xiv
Storage Reinvented

IBM XIV® Storage System

Performance Reinvented

White Paper

September 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

One of the major requirements of any SAN administration team is to provide users and applications with adequate performance levels. This task becomes increasingly difficult with demands for high performance growing while budgets for storage systems, administration efforts, and power consumption are diminishing.

This document describes how the IBM® XIV™ Storage System provides an outstanding and in many ways unprecedented solution to today's performance requirements. It does so by achieving the following:

- Providing high performance through a massively parallelized architecture, optimal exploitation of all system components (including disks, CPUs, and switches), and an innovative cache design.
- Ensuring that performance levels are kept intact when adding storage capacity, adding volumes, deleting volumes, or resizing volumes.
- Guaranteeing the same performance level, even throughout variations of the applications' access patterns
- Providing high performance without any planning or administration efforts
- Providing consistent performance levels even through hardware failures
- Maintaining high performance even while using snapshots

The XIV System: Architecture and Performance

Optimal Exploitation of All System Resources

Each logical volume in the XIV system is divided into multiple stripes of one megabyte. These stripes are spread over all the disks in the system, using a sophisticated pseudo-random distribution mechanism.

This revolutionary approach ensures that:

- All disks and modules are utilized equally, regardless of access patterns. Despite the fact that applications may access certain volumes more frequently than other volumes or access certain parts of a volume more frequently than other parts, the load on the disks and modules remains balanced perfectly.
- Pseudo-random distribution ensures consistent load balancing even after adding, deleting, or resizing volumes, as well as after adding or removing hardware

Integrating Cache and Disk in Each Module

Unlike traditional storage systems, the XIV system’s design embeds the read/write cache in the same hardware module as the disks. This unique design aspect has several advantages:

- **Distributed Cache.** The cache is implemented as a distributed cache, so that all cache units can concurrently serve host I/Os and perform cache-to-disk I/O. This ensures that cache never becomes a bottleneck. In contrast, traditional storage systems use a central memory architecture, which has significant overhead due to memory locking.
- **High Cache-to-Disk Bandwidth.** Aggressive prefetching is enabled by the fact that cache-to-disk bandwidth is the internal bandwidth of a module, providing dozens of gigabytes per second for the whole system.
- **Powerful Cache Management.** Its unique cache design enables the XIV system to read a large cache slot per each disk read, while managing least-recently-used statistics in small cache slots. This unique combination is made possible by the system’s huge processing power and high cache-to-disk bandwidth.

Huge CPU Power

Each data module has its own quad-core processor, giving the XIV system dozens of CPU cores. The system uses this vast processing power to execute advanced caching algorithms that support small cache slots, enable powerful snapshot performance, and so on. The massive CPU power ensures high performance through high cache-hit rates and minimal snapshot overhead.

High Performance without Management Effort

Unlike other storage systems, the XIV system is fully virtualized. The user has no control over the allocation of volumes to physical drives. As a result, the XIV system's high performance is gained with no planning efforts. The user does not have to allocate volumes to specific disk drives or shelves, nor is there a need to reconsider these decisions when new volumes are required, new hardware is added, or application access patterns change.

Instead, the XIV system **always** ensures optimal utilization of all resources in a way that is transparent to the hosts and storage administration team.

High Performance with Snapshots

Many storage systems can provide the required performance levels as long as snapshots are not defined. This is because snapshot functionality was added to these systems long after their initial design. As soon as snapshots are defined, performance levels in many cases degrade to unacceptable levels. Some systems solve this problem by using full copies instead of differential snapshots.

