IBM XIV® Storage System:

Reliability Reinvented

White Paper

August 2008

This document could include technical inaccuracies or typographical errors. IBM may not offer the products, services or features discussed in this document in other countries, and the product information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. The information contained in this document is current as of the initial date of publication only and is subject to change without notice. All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided "AS IS" and no warranties or guarantees are expressed or implied by IBM. Information concerning non-IBM products was obtained from the suppliers of their products, their published announcements or other publicly available sources. Questions on the capabilities of the non-IBM products should be addressed with the suppliers. IBM does not warrant that the information offered herein will meet your requirements or those of your distributors or customers. IBM provides this information "AS IS" without warranty. IBM disclaims all warranties, express or implied, including the implied warranties of noninfringement, merchantability and fitness for a particular purpose or noninfringement. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

info@xivstorage.com

www.xivstorage.com

Contents

Introduction

Reliability is a critical feature of any enterprise storage system. Today's enterprise-class reliability requirements translate to total avoidance of data loss. Data loss can have a crippling effect on any organization, causing lengthy and complicated restore processes, performance degradation, and application downtime, ultimately having a negative impact on financial performance. Avoiding data loss is typically achieved by means of various types of redundancy mechanisms that provide tolerance to the failure of individual components in the system. Such mechanisms are transparent to the user, allowing the system to continue to function non-disruptively and without affecting performance levels.

Another aspect of enterprise-class storage is availability — the quality of providing non-stop service to enable applications to be continuously operational 24 hours a day, 365 days a year. In virtually every industry and across all sizes of companies, application downtime translates to lost revenues. In other scenarios that do not require a 24/365 operation from a business standpoint, there may still be a requirement for 24/365 storage, since the effort and cost of restarting an application can be significant.

While many vendors promote their products as highly reliable and available, most do not deliver to true enterprise-class standards. For instance, even high-end storage systems are much more prone to double failures (and thus to substantial data loss) than is commonly perceived. Additionally, while many leading enterprise-class storage products allow for continuous availability of data, they do so at a significant performance penalty, sometimes so severe that it is equivalent to downtime.

The IBM XIV® Storage System delivers a new standard of reliability and availability for enterprise-class storage systems, based on the following strategies:

- Minimizing the probability of a single failure
- Providing full immunity from both data loss and downtime against a single failure
- Minimizing the probability of a double failure
- Providing real-life high-availability, translating to “always on” with high performance

This white paper explores the basic concepts behind these strategies and how they are implemented and supported by the XIV architecture.

Minimizing the Probability of Single Failure

The traditional approach to analyzing reliability is based on the false assumption that the probability of single failure is the same across all storage architectures. This analysis is predicated on the fact that all storage vendors use the same core architectural components, mainly disk drives, and that these are manufactured by the same vendors. In reality, failure rates vary depending on the type of system.

Storage vendors can implement several architectural enhancements to minimize the probability of single failure. The XIV system supports all these enhancements, including:

- **Optimal Load Balancing to Reduce Disk Drive Failures.** Optimal load balancing across all disk drives not only has performance and management advantages, but also improves reliability. Due to the mechanical nature of disk drives, these devices fail more frequently as usage increases. Also, unlike electrical components, the mechanical component failure rate is super-linear to stress levels. For example, a system of 200 disk drives performing 10,000 transactions per second on continuous 24/365 operation would have significantly more disk drive failures if 20 percent of the disk drives account for 80 percent of the transactions (as is the case with traditional storage architectures), as compared to the number of disk drive failures it would have if the load were optimally balanced (as is the case with the XIV system).
- **Aggressive Phasing-Out of Disk Drives Prone to Failure.** This capability minimizes exposure to both single and double failures. Modern disk drive technology provides advanced notice of the possibility of failure. Disk drives can detect early signs of failure, and the storage system can initiate the rebuild process before a failure actually happens. The XIV system starts the rebuild process when a problem is detected, mitigating data loss exposure for both single and double failures since the rebuild process may be completed prior to the disk drive actually failing.
- **Scrubbing.** As a continuous, background process, the XIV system tests every disk drive block in the system to detect for media errors. This approach ensures early detection of media errors and minimizes the risk of data loss from a combination of a failed disk drive and media errors on the backup copy.
- **UPS Systems Extend Component Life.** Storage systems use UPS units to maintain cached data during power failures. A side benefit of these UPS systems, which are of double-conversion type (AC-to-DC-to-AC) is that stable current is ensured, with no spikes. The XIV system employs three UPS systems in a standard configuration, extending the time to eventual individual component failure.

