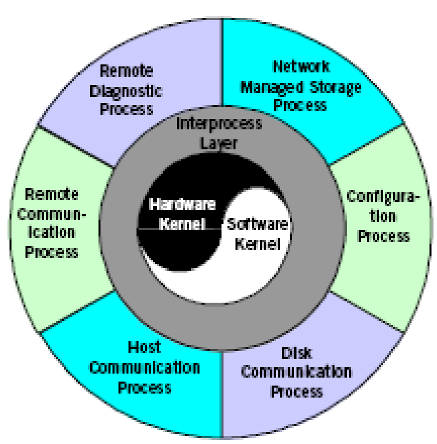
GateStor Data Systems Corporation

Performance without Compromise: Virtual Storage Architecture™

ABSTRACT:

GateStor Data Systems in now delivering enterprise-class storage services for open systems with SolidPOWER7000® *and* EPOCH® Power Cluster *family of storage servers.*
Behind every GateStor *product is the Virtual Storage Architecture™, a different kind of storage architecture, developed to address the performance and data protection tradeoffs inherent in existing RAID storage technologies. This paper presents a basic overview of the Virtual Storage Architecture, why it's different, what it does, and where it's headed. It also describes a disruptive cluster architecture that provides unequal performance characteristics and simplifies current cluster software solutions. After you read this paper it will be crystal clear the benefits of this revolutionary architecture.*



With businesses staking their futures on Information-intensive applications like Cloud Storage, data warehousing, Internet-based commerce, document processing, and teleconferencing, the demands on storage are exploding.
Storage already often accounts for 40% to 60% of computer system costs. Storage systems sales are growing at a rate that will surpass computer system sales during the next 24-36 months.
And, for applications moving to incorporate imaging, 4K and 8K video, multimedia, or simulation, storage requirements are poised to grow – not twofold or threefold – but one hundred, one thousand and even ten-thousand-fold!
For years now, companies in both the mainframe and the open systems enterprise environments have improved performance and cost-effectiveness of storage through RAID technologies that enable multiple small disks to operate together as one large disk.
But as the demands on storage grow, the limitations of RAID become self-evident.
Inherent architectural trade-offs begin to surface: performance versus data protection; read versus write performance; large versus small data transfers. All these situations limit the performance, efficiency, and cost-effectiveness of traditional RAID storage.
For many, the write penalties of RAID, wasted capacity of 1:1 mirroring, and expense/performance limitations of disk caching and buffering are becoming increasingly unacceptable.
While the falling cost of raw storage is a mitigating factor, it is far outstripped by the phenomenal growth in the volume of data – and the rising cost and complexity of *managing it* – especially in the open systems environment with its typical mix of hosts, subsystems, applications and databases.

A systems approach to storage

It was to solve these problems that the Virtual Storage Architecture was developed.
Designed from the beginning to deliver enterprise-class storage services to open systems, Storage

41 Simon St Section 2B. Nashua, NH 03060.
Phone (603) 879-0216

solutions from GateStor Data Systems are profoundly fast, efficient, *and* fault tolerant *at the same time*.

That's because behind every GateStor's solution there is an entirely new approach to managing storage: a Virtual Storage Architecture that optimizes performance, not just through its control of redundant arrays of disks, but also through intelligent, real-time management of all storage-related operations.

Unlike traditional storage which is controlled by – and dependent on – microcode located on the host, disk drive, or in the array, all GateStor systems come with powerful, embedded storage management software that's completely independent of both the associated host and drive technologies.

In the simplest terms, what enables the Virtual Storage Architecture is a powerful "storage server" – a computer powerful enough to go beyond managing disks and arrays, to managing each and every data storage/retrieval operation.

It's only logical

In essence, the power of the Virtual Storage Architecture lies in its ability to:

• Transform physical disks into a single, logical pool of continuous "virtual storage" for multiple hosts

The Virtual Storage Architecture acts as a logical "facilitator" fully separating hosts from physical storage boundaries, enabling each host to "see" just the amount of storage assigned to it, along with the proper data protection and data mirroring levels.

