



White Paper

Shared Internal Storage : An Introduction

A new storage paradigm for fully commoditized infrastructures

Executive Summary

The challenge posed by unstructured data storage requires a change in the way that storage systems are designed. This data requires both high performance and a vast quantity of storage space. Clustered storage provides an initial answer. The storage system is comprised of ordinary PC servers which play the role of storage servers. These servers are aggregated in a cluster to provide reliability, performance and capacity at a low cost thanks to the commoditized hardware that they use. This wave of commoditization in storage is quite similar to what application infrastructures have gone through in recent years. A large part of mainframe and large UNIX servers have gradually been replaced by groups of PC servers that are interconnected by an Ethernet Gigabit network. Seanodes proposes to push the commoditization of infrastructure to the limit by merging the commoditized application infrastructure with the commoditized storage infrastructure: ie - Shared Internal Storage (SIS).

SIS consists in transforming the PC servers' internal hard disks into an upmarket shared storage system, i.e. high performance, reliable and easy to manage. Each server in the infrastructure can run an application, store data or do both at the same time. SIS offers significantly better consolidation than a standard infrastructure in which the storage hardware is separate from the application hardware: it reduces infrastructure acquisition, administration and maintenance costs. SIS improves storage performance because each and every server in the infrastructure stores data and participates in processing storage I/O requests: in this way it offers much better parallelism than traditional storage solutions. Finally, SIS combined with server virtualization software brings a complete virtualization solution that reduces the need for externally shared storage systems and allows precise overall management of the Quality of Service provided by the virtual server.

But SIS is not just a concept; it exists and is currently in production. Seanodes has implemented it in its software, Exanodes. It is being used by HPC applications where it has already demonstrated a cost/performance ratio beyond comparison. The new version of Exanodes extends its field of use to the domain of virtual machines, in particular that of Hosted Services and Test and Development.

TABLE OF CONTENTS

SHARED INTERNAL STORAGE (SIS):	3
A NEW CATEGORY OF STORAGE FOR THE COMMODITIZATION REVOLUTION	3
3 TRENDS INSPIRED BY THE COMMODITIZATION REVOLUTION	3
SHARED INTERNAL STORAGE	6
<i>SIS: The Definition</i>	<i>6</i>
<i>The advantages of SIS over other forms of storage</i>	<i>6</i>
SIS CHALLENGES	9
SIS: CHALLENGES AND KEY DIFFERENTIATORS	9
EXANODES : SIS IN MOTION	10
OVERVIEW OF EXANODES	10
MAIN ADVANTAGES OF EXANODES	11
<i>High-performance storage</i>	<i>11</i>
<i>Scalable storage</i>	<i>12</i>
<i>Reliable storage</i>	<i>12</i>
<i>Simplicity</i>	<i>13</i>
<i>Cost control</i>	<i>13</i>
CONCLUSION	14
ABOUT SEANODES	15

FIGURES

Figure 1. Server Virtualization Needs Shared Storage	4
Figure 2. Relative Performance Improvements of System Components	5
Figure 3. Shared Internal Storage and Server Virtualization	6
Figure 4. Shared Internal Storage : The Path to Full Infrastructure Consolidation	8
Figure 5. Creating Shared Internal Storage with Exanodes	11
Figure 6. Exanodes 2.3 Benchmarks	12

1 Shared Internal Storage (SIS): A new category of storage for the commoditization revolution

The challenge posed by unstructured data storage imposes a change in the way that storage systems are designed. This data requires both high performance and an enormous quantity of storage space. To meet these requirements by using traditional SAN or NAS systems sends costs sky high.

To increase performance while reducing cost, the storage system architecture is going through an in-depth transformation through the use of commoditized hardware. Large monolithic storage systems with specific hardware are being replaced by clusters of small commoditized storage servers. The use of hardware from the world of PC servers results in an excellent cost/performance ratio and thanks to fault tolerance software layers this is achieved without sacrificing reliability.

The hardware commoditization revolution that we are experiencing in the storage world is similar to that which has been taking place in the application world, in datacenters, or high performance computing centers. Large UNIX, mainframe or supercomputer systems are gradually being replaced by smaller, less expensive groups of PC servers with a standardized architecture.

Seanodes proposes to push commoditization to its limits and to achieve an incomparable infrastructure cost/performance ratio through its new category of storage, Shared Internal Storage (SIS). This is the culminating point of 3 trends that define the hardware commoditization revolution.