Providing Full Immunity in Cases of Single Failure

Immunity against failure or the malfunctioning of any individual component is the basic pledge of any storage system that is positioned as enterprise-class. Upon close inspection, however, significant differences emerge regarding immunity to failures. End-users would be wise to examine vendor claims carefully.

The XIV system supports a unique approach to reliability that translates to truly **full** immunity against individual component failure. This approach consists of a combination of features, including:

- **Data Redundancy.** Each data portion is stored twice in the system, with the two copies residing in different disk drives and different modules.

- **Power Supply Redundancy.** Internally, each module has redundant power supplies and fans. Externally, the system is fed by two power sources and is supported by three UPS units.
- **Internal Connectivity Redundancy.** Based on multiple Ethernet switches, this feature is important in terms of reliability, since it provides an alternative to the common backplane, which is a single point of failure and typical of leading architectures in the industry. Indeed, although backplane failures are rare, they do happen as a result of either external (such as water- and heat-related) or internal problems.
- **Full Replaceability of Components.** At any point in time each of these redundant components can be replaced through an online procedure that is transparent to the overall functioning of the system. Unlike the XIV system, other storage architectures may require system downtime to replace faulty components (especially backplane replacement for backplane-based architectures).

Minimizing the Probability of Double Failure

Before examining the XIV system's data protection architecture, it is helpful to understand the data protection schemes of today's storage systems and their limitations.

Current Data Protection Schemes and Their Limitations

The data protection schemes available today are based on several concepts:

- Disks are grouped together for redundancy purposes (either mirroring of disk pairs or another scheme, like RAID-5/6, in larger groups)
- A hot spare disk is allocated for redundancy and used only to store data in the event of a failure
- Upon a disk failure the data protection scheme initiates a rebuild process, re-generating the lost data onto the hot spare drive

In this data protection scenario, exposure to a double failure does not depend on the time required for replacing the failed component but, rather, the time of the rebuild process.

An analysis of storage system reliability takes into account the disk Mean Time Between Failure (MTBF) and calculates the disk failure rate and probability of a second failure during the rebuild process. Done this way, the analysis yields a double failure rate that is practically infinite. However, a more accurate analysis approach reveals a higher probability of double failure.

Long, High-Stress Rebuild Process: Potential for Double Failure

While disk capacity has increased dramatically in recent years, disk bandwidth has increased only modestly. As a consequence, the time required to write the contents of a disk has increased dramatically. During a disk rebuild process, the hot spare disk is completely re-written, resulting in a long rebuild process that can take several hours and, in some cases, even more than a day.

But the impact of the rebuild process should be measured in other terms, not just time. Disks are mechanical components, with moving parts; as such, they tend to fail in proportion to actual use, not time. In a system with 300 GB disk drives, for example, the impact on the disks of reading and writing 300 GB might take only about eight hours, but is equivalent to an activity stress level of several months of storage transactions. Analyzing disk usage statistics, the disk stress in the rebuild process is equivalent to the stress level of 2-6 months of service.

Therefore, when evaluating the probability of double failure, the approach of using the disk drive MTBF and time to rebuild is not relevant, since the disk drive MTBF assumes a standard I/O load while load on the rebuild time is an order of magnitude larger.

Hidden Problems in Hot Spare Disk Drives

The standard implementation of a hot spare disk drive is a drive that sits idle until a failure occurs, at which time the whole disk drive is written to. This implementation causes an idle disk drive to undergo sudden stress: the equivalent of several months of work within several hours.

Such an implementation is vulnerable to double failures, since the previously unused hot spare disk drive may contain a hidden problem. The Self-Monitoring, Analysis and Reporting Technology (SMART) mechanism provides 24-hour advance notice of disk drives that are about to fail. For a redundant system, this early indication is enough to start a rebuild process. However, if the system is already undergoing a rebuild process, the disk drives participating in the rebuild process are at a stress level many times higher than standard. The 24-hour warning will be of no help.