Because volume and host assignments are purely logical:

- The storage area can be shared by multiple hosts (multiple heterogeneous hosts, can be supported by InfiniBand based host platforms)
- Any amount of available storage may be allocated to any host
- Physical storage capacity may be added at any time without bringing down the system.
- Data protection and mirroring levels can be assigned on a data set (vs. disk) basis
- Data protection and mirroring levels can be changed at any time, without requiring any hardware reconfiguration.

At the heart of the Virtual Storage Architecture is a computer with its own optimized operating system and layered interprocess software. The computer optimizes performance through the control and coordination of asynchronous, parallel processes throughout the system.

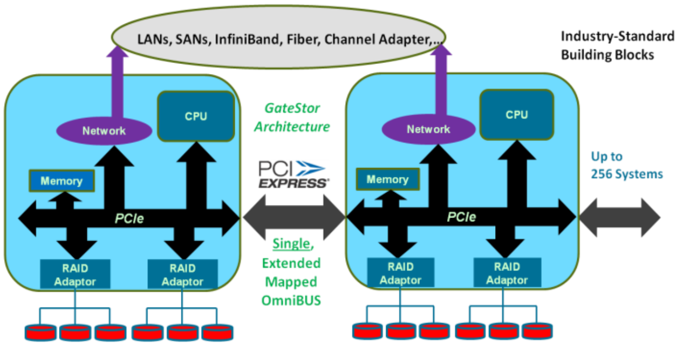
The revolutionary concept is the method on how multiple computers or nodes are interconnected to form a cluster. Traditionally clusters are done interconnecting nodes via communication networks, the overhead on network protocols affect performance severely and poses many problems that have been solved by thousands of lines of code resulting in very complex and very sizeable cluster software packages.

One huge problem is that storage clusters required data to be sectioned to the individual nodes, each node own its own storage and there are metadata directories indicating where the data is allocated on the cluster. These metadata directories often require re-direction of the client request to the node that holds the requested data, exacerbating the network usage diminishing total throughput. A huge problem occurs when data in one of the nodes is added to one node and suddenly data spill-over to the adjacent node. This is not an easy problem to solve, requiring many data movement operations and causing a huge performance degradation when this happens.

GateStor's patented cluster architecture, avoid these issues, because the cluster is not done at the network level but instead is done at the hardware bus level, avoiding the need of network protocols and providing many advantages to clusters. By using a unified bus, all the nodes in the cluster can share any resource connected to the cluster. The processor of any node can see all memories and adapter from any node connected to the unified bus, avoiding the need of segmenting data, any node can have access to any node storage even if is not its assigned storage pool. This is huge not only because solves the division of data, but even more important data can be accessed by a surviving node in case of a node failure, avoiding the need of duplication of data.

The EPOCH Power Cluster is formed by attaching SolidPower7000 nodes through the PCIe this makes a very powerful storage cluster that can grow seamlessly and performance grow linearly.

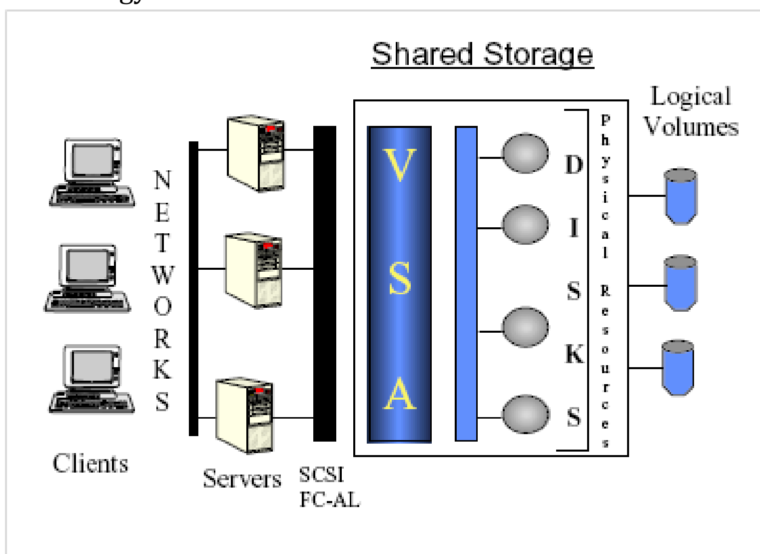
Having also the node interconnected via the PCIe, the transfer speed is 5 times faster than even the fastest network communication currently available.