3 TRENDS INSPIRED BY THE COMMODITIZATION REVOLUTION

Hardware commoditization has lead to 3 trends:

- Server Virtualization has become a must.
- Storage performance is under extreme pressure.
- The frontier between storage servers and application servers has become blurred.

The full virtualization of commoditized architectures: a must

The commoditization of datacenter hardware has lead to a shift in cost structure. Although infrastructure hardware costs are decreasing, management costs are increasing. The multiplication of the number of PC servers in the infrastructure is leading to problems managing heterogeneity and reliability.

Server Virtualization software is one answer to this problem: server virtualization is a layer of abstraction that masks the hardware complexity of the commoditized infrastructure and replaces it with an infrastructure of virtual application servers that can be created, configured or eliminated on the basis of application needs. An application running on a virtual server is no longer blocked within a physical server. Virtualization of PC servers allows consolidation of physical servers and considerably reduces management costs linked to hardware commoditization.

But this virtualization is incomplete. Server virtualization software does not appropriately virtualize internal hard disks on physical servers and therefore imposes the use of NAS or SAN shared storage. In fact a virtual server moves within a hardware infrastructure. For example, it can run on one physical server and then migrate to another physical server. If the application that is running on the virtual server writes data to the internal hard disks of the first physical server, this data will be inaccessible when the virtual server migrates to the second physical server. Consequently, a shared infrastructure - NAS or SAN - must be in place.

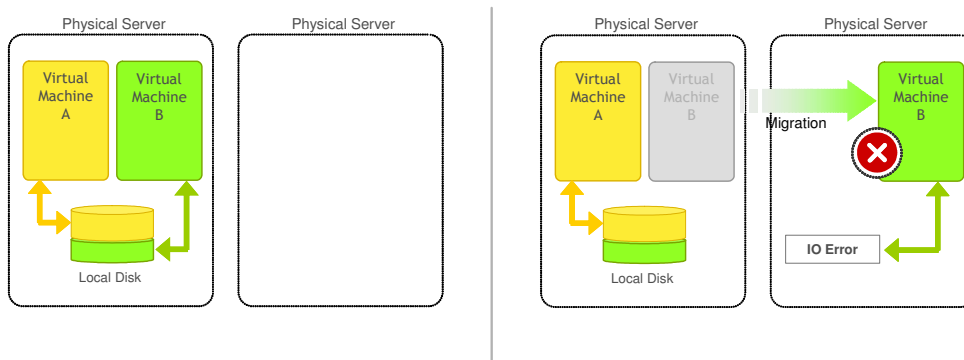
This incomplete virtualization of PC servers creates a problem: effectively it is used to consolidate application layer hardware but it necessitates adding resources to the storage part. Finally, the overall infrastructure is insufficiently consolidated because savings in hardware achieved on the application side of the infrastructure are partly lost due to the need to add hardware at the storage infrastructure level.

Commoditized infrastructure's complete virtualization (i.e. virtualization that also includes physical servers' internal hard disks) is necessary to avoid the growth of the infrastructure's storage. **Full consolidation needs full virtualization.**

Figure 1. Server Virtualization Needs Shared Storage

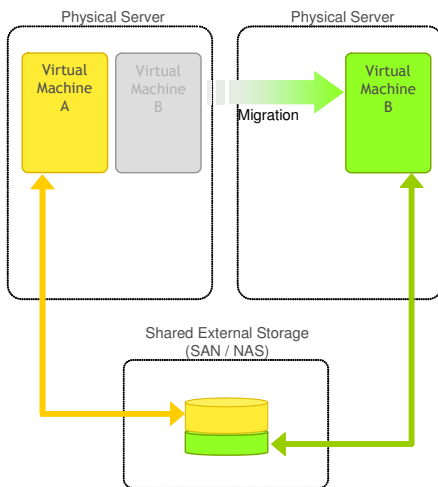
Server Virtualization Needs Shared Storage

Virtual machine migration is impossible if the data is isolated on the local disks



With shared storage, data is accessible from any server.

Therefore VM migration is possible.



