SANCluster

InfinityScale Storage

Chapter 1: Introduction

SANCluster InfinityScale Storage (www.sancluster.com) is an affordable high-performance, scale-out solution. Different from traditional data storage solutions, we have eliminated performance bottlenecks caused by SAN controllers or NAS gateways. Using only commodity hardware, on a 2 PB rack of 4U 36-drive servers equipped with 6 TB SATA disks and dual 10 GbE network, our system can deliver over 18 GBps throughput in aggregated performance.

Many repeating customers started with a few hundreds of TBs and expanded to the PB level over the years. With over 150 PBs deployed, some of the industries we support include:

Scientific computing/high-performance computing

Oil and Gas, Life sciences/Generic research, and analyses in computational physics and chemistry, material sciences, aerospace engineering and etc.

Media and Entertainment

Film Rending, IPTV, Non-linear Editing (NLE), and Media Asset Management.

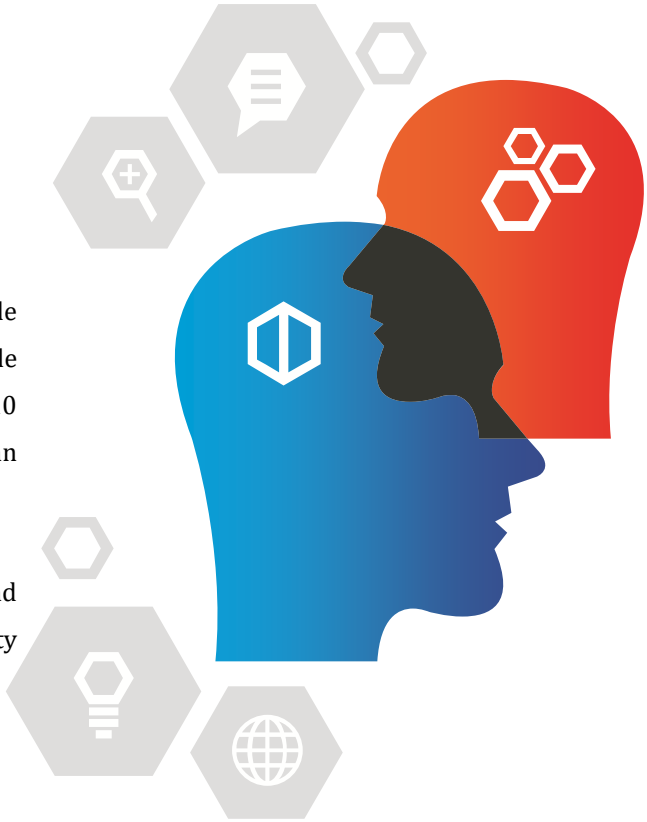
Video surveillance

High definition video recording and monitoring (24/7/365), with WORM - write once and read many technology.

Telecommunication

In a mix environments included CDN files, social network data, big data analytic and etc. Telecommunication industry needs high performance and scalable systems to address their new challenges.

❖ Key system values



High performance

Using SATA disks with dual 10 GbE network cards, the average of a single disk's performance in our system is about 50 MB/s to 60 MB/s. A single 4U 36-drive server can deliver over 1.8 GB/s throughput. A cluster of 10 of such storage nodes can provide 2 TB in capacity and more than 10*36*50 MB/s = 18 GB/s in aggregated performance.

Using SSD disks with dual 40 GbE network cards or 100 Gb Infiniband card, A cluster of 4 of such storage nodes can provide 192 TB in capacity and more than 156 GB/s write performance.

Simple SANCluster ability

"The clustered solution starts with a minimum of 3 storage nodes, few TBs in capacity and can be scaled out to 300 PB (single volume) by simply adding storage nodes. The SANCluster Storage solution is truly scale-out, expanding capacity on-demand and online, without interruption of work operations.

The software switch automatically load balances for newly-added storage nodes. Customers have the option to use 1 TB to 6 TB SATA disks or mix with SATA, SAS, and SSD disks for a Hierarchical Storage Management (HSM) solution."

Strong reliability

The system's automated self-monitoring mechanism can single out inactive hardware, either at the disk or server level, to provide high data availability, eliminating system downtime.

File-level RAID enables up to 80% capacity utilization and fast data recovery in the event of hardware failure. Typical data recovery time is only 1/5 the time of traditional RAID; 3 TB in 30 minutes.