• Control and optimize each storage operation at the data level

The storage server, with its own central or distributed intelligence and storage management software, manages a network of intelligent devices throughout the array to enable real-time optimization of storage operations.
This capability, combined with the ability to logically distribute data streams across any number or combination of physical disks enables optimal performance through:
• Highly efficient parity calculation
• Elimination of unnecessary read/write operations
• Highly parallel, asynchronous operations
• Concurrent support for multiple RAID levels (0, 1, 3, 5, 6 and 10) at the data set level and multi-dimensional RAID levels or RAID on top of RAID called levels 530 and 630
• Concurrent support for multiple levels of data mirroring (intra-mirroring, local, remote, cross-host (cluster), reflective mirroring, replicated mirroring) at the data set level
The same logical volume may be assigned to multiple hosts, enabling them to share data.
Logical volumes may also be moved back and forth between hosts, for example, processing on one host with subsequent post-processing on another. And logical volumes can overlap.
Because the allocation of logical volumes between hosts is configurable, the amount of storage allocated to any host, the data protection levels, and data mirroring levels can all be selected and applied at the data set level. The storage, protection and mirroring metrics can be effected, without requiring any physical changes to the array.

Today, SolidPOWER7000 and EPOCH nodes provide support for thousands of disk drives to be seamlessly shared by clusters of different hosts. And because storage is hidden from controller software and physical disk characteristics, all standard SATA and SAS disk technologies are supported and even legacy systems using older disk technology such as Fibre Channel and SCSI can be



easily integrated.

Performance Optimization

As we have seen, the Virtual Storage Architecture optimizes performance in several ways, including:
• Broadcasting data to all components in a single system and furthermore to all nodes connected to the Unified cluster.

• Parallel, asynchronous operations
• Virtual Memory Mapping
• Eliminating unnecessary read/write operations
• Efficient parity calculation avoiding the legacy read/modify/write performance penalty of RAID levels 5 and 6.

These optimizations are made possible by the continuous logical structure of the virtual storage space, including the treatment of memory and disk as a single, coherent space; and a technique called ***memory mapping***, which permits the staging of operations in memory, while maintaining a real-time map, or meta data, of the location and status of all data.

By making the memory page size to be equal to the stripping size, we avoid the Read/Modify/Write operation inherent on all RAID-5 and 6 modes. Every time data is written to the RAID a complete stripe is sent so no partial stripes are sent to the drives. Parity is generated on the fly when writing the entire stripe without the need of partial stripe parity calculations.

Data transfer optimization

Data transfers on every storage systems are perhaps the most time consuming operation only second to mechanical delays from drives, a method which reduces the number of data transfers required to accomplish a storage command is perhaps the most important attribute to improve overall performance. The larger the amount of data to be moved, the longer the transfer takes. The SolidPOWER7000 includes a method on which by doing a simultaneous transfer of the data on its internal system bus to all components of the system avoid the double or triple transfer of data inherent on any traditional storage system. Legacy systems perform data transfers from an I/O to system memory and from memory to one or more RAID controllers. The SolidPOWER7000 perform this data transfer all at once, reducing the overall transfer time in half or in third.
This technique is even more powerful when the system uses the Unified Bus Cluster architecture, data can be simultaneously transfer not only within the components of one node but to multiple nodes, this makes a unique advantages of been able to provide one or more copies of data to different nodes without any extra overhead or time to accomplish this feature (Very important in cloud storage applications).

