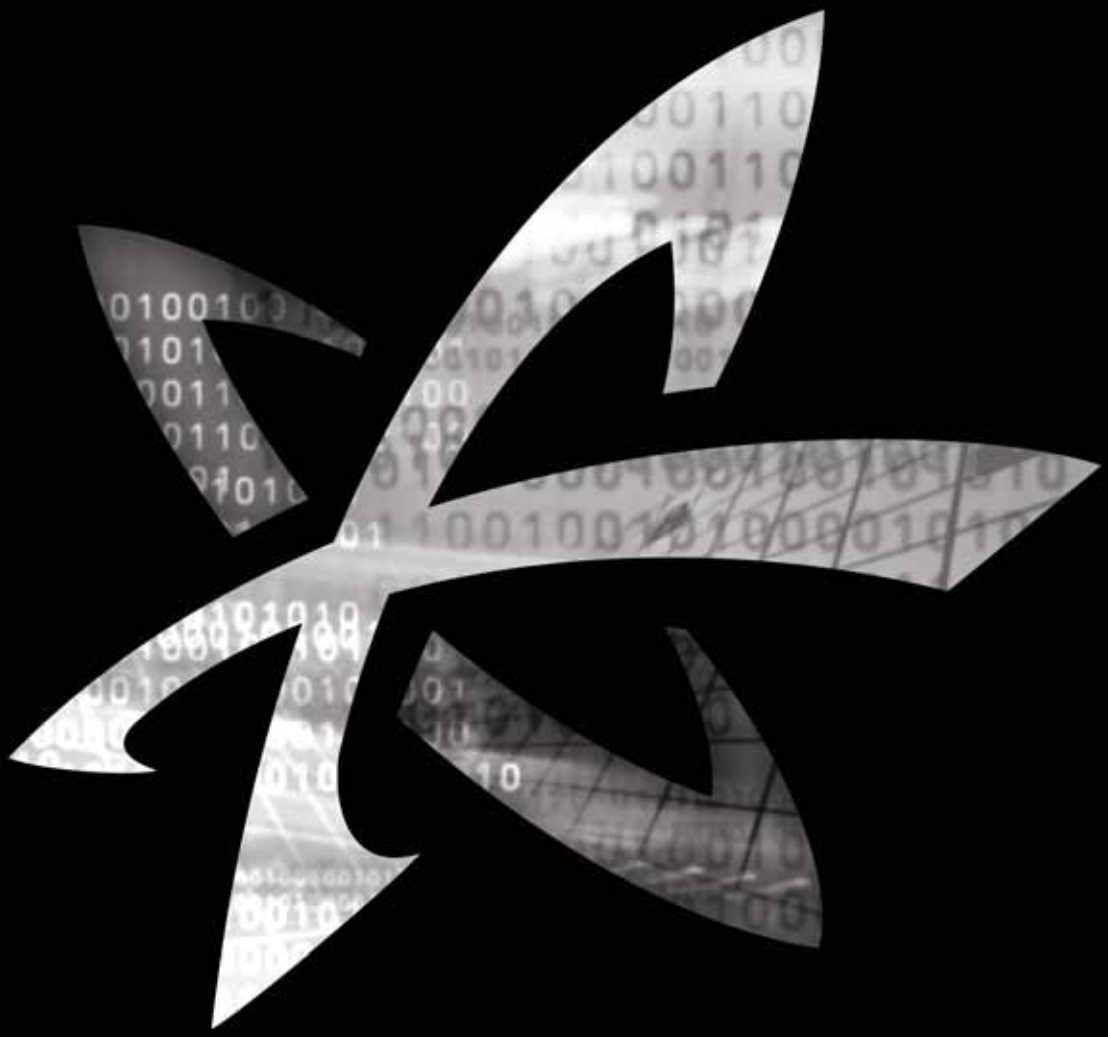

Fusion-io's Solid State Storage – A New Standard for Enterprise-Class Reliability

FUSiON-iO

Fusion-io's Solid State Storage – A New Standard for Enterprise-Class Reliability

Fusion-io offers solid state storage solutions based on NAND flash that provide a level of integrity and availability for mission-critical data that exceeds today's solid state storage solutions and significantly surpasses that of enterprise-class rotating magnetic storage devices.

With throughput and seek times many times faster than the fastest disk arrays, it is little wonder that enterprise data centers have been keen to include NAND flash as part of their server infrastructure. The primary reason NAND flash has not been widely adopted in the computer industry is its reputation for unreliability. There is a long-standing view that NAND flash storage works well for non-mission-critical applications, such as media storage devices (where the occasional bit error generally translates into a slight audio hiss or a stray errant pixel in a video), but cannot be relied upon for applications where a bit error could crash an operating system or compromise the integrity of critical data.

