

FAST Solves Disk Play-to-Air Reliability

With Flash Assisted Storage Technology (FAST), broadcasters get both high performance and high reliability

The Flash Assisted Storage Technology (FAST) architecture combines flash memory with proven SeaChange® MediaCluster® technology to remove unreliable spinning disks from the most critical part of the on-air infrastructure. A “green” solution, FAST offers higher availability, consumes less power, is less noisy and reduces the cost of long-term maintenance.

Disk reliability has long been an issue for broadcasters. When disk failures occur in play-to-air servers — the most critical part of the on-air infrastructure — it can mean going “black to air” for millions of viewers. Even with the redundancy of mirrored or parity-protected configurations, broadcast engineers must still wait for disks to rebuild while hoping the rest of the system stays intact. Now a new study by Carnegie Mellon University researchers confirms what broadcasters have suspected: disk drives fail at rates six times higher than those reported by vendors.¹

Things will get worse with HD. Moving from 15Mbps SD to 50Mbps HD, the same TV show will take up to three times more bandwidth. Three times as many disks will make it three times more likely that a disk-based server will fail.

To bring failure rates down, vendors have tried replacing disk with solid-state solutions — i.e., devices with no moving parts, which is where almost all disk failures occur. Flash is an obvious candidate since — like disk — it holds data after the power goes off. Flash also consumes less power than disk, produces less heat, and is less noisy.

That’s the thinking behind a recent SeaChange innovation — Flash Assisted Disk Technology (FAST). FAST has something other vendors don’t: MediaCluster®. FAST is essentially MediaCluster with flash modules replacing disk drives on all play-to-air servers. Due to unique data striping — over all nodes in a cluster and all modules in a node — both I/O and reliability are extremely high. The solution’s managed reads and writes optimize performance and avoid write hot spots that can burn flash out prematurely. FAST is also very economical because it combines a disk-based near-line MediaCluster for ingest with the flash-based play-to-air cluster for high availability.

¹ Bianca Schroeder and Garth A. Gibson, Computer Science Department, Carnegie Mellon University. “Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?” *In Proc. of the FAST’07: 5th USENIX Conference on File and Storage Technologies*, 2007.

DISK'S INHERENT RISK

Broadcasters' demands for a better way to store content follow years of dissatisfaction with what many see as high disk failure rates. Now research at Carnegie Mellon University has confirmed that disks do suffer very high failure rates.

Using vendor RMA (return merchandise authorization) data, the researchers measured actual disk failure rates in the field on two key benchmarks: annual failure rate (AFR) and mean time to failure (MTTF). Among the findings:

- *Disk AFRs typically exceed 1%, with 2-4% common and up to 13% on some systems*
- *Field replacement rates of systems are significantly higher than expected based on datasheet MTTFs (by 2-10 times for drives less than five years old)*
- *The rate at which disk drives fail rises steadily throughout their lifetimes, starting as early as the second year of operation, rather than holding flat as is widely expected*

By comparison, the AFR for flash drives (also from RMA data) is just 0.04% — a 100 times improvement. Broadcasters should therefore expect to replace flash drives far less often than disks. That's because total risk of system failure due to a disk failing equals the risk of one drive failing multiplied by the number of drives. In

other words, it would take at least 100 flash drives to have a combined risk equal to just one disk drive.

Some risk mitigation strategies, such as disk mirroring (RAID1) and disk rebuilding (RAID5 or RAID6), simply throw more disks at the problem. These strategies don't address disk's underlying inherent risk. One result is that these rebuilding servers are in a "degraded state" where a second disk failure may take out the entire on-air operation. Another is that the strategy itself may not work — particularly since a RAID rebuild assumes 100% data integrity on all remaining "good" disks in the server — not always a safe assumption. And even when these strategies do work, there is no compensating value (like faster I/O) to offset the costs of adding the extra disks. High failure rates pose a significant challenge in the SD environment — and even more so as broadcasters move to HD.

HD requires three to 10 times as many disk drives as SD to provide HD bandwidth and store the same number of program hours. The likelihood of a play-to-air storage failure will increase in proportion to the number of drives added and the age of the drives. If one server has an AFR of 25% then a mirrored configuration's AFR is slightly less than 1%². To achieve the five nines availability (99.999%) broadcasters expect, the AFR must be less than 0.25% — a near impossibility in light of recent research.