Easy management

Single log-on page for simple management: one part-time IT staff can easily manage PB-levels of storage.

Reasonable cost

Leverage only commodity hardware, no vendor lock-in. Hardware cost of a 2PB system can be as low as per TB.

❖ Solution Background

Supercomputing

Compared with mainstream enterprise applications, compute-intensive, high-performance computing (HPC) places very different demands on storage systems.

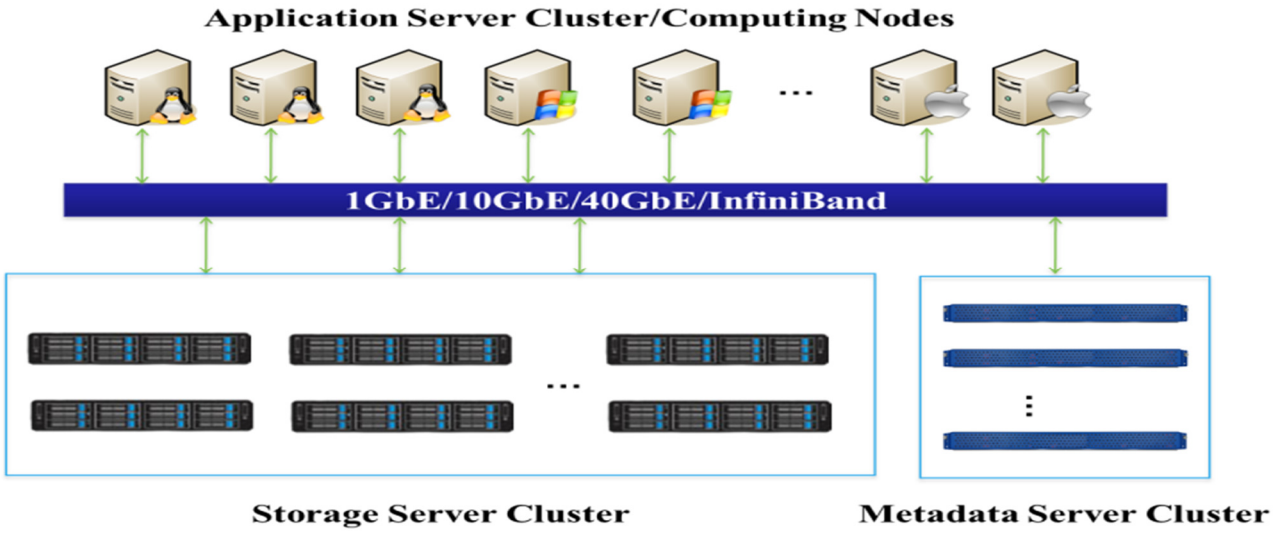
An HPC storage system needs large capacity accessible at high speed and to be highly expandable, while offering a single global namespace accessible to all users involved in the project.

Internet - Big Data

At root, the key requirements of big data storage are that it can handle very large amounts of data and keep scaling to keep up with growth, and that it can provide the input/output operations per second (IOPS) necessary to deliver data to analytics tools.

Chapter 2: System Architecture

SANCluster **InfinityScale** Storage architecture consists of two main parts: the metadata server cluster and the storage server cluster, as shown in the diagram below. A typical system will have 3 or more storage servers, and at least one pair of metadata servers. Network options include 10 GbE ,40 GbE up to 200 Gb InfiniBand. Different from typical SAN or NAS storage, there is no controller or gateway in our system. By separating metadata from storage nodes, our solution can provide the efficient handling of either large size (>10 GB) files or small size (kb-sized) files.



1. Metadata Server Cluster

The metadata server manages the functions of file creation and file open in the system. The cluster has three key functions. First, it manages the metadata of the file system such as directory (tree) structure, time of creation, owner ID, access permission, and etc. Second, it develops the global namespace, allowing each application server to access the files and coordinates the data traffic between the application server cluster and the storage server cluster. Finally, it provides the management interface for the systems.

Since the index and metadata information of a file usually are a few hundred bytes in size, we equip each metadata sever with two SSDs. Leveraging SSDs' performance as a storage medium, and with the specially designed algorithm of our proprietary file system, our system can manage more than billion files. In an actual deployment of a 5 PB system, there are 18 metadata servers actively managing over 40 billion files.

