high availability and data movement

in the hp va 7000 series arrays

hp — white paper

for information about the va 7000 series
and periodic updates to this white paper
see the HP SureStore website at
http://www.hp.com/go/storage

i n v e n t

Hewlett-Packard Company

Hewlett-Packard Product Information

high availability and data movement – in the hp va 7000 series arrays

Published: July 2001

Revision level 1.1

For the latest updates to this document see
http://www.hp.com/go/storage

Warranty

This document is supplied on an “as is” basis with no warranty and no support. Hewlett-Packard makes no express warranty, whether written or oral with respect to this document. HEWLETT-PACKARD DISCLAIMS ALL IMPLIED WARRANTIES INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.

LIMITATION OF LIABILITY: IN NO EVENT SHALL HEWLETT-PACKARD BE LIABLE FOR ERRORS CONTAINED HEREIN OR FOR ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES (INCLUDING LOST PROFIT OR LOST DATA) WHETHER BASED ON WARRANTY, CONTRACT, TORT, OR ANY OTHER LEGAL THEORY IN CONNECTION WITH THE FURNISHING, PERFORMANCE, OR USE OF THIS MATERIAL.

The information contained in this document is subject to change without notice.

No trademark, copyright, or patent licenses are expressly or implicitly granted (herein) with this white paper.

Disclaimer

All brand names and product names used in this document are trademarks, registered trademarks, or trade names of their respective holders. Hewlett-Packard is not associated with any other vendors or products mentioned in this document.

Table of contents

Designing for availability in a RAID array

Driving factors for RAID solutions

The drive for uninterrupted availability of data has created a class of fault tolerant storage technology. The RAID array takes a central roll in enabling unperturbed access to data. For a growing population of workers, loss of access to data translates directly to lost productivity. The tolerance for a loss of access to data stored on computers diminishes on a daily basis. Users of data push computers to provide non-stop availability and access to data. These machines continue to assert themselves as critical components in most business operations.

Taking a look at the technology

The term “RAID” stands for “Redundant Array of Independent Disks.” At its most basic level, this indicates that the system must store the data so that the loss of any one disk does not cause loss of data. However, many other elements play a roll in maintaining access to data. Multiple data paths and routing give the system means to work around internal electrical failures.

Variations in machinery

The VA 7000 series creates a system that enjoys a high level of fault tolerance and robustness, with numerous checks and protection of data movement.

Hewlett-Packard recently introduced a new family of RAID disk arrays known as the VA 7000 series. From a visual inspection point-of-view, the electronics in the VA 7000 series may not look very different from electronics in any other product–regardless of the application. For example, a controller board in the VA 7000 series may look similar to a motherboard in a PC. However, the care taken with the detailed operation and functionality of the VA 7000 series creates a system that enjoys a high level of fault tolerance and robustness.

This paper discusses the numerous checks and protection mechanisms surrounding the movement of data in the VA 7000 series. Any one check may or may not be implemented in any other RAID Array–the customer can tell only when the lesser device fails to function correctly.

The significant investment made by Hewlett-Packard produces reliable, efficient products.

From the highest-level view, a RAID array looks very much like something that just moves data from one point to another. Loose designs employing *off-the-shelf* RAID devices can construct the façade of availability, without sufficient rigor and robustness to truly protect the user’s data. However, Hewlett-Packard invests many millions of dollars and resources toward the development of rock solid products. ASIC integration, while time consuming, brings both a cost advantage and a validation capability to the VA 7000 series. In fact, not one byte of data moves through the VA 7000 series without the full confidence of HP’s validation circuitry.

Sharing the knowledge

Customers benefit from the robustness and availability provided by this development effort. This paper examines only a few of the availability characteristics associated with the VA 7000 series. The reader takes a journey through the VA 7000 series as they follow a ***write*** and a ***read** transaction*. Written for individuals interacting with fault-tolerant storage devices, this paper describes how and why users will benefit through HP's extensive efforts to design for reliable, highly available operation of the VA 7000 series disk array.

The life of a write transaction

Receiving the write request

The interface from the host to the VA 7000 series is ***Fibre Channel***. The first step for a *write transaction* is the SCSI command that indicates the host system wants to write data to a particular location on the array. This simple command travels through the *cyclic redundancy checker* (CRC) protected medium provided by Fibre Channel. The command then continues on its way to the processor's local memory through parity protected data buses. Should the command be corrupted either on the Fibre Channel interface or on its way to the processor, the protection mechanisms built directly into the hardware will not allow the corrupted command to propagate through the system.