² Assuming 48 hours Mean Time Between Repair.

FLASH HAS LOWER INHERENT RISK

The advantage of flash drives is that they start with a much lower AFR per unit — less than 0.04% — over millions shipped in consumer devices such as iPods and cell phones. Table 1 shows how that low flash failure rate translates to a low cluster annual failure rate (0.23%) even without RAID5 protection on the chassis level. Clusters ranging from three to nine nodes show five nines reliability, with nine nodes being the worse case. A nine-node cluster holds about 10TB of data on 24 flash memory cards, each of which holds 64GB.

The total reliability of the MediaCluster equals the sum of the subsystem failure rates, which in each subsystem equals the failure rate of each component multiplied by the number of those components (minus any redundancies). The subsystems within each node are the motherboard (1), GigE I/O cards (3), flash memory drives (24), power supplies (1+1 redundant), and fans (5+1 redundant). Component failure data is based on either RMA data or MTBF reported by vendors.

Major Subsystem Components					
	Motherboard	I/O Cards	Flash Memory	Power Supply	Fans
MTBF (hours)	105,000	408,000	21,900,000	150,000	170,000
AFR	8.34%	2.5%	0.04%	5.84%	5.15%
Number of Components	1	3	24	2 (1+1)	6 (5+1)
Subsystem AFR	8.34%	6.44%	0.96%	0.002%	0.01%
Subsystem MTBF (hours)	105,000	408,000	912,235	468,750,000	120,416,667
Single Node AFR	23.14%				
Single Node MTBF (hours)	37,850				
Single Node Availability	99.873%				
9-node Cluster AFR	0.23%				
9-node Cluster MTBF (hours)	3,730,871				
Cluster Availability	99.999%				
Definitions: Failure rate = 1/MTBF AFR = (number of hours in a year) * (failure rate) Number of hours in a year = 8,760					

Table 1: Computing the Reliability of a FAST Storage Solution

In this example for a SeaChange Broadcast Flash Memory Library FML200, availability of each server node is 99.873%. However, reliability of the worse case nine-node cluster as a whole is still 99.999% — a difference that is directly attributable to FAST.

MEDIACLUSTER MAKES FAST WORK

FAST is based on MediaCluster — a technology proven since 1996 — except that, as in the new FML200, flash memory drives replace disks. The key insight is to carefully manage reads and writes to each flash drive so its performance is optimized. Media content data are striped contiguously in two ways — across all drives in a server and also across all servers arrayed as nodes in a cluster. All content has equal and parallel access to I/O, so high ingest bandwidth is achieved while maintaining full playout performance.

With up to 24 64GB flash memory drives in each node, a nine-node FML200 cluster can scale up to over 500 hours of HD content @ 35Mbps XDCAM video (or 50Mbps video and audio combined). Single copy flash memory storage is shared among all the nodes in a cluster with N+1 redundancy, avoiding costly mirroring. And because all data, including parity data, is evenly distributed, no dedicated parity drives are needed. Service continuity is protected even if a node fails or during in-service maintenance, hot swapping of drives, system upgrades, or installation of additional base nodes.

MediaCluster also solves the problem of flash write hot spots — when writes repeatedly hit the same flash memory causing it to burn out quickly. MediaCluster eliminates write hot spots

by evenly load balancing writes across all flash modules so the same memory location may only see a few writes per day, if that. This extends the memory's lifetime to more than 10 years versus the typical five years for hard disk drives.

MediaCluster performance and reliability also come with the greater versatility inherent in solid-state devices. Flash, for example, lends itself to highly granular deployments. So instead of one large central on-air storage cluster serving all outputs, several smaller clusters can each serve a few channels, further reducing overall risk exposure. These clusters can even support “disaster recovery” sites at uplink or transmitter locations where day-to-air content is refreshed daily via WAN or satellite. Flash also tolerates environments considered too harsh for disk — so it's safer to broadcast from a moving truck, car, helicopter, airplane, or ship. That versatility is further enhanced by flash memory's green value — a 10-to-1 advantage in power efficiency over disks.