Separating metadata from the storage node has eliminated the system concern on file limitation under a single directory. SANCluster Storage has an actual customer case with tens of millions of files built in a single directory. And there is no limit on the number of directories in one system.

2. Application Server Cluster (Computing Nodes)

By installing client-side software (a MB size driver), these application servers can access the storage cluster as a local drive. Data can be stored and shared in one single storage pool.

Multiple terminals of various operating systems such as Window, Linux, Mac, and Unix can simultaneously access the storage system. We have theoretical support on cluster sizes of 40,000 servers, with actual deployments of more than 1,000 servers in a high-performance computing site.

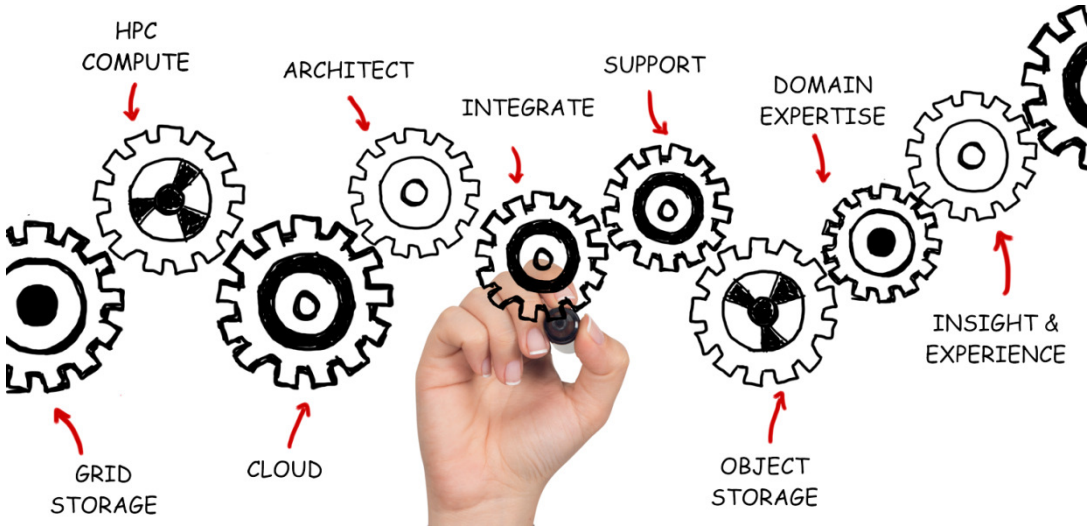
3. Storage Server Cluster

Storage servers (aka storage nodes) manage data storage and provide I/O services to application servers. Data is replicated and distributed across multiple storage servers. We offer theoretical support on cluster sizes of over 10,000 servers providing Exabyte (EB) level capacity and have actual deployments of more than 400 servers in one cluster.

When a file is being stored, the application server will first communicate with metadata server to receive instructions on where to store the file. And then the file will be sliced up, if the file is more than 64 MB, or, if it is less than 64 MB simply stored to the storage server. The metadata server is only involved, when there is a metadata-related operation. When application servers access data, they only communicate to the metadata controller once, and access the requested data from storage servers without further interaction with metadata servers.

Chapter 3: Software Suite

SANCluster **InfinityScale** Storage's solution values are derived from our proprietary software suite. The goal of our development effort is to provide customers with great value on all of the important aspects of a data storage solution, without having to choose one over the other. The values and overall deliverables of our system are: **PERFORMANCE, SCALABILITY, RELIABILITY, MANAGEABILITY and LOW COST.**



✓ Application Layer

Client-side software

MB-size drivers supporting various Windows, Linux, Unix and Mac systems, including:

- Linux: Redhat, CentOS, Suse, Opensuse, Ubuntu, Gentoo, Fedora Core
- Windows: 2000, Server 2003, XP, 2008, Win7, Win8 and upper
- Mac OS X: 10.6, 10.7, 10.8 and upper

Management side

SANCluster Storage provides a centralized management GUI to configure, manage and monitor the storage systems with a single-page log-on. One part-time technical staff can easily manage PBs of storage.

The web-based GUI interface provides a dashboard for cluster monitoring and a smart alert system. Daily performance and statistics are recorded and monitored in the dashboard. Information about the system health can be exported to different formats for review. When a failure occurs, the systems will go into a self-healing process automatically and send an alert to administrators at the same time. Usually the failed hardware does not need to be replaced immediately, but can be deferred to a later time which is more convenient to the administrators.