Processor executes from ECC protected memory

The VA 7000 series uniquely handles the write operation, ensuring protected memory and reliability.

Next, the processor in the VA 7000 series executes a program that examines the SCSI command and responds correctly. The processor, in this case, is executing out of *error-checking-and-correcting* (ECC) protected memory. Executing with ECC protection enables the processor to continue to operate reliably in the face of single bit DRAM errors. The processor is not alone in enjoying the availability characteristics of ECC memory. In fact, all accesses to all memories in the VA 7000 series are through ECC memory.

Turning back to the *write* flow, the processor notifies the host that the RAID system is ready to receive the requested *write* transfer (through the normal SCSI handshake). Then, the Fibre Channel input/output processor (IOP) in the RAID system begins to receive the data over the incoming host Fibre Channel connection. As the data arrives on the Fibre Channel, the IOP posts the data directly into the battery-backed mirrored *write cache*. The VA 7000 series is unique in how carefully it handles this important operation. As the data arrives on the Fibre Channel interface, the data travels immediately to both copies of the *write cache*.

Logically tightly-coupled mirrored memories

The mirrored memory concept used in the VA 7000 series differs from other arrays in several respects. First, in the VA 7000 series, the hardware takes responsibility to perform all mirrored operations using its high-speed dedicated board-to-board communication hardware. The hardware mirrors the memory accesses without processor intervention.

Second, the VA 7000 series hardware takes significant care to provide atomicity[1] of operations. In other words, the array mirrored accesses write to both memories, or to neither memory. A specialized protocol in the board-to-board communication interface provides this atomicity. When the IOP writes data, the data travels into a staging buffer. Initially, the data travels to the remote board, then the data is written to the other local controller buffer. Only when both boards agree that they have valid contents, does the data travel into the mirrored non-volatile cache memory. This activity takes place at hardware speeds, so the whole handshake takes mere nanoseconds.

End-to-end CRC protection

Returning to the *write transaction* flow, the Fibre Channel IOP writes data to the mirrored cache memory. The VA 7000 series's hardware provides end-to-end CRC protection on the sector data. In other arrays, CRC protection may start too late or end too soon to provide complete end-to-end protection. In the case of the VA 7000 series, the “Ends” of the transaction are pushed right to the boundary of the Fibre Channel IOP. Before the data arrives, the VA 7000 series's processor has already seeded the CRC values. As the data flows from the IOP to the cache memory, the hardware computes the CRC. Unlike some other products, *at no time* will *any* host data flow in the VA 7000 series without the associated CRC protection.

The write request resides in cache memory

Battery protection of the mirrored cache memory

Each VA 7000 series controller has a cache memory backup battery. Thus, a typical VA 7000 series with two controllers has two backup batteries. The two batteries share nothing. No component on either controller can disturb or hinder the functionality provided by the battery on the other controller. Additionally, the batteries are rigidly, physically attached to the controllers. Controllers can be safely removed from the array without fear that the cache contents will be lost.

[1] Atomicity: a transaction should be done or undone completely. In the event of a failure, all operations and procedures should be undone, and all data should rollback to its previous state.

Constant power protection throughout

At this point in the *write transaction*, the data has safely been stored in mirrored cache memory with its CRC. The processor will eventually move the data from the *write cache* to the disk. As the data sits in the battery-backed cache, many other hardware elements are actively protecting this valuable resource. For example, should a rare, but possible, fault occur on one of the controller boards such that the power rail shorts directly to ground, active circuit breakers on the controller boards prevent that corruption from taking down other boards. Or, should a fault occur such that the main power rail fluctuates, each controller protects itself with on-board DC-DC converters, therefore eliminating any fluctuation in power supplying the controller.

Moving the write data to disk

The VA 7000 series disks are formatted with 520-byte sectors, thus eliminating the piecing of sector data together as the data moves to the drives.

Continuing with the data transfer, at this point the data moves from the cache memory to the disk drives. All disks in the VA 7000 series are formatted with 520-byte sectors instead of 512-byte sectors. The extra eight bytes in each sector contain the CRC data for that sector. The data moves from the cache memory in 520-byte sectors, and the memory control provided in the VA 7000 series will treat all sectors as 520-byte singular units. Thus, there is no piecing of sector data together as the data moves to the drives. Again, the CRC protection travels with the sector data at all times. In addition, the memory controller in the VA 7000 series knows when it is handling sector data, and it will apply its ECC coverage in a block fashion over the entire sector. If any problem is detected with any words of the sector, the entire sector will not be allowed to leave the domain of the memory controller until the appropriate corrective action is taken by the VA 7000 series's processor. This type of protection is applied to all sector data accesses in the VA 7000 series.