Commodity hardware: the gap between storage performance and CPU/network performance is widening

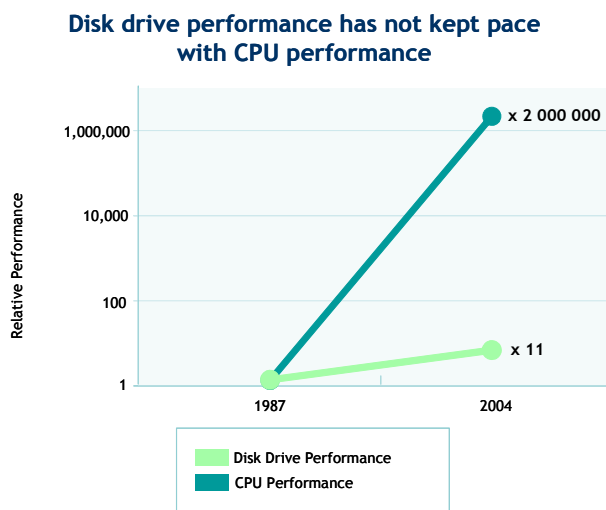
In commoditized architectures, hard disks are the only storage resources available. More effective storage resources - such as solid state disks and battery backup RAM - have not yet been commoditized and are reserved for expensive, upmarket storage systems.

There are two trends in the evolution of hard disks that are due to the basic operating principles of a hard disk and that are leading to gradual reductions in the relative performance of traditional storage systems:

- The performance of CPUs and networks has evolved considerably faster than the seek time of hard disks. Storage performances have not kept pace with CPU and network performances.
- The capacity of hard disks is rocketing: 1 TB for the Hitachi Deskstar 7K1000 but still one single disk spindle per disk, i.e. a single entry point. The size of the reservoir is growing exponentially but the pipe is still as small as ever: it is becoming more and more difficult to access data on a hard disk.

Whilst waiting for the commoditization of new storage technologies such as solid state disks, increasing performance in a commoditized architecture means opting for solutions that increase the density of the disk spindles available for each server. **Data should be attributed more disk spindles per TB.**

Figure 2. Relative Performance Improvements of System Components



Source : Seagate White Paper TP-525
(http://www.seagate.com/docs/pdf/whitepaper/economies_capacity_spd_tp.pdf)

The frontier between storage servers and application servers has become blurred

The new storage system architectures are clustered. Large monolithic storage systems with specific hardware are being replaced by clusters of small commoditized storage servers (nodes). To create a cluster node, the commoditized architecture of a PC server is perfectly adapted: effective, polyvalent and inexpensive. Commoditized hardware's lower reliability is compensated by the fact that storage clusters tolerate several complete node failures. What distinguishes these new products from one another is no longer the hardware but the software. **The new storage system is no longer hardware based - it's software based.**

The boundary line between a storage server and an application server is becoming blurred. Both are PC servers. Only the software defines the server's role. In datacenters, we are beginning to see two similar but separate infrastructures: on the one hand a set of PC servers that run applications and on the other hand a set of PC servers that store data. Why keep them separate? Combine them!

To the 3 trends described above, Shared Internal Storage (SIS) provides the perfect answer. In the next part of this document we will take you through the advantages of the SIS concept.

SHARED INTERNAL STORAGE

SIS: THE DEFINITION

Shared Internal Storage is shared, virtual storage created from internal hard disks in PC servers, whatever the server type, be they storage servers, application servers or both.

SIS consists in transforming PC servers' internal hard disks into upmarket shared storage: effective, reliable and easy to manage (not to be confused with clustered external storage which relies only on dedicated storage servers).

THE ADVANTAGES OF SIS OVER OTHER FORMS OF STORAGE

SIS is a form of storage that is completely adapted to commoditized hardware architecture and corresponds perfectly to the 3 trends described previously.

The full virtualization of commoditized architectures: a must

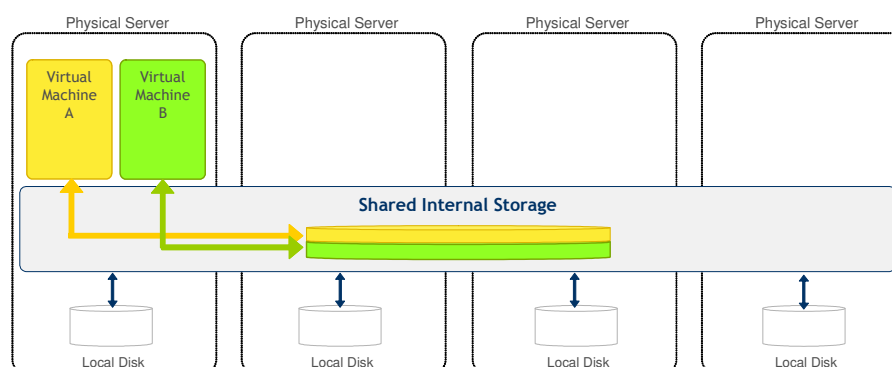
With SIS, the application virtualization layer has shared storage that reuses the servers' internal hard disks.