Shelf-Related Double Failures

Almost all storage systems implement their redundancy schemes within the disk shelf. This means that shelf problems might cause a double disk drive failure and potentially cause data loss. Typical examples for such problems include:

- **Over-Heating Problems.** Failures in the shelf’s cooling components could harm two disk drives concurrently
- **Electrical Problems.** Failures in the electrical system could cause failures in two disk drives concurrently

XIV Architecture Redundancy

The XIV system's architectural redundancy represents a significant leap in maintaining redundancy and improving reliability. It is based on several principles:

- Data on each disk drive is split into granular stripes, with each stripe mirrored on a different disk drive
- Each stripe is mirrored on two disk drives, located different modules
- Any two disk drives in the system – but not any two drives on the same module – store some of the data
- There is no disk drive dedicated for a hot spare. Instead, some of the capacity of each disk drive is reserved as hot spare.

In traditional storage architectures, when a disk drive fails, the mirrored disk drive (or, if RAID-5 protection is used, the containing disk group) contains data that is unprotected. This data is then copied to the hot spare disk drive, and the system resumes redundancy.

In the XIV system, when a disk drive fails, each of the other disk drives – except the disk drives on the same module – contains a small part of data that is unprotected. The system then copies each data stripe to another disk drive and returns to full redundancy.

The XIV process is quite different from the traditional process, in several aspects:

- All the disk drives in the system participate in the rebuild process, each rebuilding only a small piece. This distributed process leads to a 30-minute rebuild time for 1 TB drives.
- Only data allocated to volumes, and within that, only data that was actually written, is being rebuilt.
- After the rebuild process, each disk drive contains more information, as opposed to the traditional approach of a hot spare disk drive put into use after a failure

How the XIV Architecture Reduces the Probability of Double Failure

The XIV system's redundant architecture reduces the probability of a double failure by several orders of magnitude, as explained below.

Reducing Rebuild Time by Rebuilding Only Real Data

With the XIV system rebuilds only that data which has been allocated to volumes and actually written. In practice, this means that if only half of the system's capacity is allocated to volumes and, on average, only half of the volume space has been written to, then rebuild time will be one fourth the rebuild time for the full system. In terms of numbers, if a fully utilized and fully written system requires a rebuild time of 30 minutes for 1 TB disk drives, then for a half-utilized, half-written system the rebuild will take one fourth this time, or less than four minutes.

This capability to dramatically reduce rebuild time is enabled by the unique distributed grid XIV architecture, in which data modules are intelligent units that are aware of volumes and other logical structures. It is nearly impossible to implement this capability in the standard approach of implementing a storage system via a loop of disk shelves, each implementing the redundancy internally, without any awareness of logical objects such as volumes.

Rebuilding only the real data reduces considerably the rebuild time in most common situations, resulting in a significant decrease in the probability of double disk drive failure.

Reducing the Probability of a Double Failure with a Distributed Rebuild Process

The most important aspect of the XIV system's redundancy architecture is perhaps the fact that the stress of the rebuild process is spread over all disk drives. This scheme eliminates the stress that a single disk drive undergoes in a traditional rebuild process.

As a result, double failure probability can be computed using the pure statistical analysis based on rebuild time and component MTBF, making the probability of a double failure a truly rare situation.

Additionally, the short rebuild time, together with the limited load on each disk drive, makes the SMART-based disk drive failure prediction useful even through the rebuild process.

Reducing the Probability of a Double Failure by Eliminating Hot Spare Hardware Components

The XIV system's unique approach to implementing capacity over all disk drives as the hot spare, rather than using dedicated disk drives, improves reliability and eliminates many hardware failures. This approach eliminates any possibility of the hot spare disk drive producing an error in the middle of the rebuild process.

Self-Healing Even after Module Failure

Although current storage architectures support self healing for disk drive failures, they do not provide the same functionality for module failures. Once a module has failed, the system will be non-redundant until a technician replaces the broken component. Such an approach has reliability problems:

- During the time between the hardware failure and component replacement, the system is exposed to a potential second failure that could cause data loss or unavailability
- Technician error or careless handling of the system while it is non-redundant could cause data loss or unavailability. It is not uncommon for a technician to accidentally remove the functioning component instead of the failed one.