Parallel Processing

Processing is distributed among the array computer, disk adapters, and host adapters.
At any time, there can be a mixture of data blocks which belong to different requests and in different stages of processing. The array operates on any number of blocks at a time and on whichever blocks are the most convenient to operate on at that particular time.
The customized storage engine controls the flow of data throughout the channel memory.
It assembles the data blocks in their correct order for each transfer or for each commitment to disk.
This high degree of concurrency throughout the array contributes to its high level of performance.

Using memory to eliminate unnecessary read/write operations

The system uses memory to stage data for subsequent transfer to the hosts or to disks within the array. The SolidPOWER7000 maintains statistical data about each data block, such as how many times a data block in memory was referenced since it was last fetched from disk. Such information is used for a variety of purposes. One important purpose is to determine whether a read or write operation must actually go to disk, or whether it is “transient” (i.e., cancelled out by a subsequent operation) and can be discarded.
When the data is already in memory, it is directly transferred to the host on a read request.
If the data is not in memory, it is first read from disk into memory. If this data remains in memory, a subsequent read request for the same data will also be directly serviced from memory.

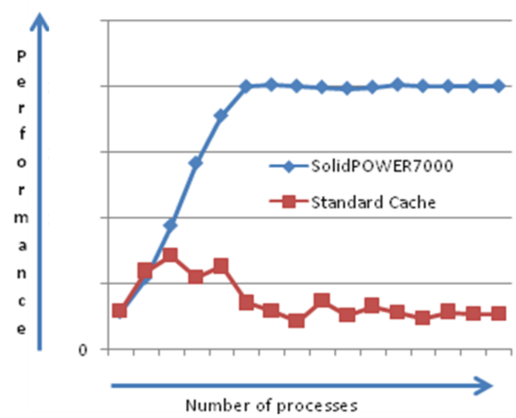
 41 Simon St Section 2B. Nashua, NH 03060. Phone (603) 879-0216

The SolidPOWER7000 may decide to read more data pages into memory than requested in anticipation of future requests; this is called *pre-fetching*.
To service a write request, data is first brought from the host into memory. If this data is to replace other data already in memory, the old data is simply discarded. New data is kept in memory as long as possible, so that there is no need to write to disk for some time and, again, so that it can be used to service subsequent read/write requests.
When new data blocks need to be brought into memory and there is not enough room, statistical data on memory usage determines which memory blocks are to be discarded.
Data is committed to disk only when necessary, such as when a memory block with modified contents has to be superseded; or when the array has idle time.
An important notion in this context is write-acknowledgment mode. The array informs the host that its data has been successfully written, at either one of two times. One, when the data is actually written to disk, called "*write-through*."
Two, when the data is secured in memory but before it is written out to disk, is called "*write-back*". Write-back mode frees up the host, allowing it to place its next request sooner.

Virtual Memory mapping vs. caching

"Caching," or holding data in memory, where access is much faster, is often used to reduce this performance penalty. Caching enables an array to dramatically improve performance – *until the cache is saturated*. Then the cache has to be flushed by writing data to disk before new data is processed.
Caching also requires that whenever the array reads a block from disk into memory, it has to write it back, and to the same location where it was read from, regardless of what operations might follow.
As a result, performance degrades sharply and becomes non-deterministic and erratic whenever cache overflow conditions are reached. Adding more expensive cache memory can allay – but more likely, only delay the problem.
In contrast to caching, memory mapping maintains a "map" of the location and status of information.
As we have seen, the Virtual Storage Architecture treats all storage – including all memory throughout the array, as well as all disks – as part of the same, continuous pool of virtual storage.
This enables the system to delay writing certain operations to disk until it sees whether it can forego them. For example, the array never writes transient parity blocks to disk. As many operations as possible are carried out in memory or are avoided altogether.
Since memory is just part of the available storage space, a modified data block in memory frees memory space occupied by the older block.
Now the memory map points to the new location.
Since memory constantly frees up evenly over time, the performance behavior of the array remains predictable at all times. When new data blocks need to be brought into memory and there is insufficient room, statistical data is used to determine which memory blocks are available.
Memory mapping uses much more efficiently memory not like caching systems that becomes a static bucket and delivers both higher and more consistent performance. The more memory you have, more predictable and better performance is achieved. GateStor systems require the use of ECC (Error Checking and Correcting) memory to ensure data loss against bit failures. It is also a requirement to have an UPS or battery backed system to guarantee power to memory. An optional non-volatile memory can be supported to protect data.