- A single infrastructure facilitates administration.
- With SIS, PC server virtualization is complete. The infrastructure virtualization software masters all the physical resources, CPU, network and storage, and overall can manage the Quality of Service (QoS) provided by the virtual servers. It no longer depends on more or less advanced NAS or SAN external storage system QoS management functionalities. It is simpler to make application requirements fit storage requirements.
- External shared storage infrastructures - such as NAS or SAN - are no longer necessary.

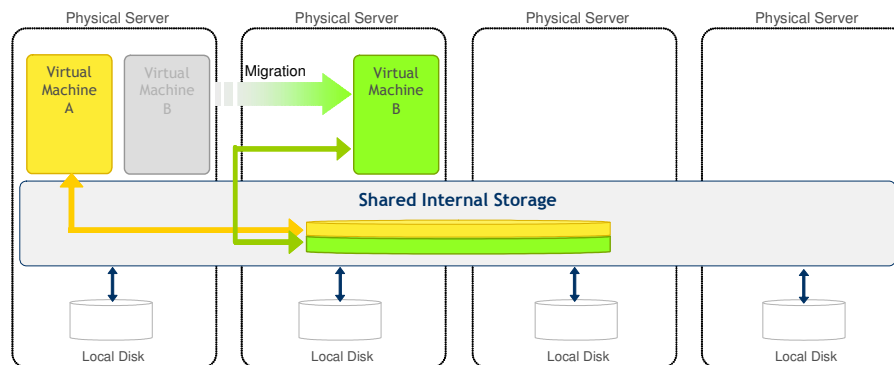
Figure 3. Shared Internal Storage and Server Virtualization

Shared Internal Storage and Server Virtualization leads to Full Virtualization

Shared Internal Storage allows Virtual Machines to share local disks.



Virtual Machine migration is possible without the need for Shared External Storage



Storage performance: reducing the gap between CPU and hard disk performance

SIS is taking shared storage performance a leap forwards:

- SIS reactivates hard disks and RAID controllers which up until now have been underexploited because they have been isolated in each server. These become shared resources and are therefore, easy to use. The resulting architecture has a very large capacity for parallelism, for more hard disks (or slots that can hold new hard disks) and therefore for more disk spindles.
- SIS is scalable because in adding a PC server, CPU and storage power is added in the same proportion.
- In comparison to a SAN or NAS solution, SIS allows a compromise to be made between the use of a server's local disks or another server's remote hard disks, which considerably reduces the data that is transferred over the network.
- All of the servers' CPUs can process storage I/O requests, which provides the storage system with considerable I/O processing power.

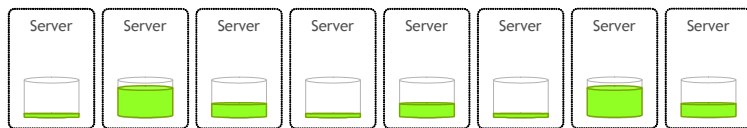
The frontier between storage servers and application servers has become blurred

With SIS, each PC server can have an application server role, a storage server role - or both. This has the following advantages:

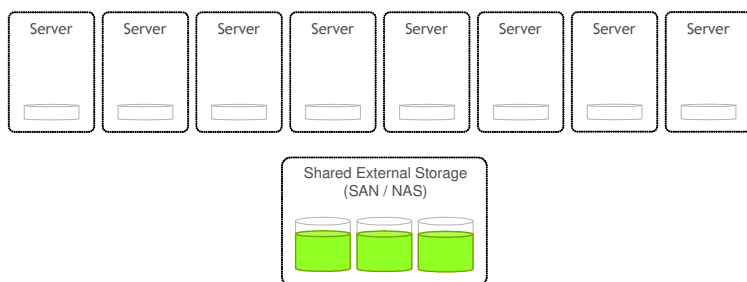
- No infrastructure duplication: SIS renders the separation of storage infrastructures and application infrastructures futile. Storage servers CPUs can now run applications, and application servers' hard disks can store data. Full infrastructure consolidation has now become a reality.
- Reactivation of seldom-used hard disks: PC servers' onboard disks can be fully exploited by SIS. Storage power increases at zero cost.
- Flexibility in infrastructure management: to change the PC server's role you just need to re-configure the SIS software. The same server can easily switch from application server role to mixed application/storage role or pure storage role. Depending on needs, the infrastructure administrator only has to activate or deactivate certain software functionalities to change a server's role.