THE SEACHANGE BROADCAST FAST ARCHITECTURE

SeaChange's broadcast FAST architecture complements how most broadcasters allocate content within their infrastructures. That is to partition content between two types of storage — near-line and play-to-air. The key difference between these storage tiers is that near-line is optimized to store large amounts of content at relatively low cost. Play-to-air is optimized for high bandwidth and high reliability.

In a typical broadcast workflow, content is ingested into low-cost, SATA-based near-line storage. As content gets closer to airtime (typically within 24 hours), automation moves

it from near-line to play-to-air storage. The link between the two must be sufficiently fast to transfer large volumes in greater than real time speed. Once the content is played and no longer needed, the automation deletes it from play-to-air storage.

In a FAST architecture (Figure 1), a majority of content is stored on a product like the SeaChange MediaLibrary™ 1G (ML1G) — a disk-based MediaCluster solution. As content comes within 24 hours of broadcast it is transferred to a flash-based solution, such as the FML200 storage cluster, using very fast I/O paths between near-line and flash storage. In this environment, no disk (and its inherently higher risk) can ever impact the broadcaster's play-

to-air performance. Extremely high play-to-air reliability and bandwidth are therefore achieved without sacrificing economy. At the same time, broadcasters benefit from all the advantages they would expect from a flash-based solid-state solution such as lower power consumption, lower noise volumes, less heat and less maintenance cost. Perhaps best of all, no longer will they need to wait for a RAID disk rebuild to finish before they know if their station will stay on the air.

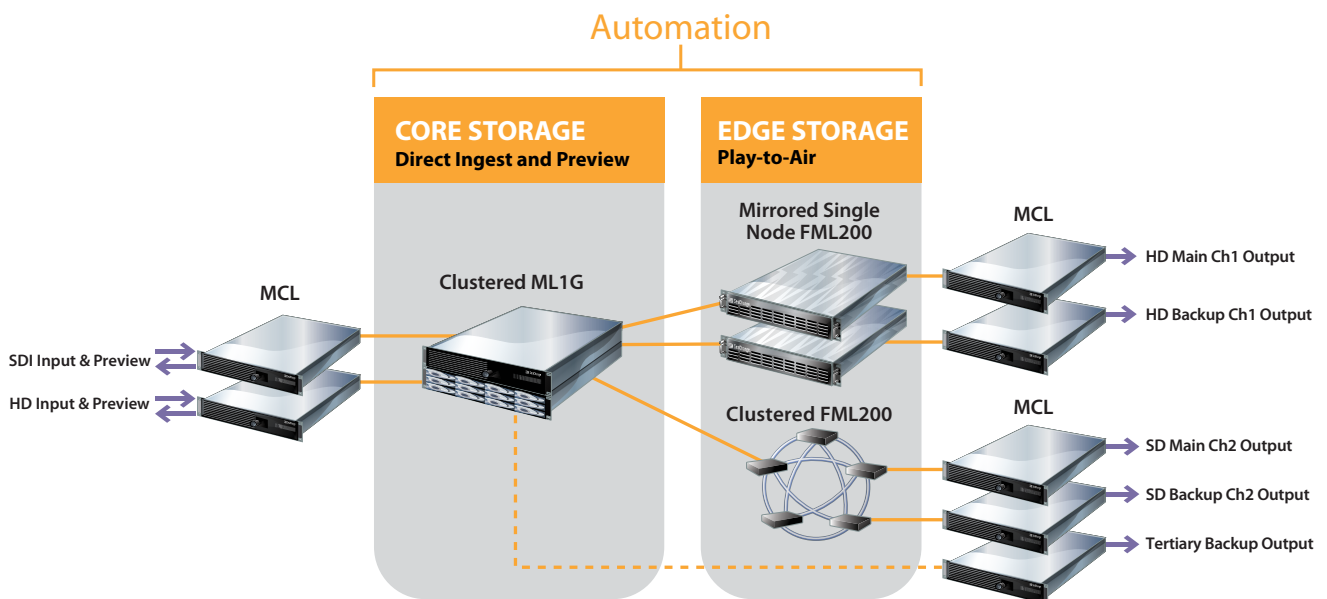


Figure 1: The FAST Architecture

BENEFITS

Most Reliable and Economical Solution – Combining flash memory with SeaChange’s patented and widely deployed MediaCluster® streaming technology, the FML200 delivers an ingest and Transmission server solution that exceeds 99.999% availability requirement, bringing a new level of reliability to the marketplace. Supporting flexible numbers of SeaChange MediaClient codecs, the FML200 provides broadcasters with a wide range of options to configure their on-air operations.

