



OPENBENCH LABS

Data Center SAN Infrastructure

# Analysis: Building Autonomic High-Performance SANs via Intelligent Storage Elements



Xiotech®

# Analysis: Building Autonomic High-Performance SANS via Intelligent Storage Elements

Author: Jack Fegreus, Ph.D.  
Managing Director  
openBench Labs  
<http://www.openBench.com>  
January 12, 2009

*Jack Fegreus is Managing Director of openBench Labs, and consults through Ridgetop Research. He currently serves as CTO of Strategic Communications, Editorial Director of Open magazine and contributes to InfoStor and Virtualization Strategy. He has served as Editor in Chief of Data Storage, BackOffice CTO, Client/Server Today, and Digital Review. Previously Jack served as a consultant to Demax Software and was IT Director at Riley Stoker Corp. Jack holds a Ph.D. in Mathematics and worked on the application of computers to symbolic logic.*


# Table of Contents

Executive Summary	04
Disk Integration Foundation	06
Performance Scenario	09
Customer Value	13

# Executive Summary

“The extraordinary reliability of an ISE™ system translates into likelihood that a typical ISE-based storage array should incur minimal service events over five years of operation.”

The mission of IT is to get the right information to the right people in time to create value or mitigate risk. That mission makes data storage the cornerstone of any IT strategic plan. Nonetheless many IT decision makers continue to attempt to resolve the explosion in the volume of data maintained on storage area networks (SANs) on a foundation of storage arrays that put an Fibre Channel (FC) interface in front of SAS or SATA disks.



**openBench Labs Test Briefing:**  
**Xiotech® Emprise™ 7000 ISE-based Storage**

- 1) Improve Storage Reliability:** Intelligent Storage Element (ISE) technology improves cooling, reduces vibration errors, and provides autonomic self-healing to minimize storage-based service events.
- 2) Improve I/O Scalability:** Expanding storage with multiple ISE modules adds local cache and processing power via dual active-active Managed Reliability Controllers, which locally manage I/O processes to scale capacity without overburdening controllers.
- 3) Simplified Storage Management:** Independent distributed cluster architecture utilizes Web Services for maximum automation of comprehensive storage services including Intelligent Provisioning.

The strategy is to reduce capital expense (CapEx) costs while maximizing reliability, availability and scalability (RAS) by making it easier for IT to manage and replace disk drives. In practice, that strategy leads to the provisioning of storage arrays with shelves of disk drives, which are striped into RAID-based physical arrays—typically RAID 5. Software for those

arrays provides the means to partition the arrays into logical volumes and present the volumes to host servers.

When the scheme to employ commodity drives in a RAID configuration was first introduced, operating expense (OpEx) costs for storage management over the life of a storage array were just two or three times greater than CapEx acquisition costs. Today, the momentum of Moore’s and Shugart’s laws has dramatically increased that differential between OpEx and CapEx costs by an order of magnitude. As a result, top IT executives are tightly focused on the OpEx side of the ledger.

While many vendors have focused on rudimentary virtualization support to provide IT decision makers with an OpEx solution, Xiotech turned to a reassessment about the fundamental components that characterize storage resources. That reassessment resulted in a radical departure from the traditional storage array: Xiotech’s Intelligent Storage Element (ISE™)—pronounced “ice”—

technology, which is able to scale from DAS to GRID configurations. Within the Xiotech product portfolio, this scalability can be seen in their two ISE-based systems: Emprise 5000 and Emprise 7000. The Emprise 5000 and 7000 systems highlight the capability of an ISE-based storage array to scale-out or scale-up respectively.

At the heart of ISE technology is a radical new multi-drive DataPac that dramatically increases reliability, availability, and scalability (RAS), along with I/O performance, while simultaneously slashing OpEx costs through the introduction of a number of technologies. For IT's bottom line, the most powerful technological impact comes in the form of autonomic self-healing storage that is service avoidant.

In particular, ISE moves the execution of low-level device-management tasks into the storage devices and takes these tasks out of the hands of system and storage administrators. As a result, ISE reaches reliability levels impossible for standard drives enclosed in a typical storage system drive bay. Statistical analysis reveals that a DataPac is able to reach a reliability level that is far superior to that of a regular disk drive in a classic storage array.

The extraordinary reliability of an ISE system translates into the likelihood that a typical ISE-based storage array should incur minimal service events over five years of operation. To back that analysis up, Xiotech offers a revolutionary 5-year warranty for its ISE hardware. For IT, ISE provides an immediate benefit via autonomic storage device management that reduces maintenance costs and increases the stability of IT systems through automation.