This chart compares the performance behavior of a cache array with one using memory mapping, as the number of processes grows. As cache memory
fills, performance degrades and subsequently becomes erratic, as cache must be constantly flushed to make room for new data



Selectable Data Protection

In addition to performance optimization through efficient parity and selectable parity distribution, the Virtual Storage Architecture supports the industry's only selectable multilevel data protection levels. Today the family of SolidPOWER7000 SUPERSERVERS is the only storage systems that enable users to select multiple concurrent levels of data protection (RAID 0, 1, 3, 5, 6, 10, 530, 630) for different data sets in shared storage.
Different data protection can be applied at the file or transfer level, irrespective of where data is physically located, to achieve the optimum protection and performance for different data sets. In addition, RAID protection levels can be dynamically changed, without affecting operations, and without requiring any physical operation.

Fault Tolerance

The probability of disk failure increases with additional capacity. Four basic techniques are used to reduce this probability, maintain data availability, and preserve data integrity. They are: double drive fail support RAID 6, ECC and non-volatile memory and multiple RAID controller levels 530 and 630 discussed above; and mirroring.

Volume Mirroring (Different from RAID-1)

Mirroring is simply creating a replica of each data block such that all operations are stored on two different disks. If one of the two disks fails, lost data is retrieved from the other. When the failed disk is replaced, a replica of all the data blocks on the good disk is made on the new disk. (In theory, any number of replicas of the same block can be made on the same number of different disks.)

Although mirroring statistically offers high reliability, it comes at the expense of losing half the raw capacity of the array. As we have seen, the amount of storage lost in mirroring is substantially higher than storage lost when the array operates in parity calculation mode.
Just as the Virtual Storage Architecture supports selectable, multiple levels of data protection for different data sets, it also enables multiple, concurrent levels of mirroring.
Six levels of mirroring are supported by the SolidPOWER7000 products. In all these levels, any data sets (which can be the entire array) may be mirrored. Furthermore, data recovery is always "hot".

1. Intra-mirroring. Copies of selected data sets are created. No two copies reside on the same physical disk. An array with intra-mirroring can withstand multiple (as high as the number of copies made) concurrent disk failures for all intra-mirrored data. Alternatively, the second copy of data may be accessed independently, but during the independent access, the added protection is suspended. No other storage system supports this file system level mirroring within an array.

2. Local Mirroring. Selected data sets of one array are copied onto another. This configuration survives the failure of an entire array for all locally mirrored data sets.

3. Cross Mirroring. Two hosts connected and running automatic failover software, such as a cluster, are then cross-connected to two mirrored arrays. The cross connections can be local or remote. In this configuration, a host and an array and multiple links can fail without any interruption of operations.

4. Remote Mirroring. Selected data sets of a primary site are copied to a secondary remote site. This configuration withstands failures caused by natural or man-made disasters at either site for all mirrored data.

5. Reflective Mirroring. Enables peer-to-peer, local or remote mirroring in which each site mirrors the other.

6. Replicated Mirroring. Enables multiple copies of critical data to be written remotely to multiple locations.

Any combination of mirroring levels can be applied to the array concurrently. Moreover, any of these combinations can be used with parity calculation.

Non-Volatile Memory

To ensure the integrity and availability of data that has been modified and still in transit, until it is securely written to disks, the Virtual Storage Architecture supports the use of non-volatile memory with any combination of mirroring techniques and parity calculation. If the entire array fails, data is recovered from the non-volatile memory, which in terms of memory mapping, like central or channel memory, is treated as an extension of the storage fabric.

Non-Volatile cache mirrors data held in memory, until the write operation is complete.

Using SSDs as extension of Memory

Optionally we can use a selected number of Solid State Drives to group them and present them to the system as an extension to system memory. These volumes are not part of user storage; they are used as a method to increase the size of available memory in the system. This option optimizes I/Os per second and highly recommended to improve response time in high IOPS applications.