The XIV system avoids these problems by expanding the self-healing and rebuild processes of module failure. The system has enough spare capacity for a failed module and three failed disk drives. Upon a module failure, a rebuild process is initiated, and the system returns to redundancy.

Only after the XIV system is redundant will a technician replace the failed component.

The XIV approach provides several advantages:

- Limits exposure to double module failure to the time it takes to rebuild the data and does not depend on maintenance team responsiveness
- Avoids human error as the cause of system downtime or data loss

Delivering Real High Availability: Continuous and with High Performance

The XIV system dramatically improves storage system availability in several ways:

- Survives single failure without affecting host I/O: every disk, module, switch, and UPS unit is redundant and protected through an active-active N+1 redundancy scheme.
- Self heals after disk drive and module failures, resulting in immunity to double failures if they do not occur within the short rebuild window
- Provides a dual power source, enabling continuous operation during power system maintenance
- Provides power redundancy via UPS units, enabling continuous service through short power outages with enough power for two de-staging cycles. The system can return to full operation after power is resumed, without wait-time for a battery recharge.

Performance Reduction Due to Rebuild Process

In many enterprise-class storage systems, performance levels can degrade significantly upon hardware failure. This is unacceptable in today's world, given that a reduction in performance levels often translates into application downtime. While traditional storage architectures may experience severe performance degradation due to hardware problems, the XIV system solves this problem.

As explained above, traditional storage architectures typically have lengthy rebuild times, during which the disk drives participating in the rebuild are under huge stress. During the rebuild process, due to the heavy I/O requirement, these systems often suffer severe performance degradation. Some systems have options for limiting the resources allocated for a rebuild, trading off better system performance for increased rebuild time. This strategy, however, increases the exposure to double failure.

The XIV system's protection scheme involves a distributed rebuild mechanism in which all disk drives participate. The rebuild process involves minimal overhead, since all disk drives are involved and each disk drive needs to rebuild only a small portion of the data. This strategy ensures that performance levels at rebuild time remain intact.

Write-Through Mode Due to Hardware Failures

Modern redundant storage architectures require that each write command is performed in two cache units before the host is acknowledged. Otherwise, a single failure in the cache module would create a data loss. Furthermore, they require redundant protection of power supply to those cache units.

Unfortunately, many storage architectures cannot guarantee protected cache after certain types of failures. A typical example is a failure in one of the cache modules, leaving its peer cache module exposed to single failure. Another example is a failure of a UPS unit, making the system vulnerable to power failures.

The solution to this problem is to use write-through mode, in which a host is acknowledged only after the information has been written to two disk drives and the write-cache is not used. This mode has a severe impact on performance and usually means downtime for application users. A technician's visit may even be necessary.

With the XIV system, write-through mode is never used. Even after failure of a UPS unit failure or module, a write request is written to a cache in two different modules.

Elimination of Data Calculation during Rebuild

Another problem with a RAID-5 or RAID-6-based rebuild is that until the rebuild process is complete, each request to read data from the failed disk must be served via multiple reads from all the disk groups and computing an XOR. This creates a huge performance hit on serving read requests. The XIV system's mirrored protection ensures that even while a rebuild is in progress, read requests are served without any overhead.

Summary

The IBM XIV Storage System introduces a world-class standard for reliability and availability that outshines that of the storage architectures that dominate the market today. The table below summarizes the key differences.

Reliability and Availability: IBM XIV Storage System vs. Traditional Systems

Aspect	Traditional Systems	XIV System	The XIV Advantage
Disk Load Balancing	High imbalance between disks	Perfect balance between disks	Fewer disk failures
Rebuild Time	6 to 25 hours	30 minutes (1 TB)	Reduced double failure probability
Rebuild Overhead	Disks subjected to an extremely high level of stress during rebuild	Rebuild stress spread evenly across all disks	Reduced double failure probability and improved performance during rebuild
Protection Scheme	Within the same shelf	Mirroring on different modules	Reduced double failure probability
Cache Mirroring	Vulnerable to failures and causing write-through mode	Always in effect, even through failures	Consistent performance through failures
Module Behavior Upon Failure	Exposure to double failure and reduced performance	Rebuild after module failure	Reduced double failure probability and maintains high performance through failures