*"Civilization advances  
by  
extending the  
number of  
important  
operations which we  
can perform without  
thinking about them."*

*Alfred North*

# Disk Integration Foundation

“Dubbed the Redundancy Allocation Grid System (RAGS) by Xiotech, this scheme stripes data at the level of drive heads, which correspond to disk surfaces.”

## TEAMING FOR MORE AND LESS

The evolution in disk technology has been far less turbulent than that of processor technology. Advances in storage technology have not experienced the effects of an essential discontinuity, such as processor technology underwent with respect to RISC technology. Nonetheless, Xiotech’s ISE technology, which should not be thought of in the context of a traditional storage array, has the potential to be such a disruptive force. This is especially true with respect to the DataPacs within each ISE Module

### ISE DATA PAC



At the heart of every ISE, there are one or two sealed DataPacs. For system and storage administrators, the DataPac is the base configurable storage unit. DataPacs come in three flavors: Performance, which provides 1.1TB of usable storage tuned for fast random access in database-driven applications; Balanced, which provides 2.4TB of usable storage tuned for fast sequential access of unstructured files; and Capacity, which provides 8TB of usable storage tuned for nearline archival storage. Most importantly, with the exception of basic mechanical and electronic drive components, DataPacs have nearly nothing in common with standard JBOD or RAID sets.

The ISE DataPac features physical innovations such as vibration reduction and improved cooling. Drives in a DataPac feature special firmware for improved performance, advanced heal-in-place capabilities, and improved system telemetry.

DataPacs start with matched, specialized Seagate Fibre Channel (FC) drives that have a completely different firmware from their off-the-shelf commercial brethren. The drives are then mounted in a cage that provides a fully optimized thermal and vibration solution. The drives are placed in a configuration that maximizes air flow for cooling. More importantly, within that configuration, the drives are mounted in opposition to their rotational direction in order to dampen rather than amplify rotational vibrations.

Rotational vibrations cause disk surfaces to wobble as they rotate. That wobbling increases drive seek times, because it takes longer for a drive actuator arm to settle the read/write head on the disk track. This directly degrades access

time and I/O operations per second (IOPS) performance. Worse yet, the vibrations can cause the actuator to vibrate off track during the transfer of data. This results in read retries and aborted writes, which degrade I/O throughput in both sequential and random access scenarios.

The problem gets geometrically worse when multiple drives are mounted in parallel in a standard storage array. The array acts like a spring, which amplifies vibrations. As a result, total vibrational forces—measured in Rads/sec<sup>2</sup>—exerted on drives frequently increase by an order of magnitude. In contrast, total rotational vibration forces on drives in a DataPac are typically less than 5 percent of the mean level found in a standard storage array.

#### **AUTONOMIC SELF-HEALING**

The FC drives are then connected to two active-active controllers, which are dubbed Managed Reliability Controllers (MRCs), in a point-to-point internal switched network, rather than the traditional arbitrated loop. The need for this sophisticated level of internal communications is a direct result of the tight integration of the MRC firmware and the special firmware used exclusively by all of the drives in a DataPac.

In a traditional storage subsystem, the drives, the drive enclosure, and the system controllers are all manufactured independently. As a result, controller and drive firmware must handle all of the compatibility issues that must be addressed to ensure that all of the devices will interoperate. Not only does this create significant processing overhead, it reduces the useful knowledge about the components to a lowest common denominator, which is typified by the standard SCSI control set.

ISE technology, on the other hand, exploits detailed specific knowledge about the internal structure of all of the components to fully leverage very advanced drive telemetry. Relieved of the burden of device compatibility issues, ISE moves I/O processing and cache down to its MRCs in order to implement a very sophisticated striping system. Dubbed the Redundancy Allocation Grid System (RAGS) by Xiotech, this scheme stripes data at the level of drive heads, which correspond to disk surfaces.

In effect, this head-level grid system turns the DataPac into a highly reliable superdisk. RAGS greatly increases the degrees of freedom to enable precise access to data and boost drive-level performance on the order of 25 percent. What's more, it significantly reduces data exposure on a drive, as well as the time needed to rebuild a large logical drive. With RAGS, only the surfaces of affected heads with allocated space, as opposed to entire disk drives, are rebuilt in very fast parallel processes.