The life of a read transaction

Obtaining the data from disk

Reading map information with read-compare technology

To read data from the array, the host requests data through the front end Fibre Channel IOP. The processor in the VA 7000 series determines the location of the data in the pool of storage attached to the backend of the VA 7000 series. The data is located by looking at the mapping of the host address to disk storage location. The memory control function in the VA 7000 series performs special *read-and-compare* operations for this important lookup operation. The memory control logic knows when the processor accesses the mapping information.

The memory control logic then launches an independent *read operation* on each controller board each time the processor accesses the mapping information. Each memory controller will apply the normal ECC protection to the reads, and once that has been completed, the two read results are compared. The processors will receive the map contents *only* when the two controller boards fully agree on the contents of the mapping information. All of these operations are performed with fast, low latency hardware support. Thus, the processor will not experience any performance impact associated with this additional layer of robustness.

Orthogonal validation of operating conditions

The VA 7000 series provides logic that monitors and validates the operating conditions.

With the mapping information, the processor requests data through the back-end Fibre Channel IOP to the appropriate location on the appropriate disk. The Fibre Channel IOP, like all digital devices, requires that it operate within a certain voltage range (3.6 to 3.0 volts) and temperature range (0C to 65C). When any digital device operates in an environment outside its specifications, the resulting behavior may be undefined. The VA 7000 series provides logic that monitors and validates the operating conditions provided to these devices. In the case of voltage sag, an independent circuit detects that the voltage has fallen below an acceptable level and immediately halts operations on the controller board. Similarly, each controller board contains multiple temperature sensors. The user has access to the readings on these sensors. If the temperature rises to a point where the digital logic may no longer operate correctly, the implicated controller is shutdown. Through careful monitoring of the conditions experienced by all of the devices in the VA 7000 series, and then taking appropriate protective actions as necessary, the VA 7000 series drastically reduces the opportunity for undefined behavior to compromise the integrity of the system.

The VA 7000 series has multiple paths to every drive and multiple choices on how to reach a given drive. At no time will data travel through the system without CRC protection.

In the normal case, the drive returns the 520-byte sector data without difficulty. Incidentally, the VA 7000 series's processor has a number of choices on exactly how to reach any given drive. The VA 7000 series has multiple paths to every drive in the system. Once data has been transferred, the VA 7000 series's processor is notified of the completed request. The processor then commands the host IOP to transmit the requested data back to the host system. However, before the processor can complete that command, it will store the end-to-end CRC seeds into the "seed-table" in the VA 7000 series's memory.

The 520-byte data sits in the *read cache* until the host Fibre Channel IOP reads the data from the cache. Just like the ***write case*** where the CRC computation occurs just as the data enters the VA 7000 series, the end-to-end CRC is validated precisely as the data flows to the host IOP. At no time will data internal to the VA 7000 series travel through the system without the CRC protection.

Improving fault identification and handling

Adding robustness to any system requires the provision for careful testing of the features. In the case of the VA 7000 series, special fault injection testing logic insures that the overall system responds correctly to the faults that trip detection circuits. In many cases, fault injection capability must be designed into some of the electronics in order to validate the overall system's response to failure.

For example, heating resistors are built into the board (but do not ship) so that an independent party (a test system) can cause a particular temperature sensor to read an abnormally high reading. This externally injected fault causes the processor to take action to prevent thermal runaway. Without the benefit of this rigorous testing with externally injected faults, many products do not behave correctly in the face of actual failures in production operation.

The VA 7000 series's circuitry also has a fault injection interface that can be operated through a special cable from a separate test system. Once the external test system provides the appropriate passwords to the VA 7000 series's hardware, the external test system injects a fault by requesting that adjustments be made to certain elements in the system. For example, a bit in a particular data word can be adjusted to check the validity of the system's response to ECC corrections.

The designers of the VA 7000 series recognized that robustness features are useless without the proper validation. External systems test fault response better than any internally driven tool. The external system does not care exactly where in the code the controller processor happens to be. With the VA 7000 series in a random state, the external system injects the error at a random time and in a random way. Lesser designs rely on tests with internally injected faults where the processor is at a given point in its code, injecting a given error. The fault coverage of internally generated events covers only a fraction of the fault coverage offered by testing with external fault creation.

Conclusion

By designing the VA 7000 series from the very beginning with the capability to support complete end-to-end CRC and exhaustive robust test mechanisms, HP has created an array with superior data protection and well proven error-handling capabilities. HP's extensive design and testing efforts benefit the user through extreme reliability of the array components in production use.