Reduced Operational Costs – Broadcasters have been so far burdened with the high operational costs associated with disk failures, disk replacements and long disk rebuild times as well as high power consumption/thermal dissipation, large space requirements and noise. Compared to spinning hard disk drives, flash memory:

- has no moving parts
- is 100 times more reliable
- consumes 10 times less power

With the FML200, broadcasters can now avoid disk failure and the headaches associated with disk rebuilds. With less power to consume, less heat to dissipate and less noise, the FML200 will move broadcasters a step closer to the green operating environments that everyone desires.

High Channel Density and Small Form Factor – The FML200 comes in a compact 2RU package, capable of supporting up to 3 GigE IP accelerator cards. Of the three accelerator cards, two of them can support large numbers of SD and HD MediaClient codecs. The third accelerator card supports FTP/CIFS file transfer. Each FML200 server can host up to 24 flash memory drives. As the flash memory drives’ capacity expands, the FML200 will further increase its storage density.

Full HD Now – The FML200 is fully compatible with SeaChange’s widely deployed broadcast-quality MediaClient codecs. With the latest MCL 6000 software-based flexible codec, broadcasters have the flexibility to start in SD and upgrade to HD software or simply go HD directly.

Fully Integrated with SeaChange’s RAID2® Disk-based Storage Servers – SeaChange provides smart automation-based file prefetching between the FML200 and the backend SeaChange servers (e.g. online BMLex) and also supports FTP for ingest and API control.

USE CASES

Day of Air Server – provides high reliability right where you need it: on your Transmission Payout Server. Our massive Near-line Archive offering supplies the Green Machine continuously with new content.

Divide and Conquer – you can use several small Green Machines to support only a small number of outputs each. You would not need to put all your playout “eggs in one basket”.

Remote Payout Server – you can place a number of ‘Disaster Recovery’ channels at your uplink or transmitter site, forward the content via your satellite link or WAN to the transmit site, store day-of-air content in the Flash Memory Server. In the event you lose your playout facility due to fire, flood, power loss, etc., the Green Machines will be ready to play your critical programming to air out of the DR site.

Portable Record and Payout for Mobile Applications – due to its high reliability, it thrives in environments that you would not dare to install a disk-based storage device, such as in a moving truck or car, in a helicopter or airplane, or ship-board.

SPECIFICATIONS

SERVER SPECIFICATION

- 2U rack-mountable server
- 64GB flash memory drives
- 24x drives per server
- 1.04TB usable storage per server with 64 GB flash memory
- 109 hours of 15Mbps SD (~21Mbps video plus audio)
- 35 hours of 50Mbps HD (~65Mbps video plus audio)
- Redundant power supply

SOFTWARE AND CAPABILITIES

- Cluster Management GUI and command line interface
- System snapshot for upgrades, server replacements, and cluster expansions
- SNMP and alarms package
- Jumbo frame support for optimal bandwidth
- Supports SeaChange MediaClient for both SD and HD

ABOUT SEACHANGE

SeaChange International is a leading provider of software applications, services and integrated solutions that deliver a high-quality television experience across TVs, PCs and mobile devices. By partnering with leading cable and telco companies, SeaChange enables in-home and mobile entertainment, as well as advanced advertising solutions, allowing broadband operators to differentiate their offerings and create strong customer loyalty.

Visit www.schange.com today.



SeaChange International, Inc.
50 Nagog Park, Acton, MA 01720 USA
T 1.978.897.0100 F 1.978.897.0132
www.schange.com

3.26_2010

©2010 SeaChange International, Inc. SeaChange, MediaCluster and RAID² are registered trademarks and MediaLibrary is a trademark of SeaChange International, Inc. All other marks are the property of their respective owners. While every effort is made to ensure the information given is accurate, SeaChange does not accept liability for any errors or mistakes which may arise. All features, specifications, system requirements and/or compatibility with third party products described herein are subject to change at any time without notice.