What's more, with precise knowledge about the underlying components, ISE is able to reduce the rate at which DataPac components fail, repair many

component failures in-situ, and minimize the impact of any failures that cannot be repaired. This remedial reconditioning extends to such capabilities as remanufacturing disks through head sparing and depopulation, reformatting low-level track data, and even rewriting servo and data tracks. As a result of this heal-in-place technology Xiotech is able to provide a five-year warranty on its ISE-based Emprise storage system.

#### **SOLVING FOR UNRECOVERABLE BIT ERRORS**

While the details of ISE reliability with regards to the prevention and repair of errors may seem a bit esoteric to many, ISE technology resolves an important storage problem that most vendors are addressing in an oblique manner: the Unrecoverable Bit Error Rate (UBER) of a disk drive. Unlike the recoverable retries and aborts that occur with vibration problems, unrecoverable bit errors are just that: unrecoverable without some sort of reformatting or remanufacturing intervention.

Already, this is a very serious problem for large SATA drives, which typically sport an UBER of  $10^{14}$ —more expensive SCSI and FC drives typically have an UBER of  $10^{15}$ . That  $10^{14}$  bits equates to 12.5TB of data. So for a typical 500GB SATA drive, the odds of incurring an unrecoverable disk error while scanning the entire drive are 1 in 25, or just 4 percent. While UBER odds may appear pretty safe, real problems occur when multiple 500GB drives are grouped in a RAID configuration to create a very large logical disk.

For our large logical disk pool, the probability of a bit error is now the probability that there will be just one occurrence of a bit error over a series of mutually exclusive disk scans. That means a five-drive RAID configuration will have a 1 in 5, or 20 percent, chance of encountering an unrecoverable error. Those odds look a bit more dodgy and explain the upsurge in array support for RAID-6 which incurs a big performance penalty on writes.

# Performance Scenario

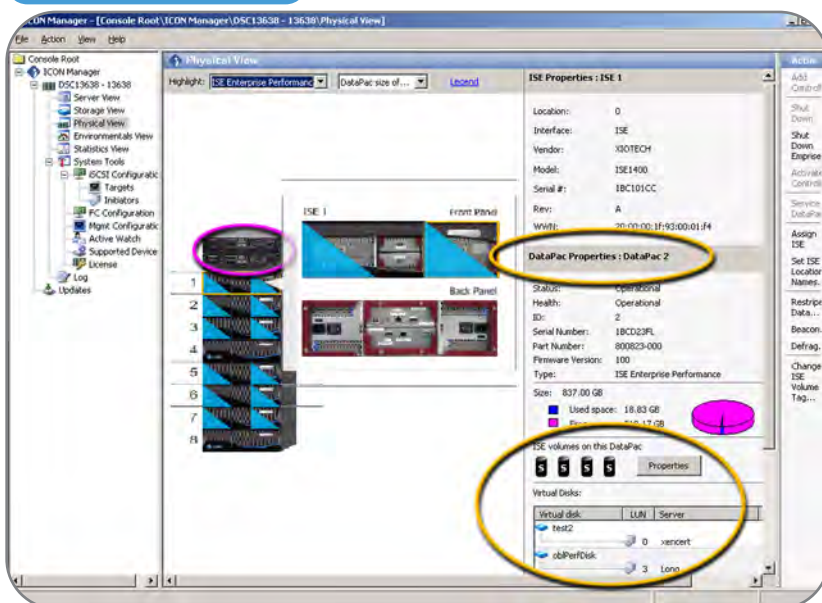
“With the DCN layering software-based RAID-5 on top of the ISE RAGS scheme, we expected to encounter some level of I/O penalty on writes; but we encountered quite the contrary.”

## SCALE-UP TEST SCENARIO

To test the performance of an ISE-based system, we used an Emprise 7000 system in a 4Gbit per second SAN environment. Emprise 7000 is designed to scale up as a single logical storage system from one terabyte to one petabyte via the addition of ISE modules. To do this, Emprise 7000 creates RAID-based volumes using the virtual volumes associated with ISE DataPacs and large 512KB stripe segments. What makes this scale-up scheme work is the fact that by adding ISE modules for storage capacity, cache and data processing power are also being added.