System architects face a number of storage-related challenges and NAND flash technology presents its own set of unique problems. But Fusion-io has developed patent-pending techniques to create NAND flash-based storage with reliability equal to or exceeding that of disk-based storage. This paper describes several inventions and advancements Fusion-io has introduced to ensure data is not corrupted or lost. Additionally, this paper discusses the probability of catastrophic storage device failure and how Fusion-io's architecture ensures predicable, controlled management of early device failure, long-term device attrition and data changes due to external and data transport interference.

NAND Flash

Flash memory chips are a non-volatile storage medium (i.e., they can retain their information even in the absence of power). The most common types of flash chips are silicon-based NOR and NAND, named after the types of logic gates used in their design. NAND flash, introduced in 1989, has become the most commonly used type of flash chip, due to its quicker write speed. Flash memory continues to grow in popularity as its price steadily declines, its storage capacity increases, and its physical size continues to decrease.

In Fusion-io's storage devices, NAND flash chips are stacked several at a time (to increase density), operated in parallel (to increase throughput) and mounted on a printed circuit board (PCB) that plugs into a PCI-Express (PCIe) slot on the server or in the CPU. The flash media is integrated with the controller onto a single PCI-Express card.

NAND flash, as a storage medium, offers a number of benefits in comparison to rotating magnetic storage devices (aka HDD, Hard Disk Drives). NAND flash has no moving parts and is therefore significantly less prone to shock or movement disturbance. It is a high speed solution in both latency and throughput. Temperature and humidity resistance mean that it can operate in a number of different environments. Finally, NAND flash consumes significantly less power than rotating magnetic storage devices, particularly when you take into account secondary power requirements for device cooling.

However, NAND flash does introduce a number of potential failure points including:

- Media – Media failures can occur on the NAND flash chips themselves.
- Transport – Transport errors can occur anywhere along the path carrying data from the CPU through to the NAND flash chips.
- Management – There is a small chance that management problems can occur within the logic of the device itself. The code that controls the operation can contain technical problems that can result in data failures.
- External – External problems can affect any part of the process.
- Device Failure – Catastrophic hardware failure can also occur. This includes the possibility of internal short circuits and open circuits within the memory array itself.

Protecting the Data

Implementing a variety of design and architectural strategies for protecting data integrity, Fusion-io's NAND flash devices greatly exceed the reliability of rotating magnetic media storage devices, while providing performance that is orders of magnitude better. Fusion-io protects your data at every step, ensuring that nothing is lost or corrupted in transit or on the media.

Data Integrity

Data integrity means having a high degree of confidence that what you put into a storage system is exactly what you get out when you request that data and it is the most important function of a storage system. While being moved from a computer's RAM or CPU to the Fusion-io device, several proven industry-standard approaches are used to ensure data integrity. The CPU, chipset, and RAM use SECDED (Single Error Correct Double Error Detect) or chipkill (method for on-the-fly replacement of a failed chip) to ensure accuracy. Once data is written to the storage medium, it is again checked for accuracy.

When data is read from the storage medium, error correction techniques are again employed to ensure that the data being retrieved is correct. The device can correct a substantial portion of the data being read. NAND's reputation for unreliability is based on studies that show potential data loss without utilizing error correction – or less correction than that employed by the Fusion-io device. Using the methods described here, Fusion-io devices can produce results that exceed target error probability by about four times. Fusion-io's devices also use a patent-pending approach when writing data, which allows the data's path to be reconstructed from information generated during the write process.

Data Availability

Data availability means having a high degree of confidence that data stored will not be lost, either while in transition to the storage device or after it has been written to the media.

Fusion-io employs a wide variety of techniques to overcome some of the common problems associated with data availability in general, and also addresses some that are particular to NAND flash as a storage medium. Generally speaking, NAND flash is substantially more reliable than rotating magnetic media. It eliminates the chance of mechanical failure

(the failure associated with moving parts). There is, however, a chance of bad chips and chip wear-out. Fusion-io mitigates this risk using a variety of approaches.

Fusion-io's redundant, patent-pending approach to writing data allows data to be rebuilt at a very high rate of speed, ensuring rapid data availability. Data is also regularly moved and checked for accuracy to ensure that it does not deteriorate on the flash chip. This also consolidates good data and reallocates space on the drive to ensure greater data availability. This system also spreads data evenly across the device, ensuring uniform wear across all chips.

Additionally, Fusion-io uses multiple error correction code (ECC) techniques to identify and correct faulty data. Using ECCs, the device controller can correct up to 11 missing or incorrect bits out of every 240 bytes. One of the biggest benefits of ECC routines is that it they allow the device to predict the likelihood of failure on individual chips. When a particular area of a chip has passed a set unreliability threshold, its data can be moved and that are will be taken out of service. The controller continues to identify and remove bad blocks, regions of chips or even entire chips so that ordinary wear-out does not cause catastrophic failure rather a very predictable wear-out.