Figure 4. Shared Internal Storage : The Path to Full Infrastructure Consolidation

**The Path to Full Infrastructure Consolidation :
 From Direct Attached Storage to Shared Internal Storage**

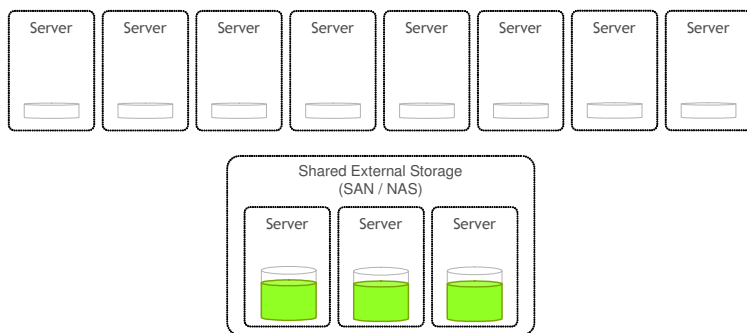


Direct Attached Storage
 Resources can't be shared.
 Storage capacity can't be optimized.

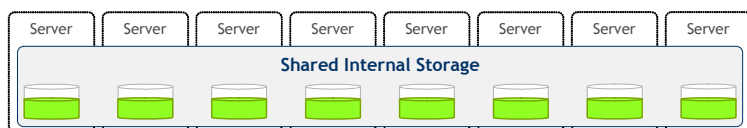


Disk storage consolidation occurs thanks to shared external storage:
 - Disk capacity is optimized
 - Management is easier

New shared external storage systems take advantage of a clustered architecture to increase the performance/cost ratio.



But Clustered storage leads to infrastructure duplication: within the infrastructure all servers have similar architectures. Some servers are storage dedicated while others are application dedicated.
 - Application servers' disks are underused
 - Storage servers' CPUs are underused



Full Infrastructure Consolidation takes place through Shared Internal Storage : each and every server is both an application and storage server
 - All disks are fully used
 - All CPU's are fully used

Storage and CPU resources are fully optimized

2 SIS CHALLENGES

In comparison with an external clustered storage system, developing a SIS system requires taking on extra challenges.

SIS: CHALLENGES AND KEY DIFFERENTIATORS

A small foot-print required: the storage software shares the same CPU, RAM and network resources as the application

- Extremely simple client/server protocol: in a system that is as parallel as SIS, a large quantity of I/O requests are exchanged between clients and servers. It is important for the exchange protocol to be simple in order to save CPU and network bandwidth. For example, the iSCSI protocol is not well adapted to SIS. iSCSI is an excellent protocol to provide interoperability between clients and storage servers developed by different companies, but interoperability is not very useful in a SIS where communication between clients and storage servers is part of the storage system internal software layers. Using only low-level network APIs to reduce the foot print on the system is therefore better adapted.
- In the design of the SIS, care must be taken to make sure that the memory resource, CPU and network resource consumption is independent to the number of storage servers or clients because in comparison to a traditional storage system the number of servers is considerable.
- A SIS system should allow the user to specify the amount of CPU and network bandwidth attributed to the storage system. This allows the user to reserve resources for the application that is running locally on the server

Application servers' internal hard disks are less reliable

- SIS reuses hard disks located on application servers. These disks are less reliable because an application server is, for example, more subject to rebooting after software updates or bugs than a server dedicated to storage whose software is less complex and varies little. A SIS system should be particularly failure tolerant. This is attained:
 - by integrating RAIN systems (Redundant Array of Independent Nodes) that tolerate numerous simultaneous server and disk failures and cumulative failures.
 - by integrating highly effective data rebuilding mechanisms to cope with disk failures: immediate rebuilding of data on hard disks' spare zones, parallel rebuilding and the rebuilding of the data which has modified since the last failure, only.