To coordinate scale-up as a single logical system, Emprise 7000 uses one or two Dimensional Controller Nodes (DCNs). The notion of clustered DCNs was first introduced in the previous generation of Xiotech storage devices: the Magnitude 3D® family, which created a single virtual pool of storage for dynamic provisioning utilizing standard storage drives. To extend all of the Magnitude 3D system’s advanced software, including the ICON Manager interface, to Emprise 7000, Xiotech introduced a clever logical construct that assigns four virtual storage volumes to each ISE DataPac—volumes 1, 3, 5, and 7 are assigned to DataPac 1 and volumes 0, 2, 4, and 6 to DataPac 2 of each ISE module.

### EMPRISE 7000 TEST BED

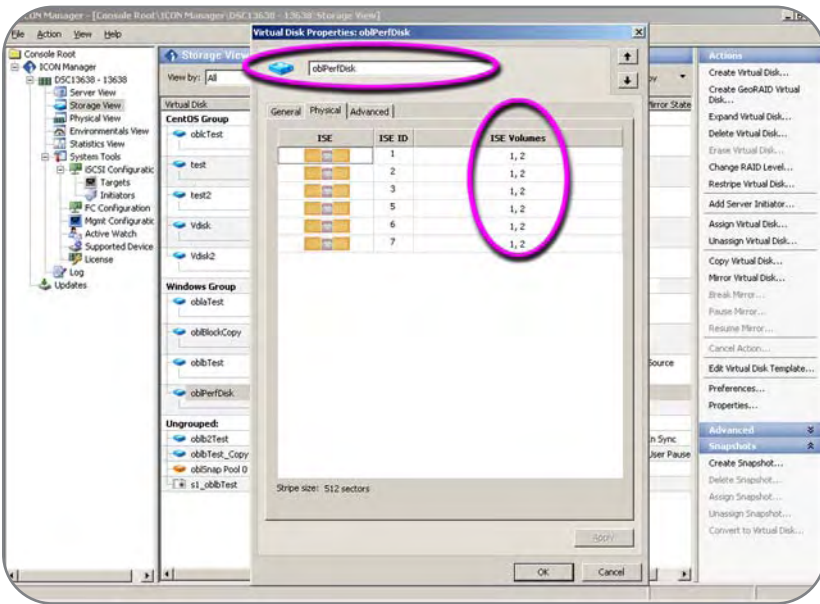


Our test Emprise 7000 storage system was provisioned with two, clustered DCNs and eight ISE modules containing two DataPacs: six modules were configured with 12 Performance DataPacs, one with two Balanced DataPacs, and one with two Capacity DataPacs. We then used ICON Manager to create logical disk volumes for our host server. For greater data redundancy and better load balancing,

Using the Physical View within the ICON Manager interface, we were able to view the Emprise 7000 test configuration with two DCNs and eight ISE modules. Highlighting the Performance Tier DataPacs, we were able to drill down and see which logical volumes that were using any of the DataPacs. In particular, our test volume, dubbed *ob1PerfDisk*, was using DataPac 2 of ISE 1.

we set a policy within ICON Manager to configure new logical volumes for host servers as RAID-5 volumes. Our policy utilized the DCN's construct of a logical volume—four logical volumes are automatically assigned to each DataPac—as the foundation for RAID sets. In particular, our policy called for using a logical volume in each of six Performance DataPacs.

**OBLPERFDISK TEST VOLUME**



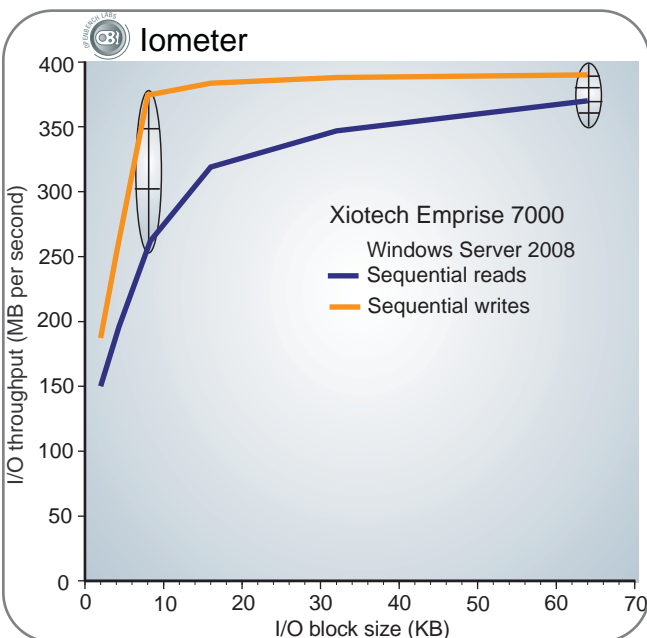
Using that policy, we provisioned our host server with a RAID-5 volume for I/O testing that we dubbed *oblPerfDisk*. With our test volume provisioned on our server, we were ready to run a series of I/O tests using the Intel® Iometer open source benchmark. In our tests, we examined sequential throughput in MB/sec and random access throughput in IOPS on an ISE-based storage system.