The XIV system has been designed from inception to support snapshots. Its combination of innovative replication algorithms and massive CPU and cache power keep the impact of snapshots on performance to a minimum. Specifically, it achieves this as follows:

- The traditional copy-on-write technique is replaced by the more efficient redirect-on-write technique, eliminating unnecessary copies
- Redirect-on-write is always performed within the same module where data is being copied between disks. This architecture provides a huge performance boost compared with the traditional method of copying between modules.
- Snapshot write overhead does not depend on the number of snapshots or volume size
- Zero read overhead for volumes and snapshots
- Zero overhead when writing in unformatted areas

Disk Mirroring vs. Parity-based Protection

Today's storage administrators face the dilemma of deciding which protection scheme to choose for their data: mirroring or parity-based. The XIV system uses mirroring protection, in which each piece of data is written on two disks. When comparing the XIV system to other systems, keep in mind that the propose configurations of other systems often involve RAID-5 or even RAID-6 protections, which create several performance problems:

- Each host write translates into two disk writes and two disk reads (or even three writes and three reads in RAID-6) compared to two disk writes in mirroring.
- RAID-5/6-based rebuild time is much longer, hence extending the time of reduced performance due to disk rebuild whenever a disk fails.
- With RAID-5/6, upon a rebuild, each read request to the failed area is served through multiple reads and computing an XOR, creating a huge performance overhead.

The XIV system architecture is shown in the following diagram:

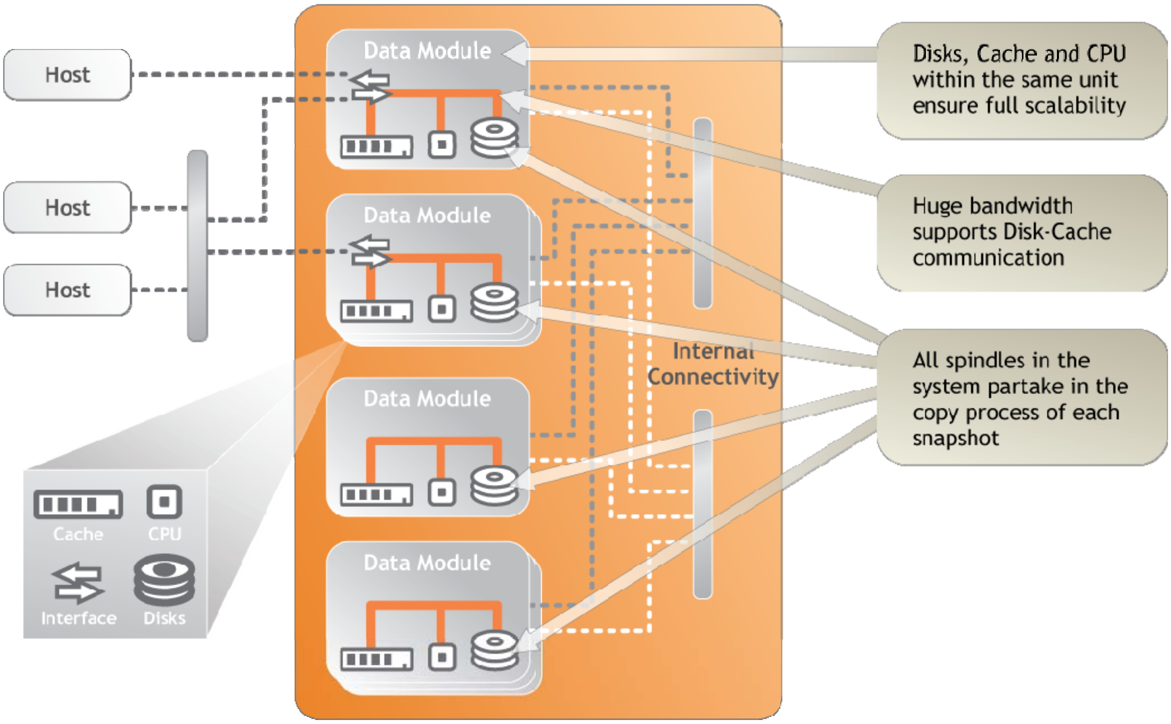
Host
Host
Host
Data Module
Data Module
Data Module
Data Module
Cache
CPU
Interface
Disks
Internal Connectivity
Disks, Cache and CPU within the same unit ensure full scalability
Huge bandwidth supports Disk-Cache communication
All spindles in the system partake in the copy process of each snapshot

Figure 1: XIV Architecture

Maintaining Performance Consistency through Failures

In many storage systems, even those considered tier-1, performance levels can degrade significantly upon a hardware failure. This is unacceptable in today's world, since a reduction in performance levels means, in many cases, downtime for the applications.