Hardware is fully commoditized

- A SIS system should operate efficiently on any commoditized server.
- It should be capable of exploiting any type of block device: SATA/SCSI/SAS disk drives, RAM disks or flash disks, internal RAID or DAS, and any type of interconnection network: Gigabit Ethernet or Infiniband.
- It should be unintrusive to the servers' OS, install easily and be compatible with a large number of versions of different OSs.
- The design of a SIS cannot rely on specific non-commoditized hardware (for example NVRAM -Non Volatile RAM- or ASICs specialized in accelerating storage performances).
- A fully commoditized architecture is heterogeneous: the servers do not have the same CPUs, memory capacity or internal disk storage capacity. SIS should allow easy management of this heterogeneity.

Each server is a potential storage server

- Each and every machine in the infrastructure can be a storage server. A SIS system is therefore distributed over a number of servers far greater (hundreds of servers) than those found in external Clustered systems.
- In a SIS, there are as many storage servers as storage clients. If I/O scheduling is carried out at client level as is the case in iSCSI or in the SAN Fibre Channel, each server receives few queued requests. Server-based I/O scheduling algorithms are necessary in order to provide performance. In this case, the clients send all their waiting requests to the servers and it is the servers that decide on the processing order of these requests.
- A SIS should operate in perfect symmetry. A machine which specializes in a specific task would behave differently to other machines or would require specific hardware. All machines run the same software.

3 EXANODES : SIS IN MOTION

But SIS is not just a concept; it exists and it is being exploited in production. Seanodes has implemented it in its software, Exanodes. Exanodes can create a SIS from any group of Linux servers. This section briefly describes the product as well as its features/benefits. For a detailed presentation of Exanodes, refer to the whitepaper "Exanodes: A new paradigm for Linux Cluster Storage".

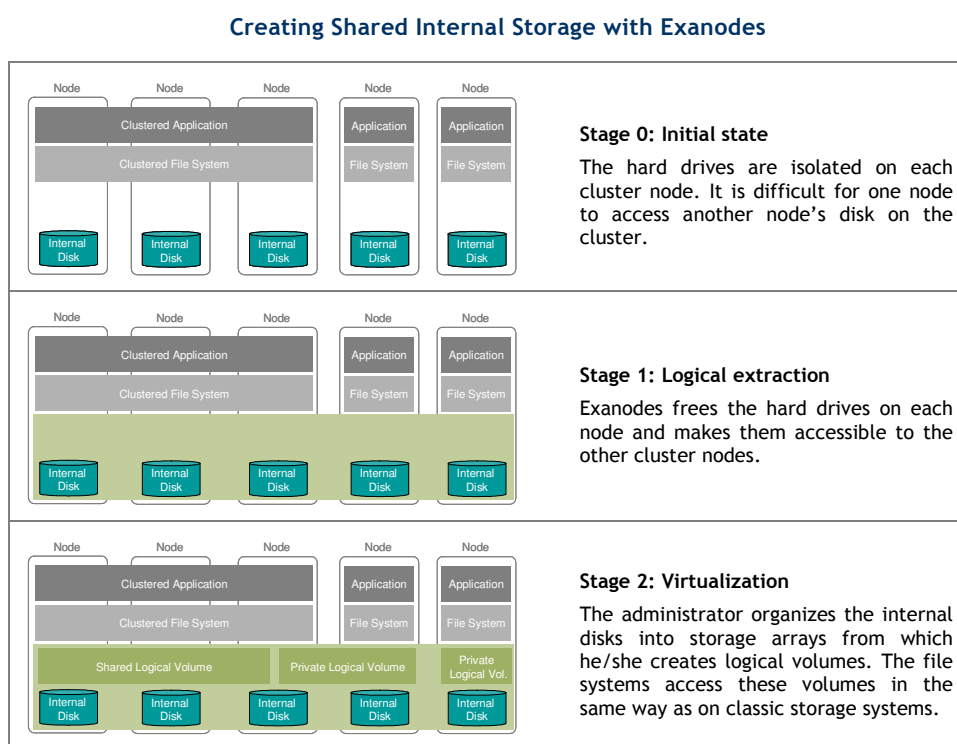
OVERVIEW OF EXANODES

Exanodes creates virtual storage subsystems on the cluster nodes unused disk space. These virtual arrays are accessed via the interconnection network. Applications then have at their disposal shared high-performance storage volumes to store their data.