Using Flash

The SolidPOWER7000 Flash uses all flash storage. The Virtual Storage Architecture optimizes the life of the NVMe flash and/or SSD drives by avoiding multiple and constant writes to the flash storage. As previously discussed, the VSA does not update constantly blocks of information on the permanent storage, it only writes when it has a complete page of information and avoids the numerous read/modify/writes inherent from generating parity to disk arrays.
The SolidPower7000 is architected for maximum bandwidth and can take advantage of the substantial speed difference from SSDs and NVMe Flash drives. The SolidPOWER7000 Flash uses NF1 NVMe drives these are flash memory sticks that are mounted like small disk drives and have an NVMe interface, so each module is treated as a superfast drive and can pack currently up to 512TB into a one 1U tray and provide over 2Million IOPS.

Enabling the Open Enterprise

The Virtual Storage Architecture of the SolidPOWER7000 system was designed from the beginning as an open solution in an open systems environment.
SolidPOWER7000 systems are totally independent of host and drive technologies (InfiniBand, 100GbE, Fibre Channel, OmniBUS®, SATA and SAS)
In addition, all storage resource statistics can be accessed from any SNMP management station – either from a central site or from multiple distributed management centers throughout the enterprise.
Finally, to enable customers to protect their storage investment in the face of relentless and unpredictable business and technological change, the SolidPOWER7000 was designed to support scaling across multiple dimensions, including: performance, capacity, data protection, data availability, and management.

Future Directions

The corporate business world is aggressively focusing on information as a means of providing world-class service to their customers and gaining competitive advantage. As a result of this information push, the open system enterprise has an impending need for continuous operation of its information systems and continuous access to mission critical data. This pressing need places a burden on the storage vendors to provide a robust solution that not only retains investment in current RAID technology, but that also provides primary features that equate to continuous operation such as dynamic on-line expansion of array capacity, dynamic reallocation of storage, host connectivity, OS upgrades, hardware “hot-swap”, volume re-partitioning, and data protection levels.
Because of the unique and flexible architecture inherent to our Virtual Storage Architecture on the SolidPOWER7000 product line, GateStor Data Systems is well-positioned to meeting the future needs of the open system enterprise, by providing storage solutions which are both adaptable and extensible to continuous around-the-clock operations.
By allowing client systems or servers to connect directly to the Unified Bus, we can improve overall throughput and response time ten times fold. This requires adding a GateStor’s Unified Bus HBA on the client servers and the appropriated drivers. This will allow client computers not only to access the storage cluster faster but if desired to share the resources pool of the entire cluster.

Summary

In summary, the Virtual Storage Architecture has proven its viability and longevity over multiple generations of high performance, enterprise storage solutions for open systems. Today, only SolidPOWER7000 solutions built on the Virtual Storage Architecture give customers the ability to:

• Optimize performance for all types of data sizes and types through both automatic and selectable parity mode switching.

•Provide massive storage solutions, each node can support up to 1024 drives, with current disk capacity each node supports up to 4 PetaBytes of raw capacity multiplied by 256 nodes for a total cluster storage space of up to an Exabyte.

• Unified Bus Cluster Support thousands of concurrent host connections, providing enterprise-class storage services for any computing platform that supports industry standard interfaces.

• Define any amount of storage as a logical volume and allocate any amount of storage to any host.

• Select multiple concurrent levels of data protection (RAID 0, 1, 3, 5, 6, 10, 530, 560) for any logical data set – and to change data protection levels dynamically, in real time, without affecting operations and without any physical changes.

• Select any combination of concurrent multi-level fault-tolerant mirroring for any logical data set – and to change mirroring and other fault-tolerant characteristics dynamically, in real-time, without affecting operations and without any physical changes.

• Monitor all storage resource statistics from any SNMP management station, either from a central site or from multiple distributed management centers throughout the enterprise.

For more information please contact your GateStor Product Dealer.

 41 Simon St Section 2B. Nashua, NH 03060. Phone (603) 879-0216