This section shows how traditional architectures create performance degradation due to hardware problems and how the XIV system solves this problem.

Traditional storage: Degradation during the Rebuild Process

The current, traditional storage implementation of redundancy involves a redundant disk group, either mirrored pairs of disks or RAID-5 disk groups. Each such group has a hot spare disk, which is used to rebuild the redundancy upon a failure.

The enormous increase in disk capacity in recent years has not, unfortunately, been matched by an increase in disk bandwidth. As a result, disk rebuild time has increased to several hours, to as many as 15, depending on disk size and protection scheme. During this time, the system suffers from severe performance degradation due to the heavy I/O requirement of the rebuild process. Some systems offer a way to limit the resources allocated for a rebuild, thus ensuring more system performance, but wind up increasing rebuild time, thereby increasing exposure to double failure.

The XIV system's disk failure protection scheme enables a distributed rebuild mechanism in which all disks participate. This ensures an extremely short rebuild time, 30 minutes for a 1 TB drive. Furthermore, the overhead of the rebuild process is minimal, since all disks participate in the rebuild and each disk only needs to rebuild a small portion. This ensures that performance levels at rebuild time remain intact.

Another problem with a RAID-5 or RAID-6-based rebuild is that until the rebuild process is over, each request to read data from the failed disk must be served via multiple reads from all the disk groups and computing XOR. This creates a huge performance impact 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.

Traditional storage: Degradation Due to Write-through Mode

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

Unfortunately, many storage architectures cannot guarantee protected cache after certain types of failures. A typical example is the failure of a cache module, which leaves the peer cache module exposed to a single failure. Another example is the failure of a UPS module, which makes the system vulnerable to power failures.

The common 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 disks and without using write-cache. This mode has a severe impact on performance and usually means a slowdown or stoppage of service to the application host. Unfortunately, it takes a technician's visit to overcome such a problem.

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

The XIV System: Performance in the Field

The performance of the XIV system has been proven in the field, demonstrating dramatic increases in comparison to other tier-1 storage systems. Several examples are given below.

Scenario #1: Write-intensive Database

A leading bank was trying to contend with a performance-demanding application based on a 7 TB Oracle database with an extremely write-intensive I/O. The application practically failed when running on a leading tier-1 storage system. When migrated to another tier-1 storage system, equipped with 240 FC 146 GB 15K ROM drives, the application managed to provide an adequate performance level, but no more. Snapshots were not possible without compromising performance to unacceptable levels; as a result, backup procedures were complex and limited.

Migrating the application to the XIV system gave the customer a dramatic increase in performance (for example, queries could now be performed in one-third of the time), while enabling the ongoing use **of 28 differential snapshots.** The gains were many: a much better response time to users, simplified physical backup procedures, and 28 levels of logical backup snapshots.

Scenario #2: E-mail Appliances

Two leading ISPs compared the XIV system against a well-known tier-1 system running POP e-mail storage for a group of e-mail appliances. The existing system required an independent interface card per each e-mail appliance, making the solution much more expensive and complex.

The XIV system was able to handle five e-mail appliances on a single interface port, with no degradation in performance.

Scenario #3: Voice-recording Application

A world leader in voice recording systems compared the XIV system with a system made up entirely of 146GB 15K RPM FC drives. The customer found that, with the XIV system, the same set of servers could support three times more clients (12,000 instead of 4,000), consequently reducing the total cost of the solution by an order of magnitude.

Scenario #4: E-mail Server

A leading telecom company tested Microsoft® Exchange server performance on various storage systems and saw a striking gap between XIV and another leading tier-1 system. After sharing this information with that vendor’s top support engineers, the customer was told that since the Exchange metadata was spanned across only 18 disk drives, performance was limited. The customer asked the vendor to lay out the volume on more disk drives. The response was that doing so was technically impossible. This example illustrates how XIV’s ease of management provided real life high performance, while other vendors did not manage to exploit the full power of the physical components due to management limitations.

Summary

As presented above, the XIV system provides:

- Unmatched performance levels, setting a new standard for SAN storage
- High performance levels **without** manual planning or a configuration process
- High performance levels that are consistently maintained, even upon hardware failure
- Industry breakthrough: snapshots **with** high performance