Exanodes processes all data requests sent to the storage arrays created in this way. It also delivers the intelligence required for data allocation in order to balance the load evenly on all disks, depending on the characteristics of the application workload. For increased concurrency between the processing of requests, Exanodes runs simultaneously and symmetrically on all application servers.

Exanodes provides the file system with storage volumes that are accessible in block mode. These volumes have all the features of LUNs that are generated by a classic storage array. Exanodes makes it possible to partition the virtual array (definition of LUNs), and allocate partitions to specific nodes or mask them (LUN masking), in a totally seamless way.

Figure 5. Creating Shared Internal Storage with Exanodes



MAIN ADVANTAGES OF EXANODES

Exanodes gives commoditized server users a high-performance clustered storage system that is economical and easy to manage.

HIGH-PERFORMANCE STORAGE

Storage system with high parallelism

When configured in symmetric mode, each computing node is also a storage node. The processing of inputs/outputs is done simultaneously on a large number of nodes. The application therefore benefits from cluster parallelism for computing as well as for storage.

Fine tuning for better performance levels

Applications have diverse requirements and need different storage system configurations. Consequently, some applications tend to perform a lot of small input/output operations, randomly spread out over the storage volume while other types of applications tend to generate large sequential inputs/outputs.

In order to deliver the storage solution that matches the needs of the application, Exanodes' various technological components can be fine tuned to achieve higher performances (with respect to data layout, use of the network, balance between I/Os, computing, etc.).

Exanodes' high parallelism provides excellent performances on a 32 application server platform as shown in the table below

Figure 6. Exanodes 2.3 Benchmarks

Exanodes 2.3 Benchmarks

Computing Platform	
Servers	32 servers HP DL 145 (2 disk drives 36 GB)
Network	Gigabit Ethernet: HP ProCurve
OS	Linux RedHat EL3, kernel 2.6.9-42el5mp
FS	Ext3
Dedicated Hardware Storage Platform	
Nonexistent : Replaced by Exanodes 2.3 and application servers' internal disks	
Benchmark	
Bonnie++	
Shared Internal Storage Performance	2.3 GB/s

SCALABLE STORAGE

Natural scalability

With Exanodes, the storage architecture is naturally linked to the scaling of the computing architecture: adding an application server increases both computing and storage performance.

Highly adaptable hardware configuration

Exanodes makes it possible to combine and use all types of storage devices accessible in block mode, such as SATA or SCSI disks, software and hardware RAID, etc. This way, the storage system naturally scales with the latest available technologies, which give it the best capacities and performance levels. It's the guarantee of upgrading your clusters at the same time as new technologies become available while increasing performance levels.

RELIABLE STORAGE

Data availability

To increase data availability, Exanodes provides a RAIN (Redundant Array of Independent Nodes) storage system. All data is replicated on two disks belonging to two different nodes. When a node fails, Exanodes' RAIN system is able to carry on processing the requests, because data on the failed node's disks is replicated on the surviving nodes' disks. Applications therefore have uninterrupted access to their data. Intelligent data rebuilding mechanisms allow to rapidly tolerate new disk failures.

Data protection

Exanodes also virtualizes any internal RAIDs on the commoditized servers. By combining the protection of internal RAIDs with that of the RAIN, Exanodes provides highly reliable storage which can withstand numerous disk failures.

SIMPLICITY

Exanodes stands out from other high-performance storage solutions as it is a storage system that is simple to set up, maintain and administer on a day-to-day basis. It simplifies several aspects of application management such as the redevelopment of applications designed for an SMP architecture into cluster-type applications.

Shared storage system

Exanodes has all the advantages of a storage system shared over a pool of servers. In this way data management is simplified. With Exanodes, applications can, for instance, be run on any server in the infrastructure and access their data in a highly effective way. Within the same infrastructure, applications become truly mobile because it is no longer necessary to copy data to the nodes where it is being used. On the other hand, computing results are accessible from any node, taking away the need for staging out results after the computing phases.

