's Profile Server, Pinnacle's MediaStream Server.

• Class 3: Leitch VR Series Broadcast Video Server, SeaChange MediaCluster, Pinnacle's Vortex News.

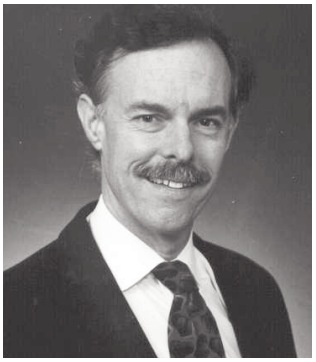
• Class 4: Hewlett-Packard, Sun, Compaq, IBM, Dell, etc.

Of course, hybrid variations exist. It's possible to create a Class 2 architecture out of Class 3 "nodes." There is no end to the esoteric combinations. Keep your solutions simple and the likelihood of a trouble free system improves.

Conclusion

The mobile illustration of the balanced server is useful for evaluating products of all types. With the concepts developed in this paper, you should be able to improve your aim when asking questions about a particular server. Do not overlook any of the elements that make up the mobile and you may indeed find the "perfect" server for your needs.

THE AUTHOR



Al Kovalick worked for Hewlett-Packard for 25 years as a designer, system architect, and technical strategist. From 1980 to 1992 he specialized in signal processing systems, hardware and software design for realtime signal analysis and synthesis. He joined Pinnacle Systems as the CTO of the Broadcast Solutions Division in 1999.

For the last eight years, Kovalick has concentrated on digital video for broadcast, post-production, and video-on-demand applications. He is involved in strategic planning, technology assessment, standards, and systems architecture for video servers and other broadcast products.

Kovalick is actively involved in SMPTE, contributing to various standards working groups. He took a leadership role in writing the SMPTE 273 Standard on Status Monitoring and Diagnostics. He was awarded a Journal Certificate for a paper published in the *SMPTE Journal*, "Reference Architecture for Digital Video/Audio Transfer and Streaming," Aug. 1998.

Kovalick was a founding member of DAVIC, the SMPTE/EBU Task Force, and the ProMPEG Forum. He has been awarded 18 patents.