From ICON Manager's Physical View, we were able to list all of the DataPacs that were hosting our logical test volume *oblPerfDisk*. In particular, our logical drive utilized all of the Performance Tier DataPacs in the system (8 DataPacs within four ISE modules)..

Initially, the DCN striped *oblPerfDisk* across logical volume 1 of six Performance Tier DataPacs. Later, we expanded *oblPerfDisk*, which created a second stripe set using volume 2 of the remaining six Performance DataPacs.

**STREAMING TO WEB 2.0**

**SEQUENTIAL I/O THROUGHPUT**



We began our tests with an examination of streaming read and write I/O performance using a single logical drive. Throughput rapidly converged on wire speed—400MB/sec—for large-block I/O. Often backup applications and business intelligence applications, such as On-Line Analytical Processing (OLAP), which utilizes multidimensional database cubes, stream data in 64KB blocks. More importantly, read performance using 8KB blocks, the default for most applications running on Windows®, was over 260MB per second. That level of performance puts I/O throughput with an ISE-based Emprise 7000 system well into the highest level required for I/O intensive

To test sequential access file throughput openBench Labs used Iometer with a 20GB test file. Both read and write throughput was measured using 2, 4, 8, 16, 32, and 64KB block sizes. Both read and write throughput converged on the throughput limit of our HBA: 400MB per second. Remarkable, writes rose to that limit faster in our software-based RAID 5 environment.

applications such as nonlinear editing of high definition video.

While streaming read throughput was impressive, what really stole the show was the performance of the Emprise 7000 system when performing writes. With the DCN layering software-based RAID-5 on top of the ISE RAGS scheme, we expected to encounter some level of I/O penalty on writes; but we encountered quite the contrary. Write performance not only exceeded read performance for all block sizes, it closely converged to wire speed with 8KB blocks as throughput reached 375MB/sec.

That level of write performance is much more characteristic of a Linux OS, which attempts to bundle all small I/O requests into 128KB blocks. What's more, the Real Time Operating System (RTOS) running the ISE, as well as the DCN RTOS, were developed from a Linux-based kernel. That helps explain the very large stripe size and remarkable write throughput.

### CACHE TRANSACTIONS

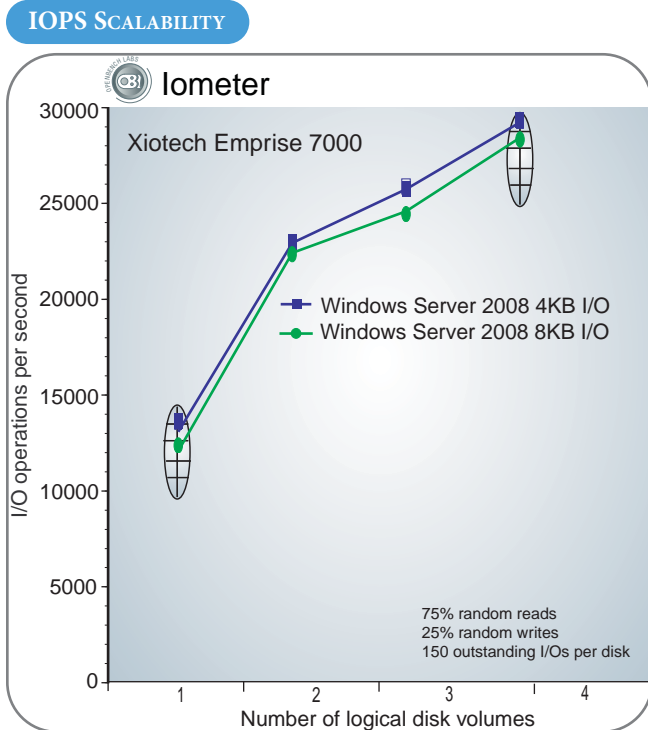
For streaming digital content applications, storage throughput and capacity go hand-in-hand as primary concerns. Applications built on Oracle or SQL Server, however, generate large numbers of I/O operations that transfer data using small block sizes from a multitude of locations dispersed randomly across a logical disk. Commercial applications that rely on transaction processing include such staples as SAP and Microsoft Exchange, which employs a JET b-tree database structure as the main mailbox repository. As a result, random access throughput, which puts a premium on the minimization of I/O latency, is an ideal benchmark to test the value proposition of ISE-based storage.