Device Longevity

The majority of this paper has concentrated on NAND flash in an enterprise-class storage device, and how to leverage its strengths while overcoming its weaknesses. NAND flash, however, is only part of a Fusion-io's storage device. The flash chips reside on a PCIe adapter card that has a number of other parts as well, all of which are susceptible to failure. The life of a NAND flash storage device can be estimated by examining the failure rate of its component parts. Wear-out is generally a function of having lost enough storage cells that both capacity and reliability drop below acceptable thresholds. This can be assessed by evaluating and keeping a record of the amount of errors detected at each physical location.

NAND flash wears out at a predictable rate as described by the formulas below. Effective use of wear-leveling strategies employed by Fusion-io can significantly improve the life expectancy of its drives. Please note that the formulas are applied to both MLC and SLC NAND-based non-volatile memory technologies. Single-Level Cell (SLC) NAND and Multi-Level Cell (MLC) NAND offer capabilities that serve two very different types of applications – respectively, those requiring high performance at an attractive cost-per-bit and those seeking even higher performance over time, that are less cost-sensitive:

Average-lifetime = lifetime / read-write- ratio

TYPE / WRITE DUTY	AVERAGE ESTIMATED LIFETIME FORMULA
SLC flash @ 40% write duty	25 calendar years
MLC flash @ 20% write duty	10 calendar years
MLC flash @ 40% write duty	5 calendar years

Average estimated lifetime based on Fusion-io lab testing

The read/write ratio is difficult to predict, and will vary considerably from environment to environment. As a point of reference, the International Disk-drive Equipment and Materials Association (IDEMA), an industry trade group that publishes storage device standards, recommends a read/write ratio of 60%/40% for its server-class device reliability testing (IDEMA Standards, Document R3-98).

Flashback Protection

Enterprises have long sought to take advantage of the speed, size, low-power and high-performance of NAND Flash because of its potential to change the way they manage large amounts of active data. The primary objection to NAND flash has been the reliability of the medium. Fusion-io has eliminated this barrier by inventing a revolutionary self-healing technology, known as Flashback Protection, in our controllers that instantaneously restores, corrects and resurrects lost data in the flash-based storage sub-system. Flashback Protection is accomplished by collectively using advanced bit error correction, proactive data integrity monitoring of stored data and the recent addition of a dedicated chip to repair failed devices.

Fusion-io is the first and only company to bring RAID-class redundancy and reliability using Flashback Protection down to the card level. The Flashback Protection system allows users to diagnose and correct system errors. Fusion-io integrates dedicated NAND flash chips, which offer information that enables the detection of single bit errors. This technique eliminates data loss due to chip failures and extends the usable lifetime of the NAND flash-based storage device. The NAND flash chips on Fusion-io's products contain an innovative storage architecture that enable it to deliver the performance, and now the reliability, of a storage area network (SAN) at a fraction of the power, size and cost of traditional disk arrays.

Controlled Predictable Usage Versus Catastrophic Failure

Among the greatest reliability benefits of the Fusion-io storage device is its ability to:

- Restore and Protect data
- Monitor and predict media wear-out
- Correct bad data as necessary
- Take blocks out of service when their failure rate becomes unacceptable
- Replace bad chips on-the-fly
- Move the data to a known good location (and update corresponding mapping information)

Data stored on the Fusion-io medium is double protected using both ECCs and parity data on the redundant chip. The net effect is that wear-out of the device, instead of being catastrophic, is predictable and incremental.

A Fusion-io device provides advanced warning prior to wear-out. Fusion-io supports today's monitoring management functions to measure and report on the device's status and usable life. In almost all cases, device upgrade is a smooth and predictable process, rather than an emergency situation.

Fusion-io protects your data at every stage of its path from your applications to the NAND flash storage medium, ensuring that nothing is lost or corrupted along the way or while the data is being stored. Data is checked multiple times, using several error detection methods. Once it reaches the storage medium, it is stored with robust error correction encoding that lets the flash device not only identify but correct bit errors. Fusion-io's data integrity design target is a 1 in 10^{30} probability of undetected bad data and a 1 in 10^{20} probability of uncorrectable data, as compared to a 1 in 10^{16} probability of undetected or uncorrectable errors for rotating magnetic storage devices.

Conclusion

Now with Fusion-io's comprehensive approach to data integrity, it is safe to exploit the exponential performance gains and many other benefits offered by NAND flash storage. The storage architecture pioneered by Fusion-io ensures predictable, controlled mitigation of early device failure, long-term device attrition and data changes due to external and data transport interference—issues that have up to now limited the adoption of NAND flash-based storage at the enterprise level. Fusion-io's NAND flash devices exceed the reliability of rotating magnetic media storage devices while providing an order of magnitude performance improvement.

Flash Back Block Diagram