Natural compliance with standards

Exanodes has an extensive compatibility matrix. It supports all types of block devices (disk partitions or whole disks, PATA, SATA, SCSI, JBOD, hardware RAID, software RAID, RAM disks, etc.) as well as the major high-performance networks (Infiniband, Gigabit Ethernet and SCI), Linux private file systems (ext2, ext3, XFS, JFS, etc.) and clustered file systems (GFS, Lustre, GPFS, etc.). This way, users retain the freedom to choose the software and hardware technologies best suited to the requirements of their applications in order to optimize each layer of their infrastructure.

No change to existing system

Exanodes is non-intrusive. Applications access Exanodes' high-performance storage area without any need for reprogramming. Moreover, Exanodes is installed without any need to modify the Linux kernel.

Simplified administration

The day-to-day administration of Exanodes is very simple. Numerous functionalities are available for the efficient management of the various storage volumes (supervision of logical volumes and file systems, hot-resizable logical volumes, management of nodes accessing storage resources, etc.).

The user has the choice between 2 administration modes:

- Administration via "command lines"
- A GUI mode that simplifies the administration and supervision of the storage infrastructure

COST CONTROL

Switching from traditional to commoditized architecture is only truly worthwhile if storage system acquisition costs remain reasonable and if the administration of the whole system is not too costly in terms of man-hours. For both of these concerns, Exanodes provides an appropriate answer.

Considerable reduction of external storage needs (SAN or NAS)

Exanodes drastically reduces storage infrastructure acquisition costs. This is because external storage needs (SAN or NAS) are greatly reduced as applications are no longer required to access these systems directly. External storage is confined to less demanding uses: archiving of computing results, data distribution for workflows involving several clusters, backup management, etc.

Reduction in operating costs

The simplicity of the Exanodes solution (integration and day-to-day management) increases the amount of storage and computing nodes that one person can administer. In addition to acquisition costs, substantial savings can thus be made throughout the use of the infrastructure.

4 CONCLUSION

The challenge posed by unstructured data storage requires a change in the way that storage systems are designed. This data requires both high performance and a vast quantity of storage space. Clustered storage provides an initial answer. The storage system is comprised of ordinary PC servers which play the role of storage servers. These servers are aggregated in a cluster to provide reliability, performance and capacity at a low cost thanks to the commoditized hardware that they use. This wave of commoditization in storage is quite similar to what application infrastructures have gone through in recent years. A large part of mainframe and large UNIX servers have gradually been replaced by groups of PC servers that are interconnected by an Ethernet Gigabit network. Seanodes proposes to push the commoditization of infrastructure to the limit by merging the commoditized application infrastructure with the commoditized storage infrastructure: ie - Shared Internal Storage (SIS).

SIS consists in transforming the PC servers' internal hard disks into an upmarket shared storage system, i.e. high performance, reliable and easy to manage. Each server in the infrastructure can run an application, store data or do both at the same time. SIS offers significantly better consolidation than a standard infrastructure in which the storage hardware is separate from the application hardware: it reduces infrastructure acquisition, administration and maintenance costs. SIS improves storage performance because each and every server in the infrastructure stores data and participates in processing storage I/O requests: in this way it offers much better parallelism than traditional storage solutions. Finally, SIS combined with server virtualization software brings a complete virtualization solution that reduces the need for externally shared storage systems and allows precise overall management of the Quality of Service provided by the virtual server.

But SIS is not just a concept; it exists and is currently in production. Seanodes has implemented it in its software, Exanodes. It is being used by HPC applications where it has already demonstrated a cost/performance ratio beyond comparison. The new version of Exanodes extends its field of use to the domain of virtual machines, in particular that of Hosted Services and Test and Development.

ABOUT SEANODES

Seanodes was founded at the end of 2002, as a result of several years of research within IRIT, l'Institut de Recherche en Informatique de Toulouse. Inventor of the Shared Internal Storage (SIS) concept, Seanodes has broken new ground in the shared storage world with the launch of its first product, Exanodes.

Seanodes' ambition is to extend the principles of commoditized distributed open systems to the world of data storage. Exanodes, the first product released by Seanodes makes it possible to drastically reduce storage system costs while significantly improving performance.

Seanodes has received several awards since its creation - including first prize in the 5th *Concours National de Création d'Entreprises de Technologies Innovantes* which awards the most promising young innovative French start-up of the year in the technology sector - for its patented technology and revolutionary product, Exanodes.

www.seanodes.com

Sales and Product Information

sales@seanodes.com

Channel & OEM

channel@seanodes.com