Transaction-processing applications have been traditionally among the hardest for IT to provision in an optimal fashion. Most often, these are mission-critical applications that naturally tend to be over provisioned. On top of that, transaction-processing applications have nothing approaching steady-state characteristics. These applications are very prone to having extreme spikes in processing. While not common, it is not unusual for a transaction-processing application to experience a peak load spike that increases IOPS by an order of magnitude.

That makes systems running transaction-processing applications very difficult for IT to consolidate. Worse yet for IT, those spikes also make these systems ideal targets for virtualization to improve overall resource utilization. Especially on the storage side of the resource equation, this makes understanding likely IOPS levels and the ability to scale those levels for consolidation very important. The typical load experienced by most transaction-processing systems averages around 1,000 IOPS. Heavy processing spikes can raise that level to around 10,000 IOPS.

For system and storage administrators, most FC-based arrays will meet those

requirements for one host system running a transaction processing application. The real question in today's IT environment is how well a storage system can scale to support multiple physical or virtual systems running transaction-processing applications.



On a quad-processor test server, we tested the scalability of the Emprise 7000 system using four 25GB logical volumes created on the Performance Tier of storage. In our random access tests we generated I/O streams that were populated with both read and write requests in a 75-to-25 ratio. This was done using 4KB transactions, which are used by Microsoft Exchange, and 8KB transactions, which characterize I/O access with Oracle and SQL Server.

To test random access throughput for database-driven applications, openBench Labs used Iometer on four 25GB test volumes. I/O streams involved 75% reads and 25% writes using 4KB and 8KB requests. With IOPS rates for 4KB and 8KB I/O in lock step, the performance profile of our Emprise 7000 was more in line with that of RAM disk arrays.

With one process and one logical volume, we measured just under 14,000 IOPS with an average access time of 10ms using 4KB I/O. Using 8KB I/O requests, IOPS averaged just under 13,000 with an average access time of 11ms. In both cases, Emprise 7000 easily fulfills the needs of the vast

majority of database-driven applications.

Most importantly, IOPS performance for 4KB requests is typically about 12 percent greater than 8KB data requests on most mechanical disk-based storage systems and virtually identical using solid-state RAM disks. With small I/O transfers, the difference in performance is directly attributable to data access time, which is a major component of the value proposition of ISE performance. That value proposition was clearly validated, as the differential between 4KB and 8KB requests was under 8 percent using the ISE-based Emprise 7000.

Even more remarkably, IOPS performance for 4KB and 8KB scaled in near lockstep. With four logical volumes and four I/O processes, IOPS performance reached 30,000 with an average access time of 23ms for both 8KB and 4KB requests. At this point for our tests, server CPU scalability was becoming the bottleneck and not I/O processing by our Emprise 7000 system. This behavior is quite comparable to that of arrays built on solid state or RAM disks. With no rotational component to data access, IOPS performance is throttled only by data throughput limitations.

## Customer Value

“By approaching disk drives not as a JBOD collection, but as grid of storage surfaces that requires a radical rethinking of firmware, Xiotech has improved the operating environment for storage and dramatically raised performance.”

For CIOs, the top-of-mind issue is how to reduce the cost of IT operations. Universally, the leading solutions identified by IT decision makers are resource consolidation and virtualization. These IT decision makers view both solutions as excellent ways to reduce the cost of IT operations through efficient and effective utilization of IT resources. Nonetheless, these strategies actually serve to exacerbate the impact of failed disk drives and insufficient performance.

Xiotech's ISE technology provides an innovative solution that radically alters the reliability and performance of disk-based storage systems. By approaching disk drives not as a JBOD collection, but as a grid of storage surfaces that requires a radical rethinking of firmware, Xiotech has improved the operating environment for storage and dramatically raised performance.

The Emprise 7000 storage system then builds on ISE technology, using a distributed cluster architecture to scale-up ISE modules and simplify both storage and server management with a comprehensive collection of services. Using these services, system and storage administrators can then readily configure logical drives that can cost-effectively support Web 2.0 streaming rich media, high-profile database-driven transaction processing application, as well as archival applications such as disk-to-disk nearline backup.