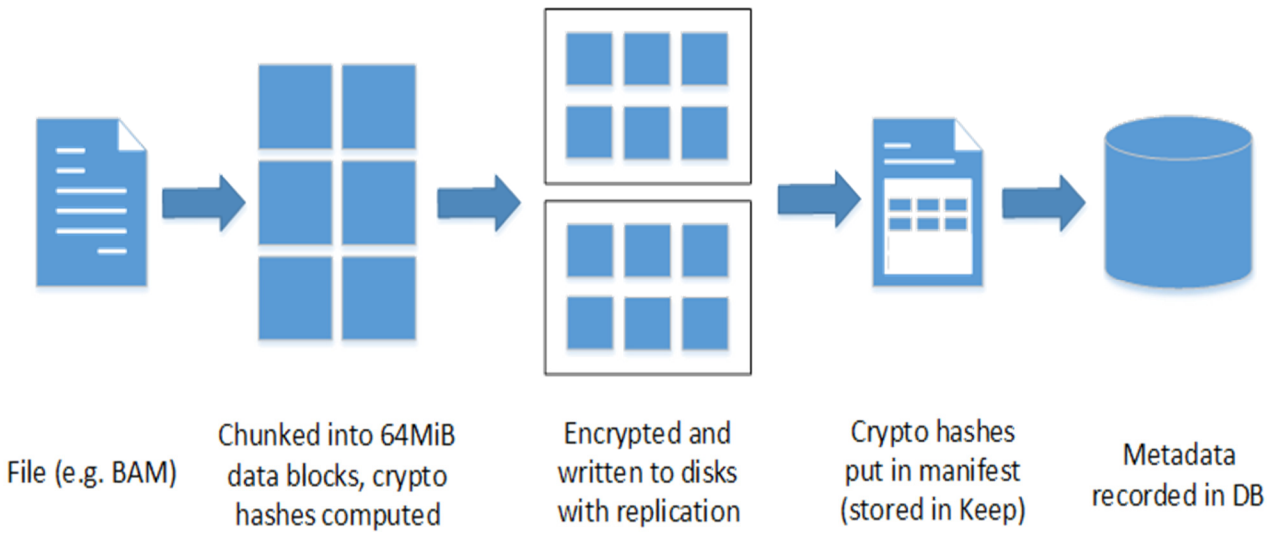
✓ Base Layer

Proprietary file system – SANCluster InfinityScale file system

InfinityScale file system is a proprietary and completely POSIX-compliant cluster file system. It replicates and distributes files across nodes in the storage server cluster.

Key Functions of SANCluster **InfinityScale** file system is:

- Virtualizes storage resources across all available storage servers into a unified storage pool and provides a single global namespace
- Controls metadata servers to provide support on data dispatch, including metadata-intensive applications and a large amount of concurrent file access
- Manages the system's automated self-monitoring mechanism, consistently checking system condition and performance level, it can proactively single out near-inactive stage hardware, either disks or servers, to provide high data availability, eliminating system downtime
- Provides system maintenance and upgrade



Filesystem working diagram

✓ Functional Features

- **Node communications**

  Communications among storage nodes, running Heartbeat, nodes communicate by exchanging packets at the rate of approximately 2 Hz (twice per second). When Heartbeat is used in our multi-node system, one machine is designated as the primary node and the other as the secondary node. If the primary node fails or requires downtime, the secondary node can take over the primary role. The communication is also used by storage nodes for data recovery.

- **Cache management**

  All data being written and read must go through cache. SANCluster **InfinityScale** has a caching algorithm in the system so frequently accessed data can be placed in cache. Cache can be created in all the nodes in the system, including application / computing nodes, storage nodes and metadata servers.

- **Automatic alert**

  When system has certain failure or abnormal signs, built-in monitoring feature will automatically send out alerts, either on the monitoring page or through administrative email.

- **Capacity constraint**

  The system can set capacity constraint on any of the directories. Once reaching the upper limit of capacity, no data can be written onto the directory.

- **Data equalization**

  All data being stored can be managed by the load balancing mechanism. If any of the disks in the system are not being fully utilized, with the load-balance switch being turned on, the system will start migrating data among storage disks until the capacity of all disks is balanced.

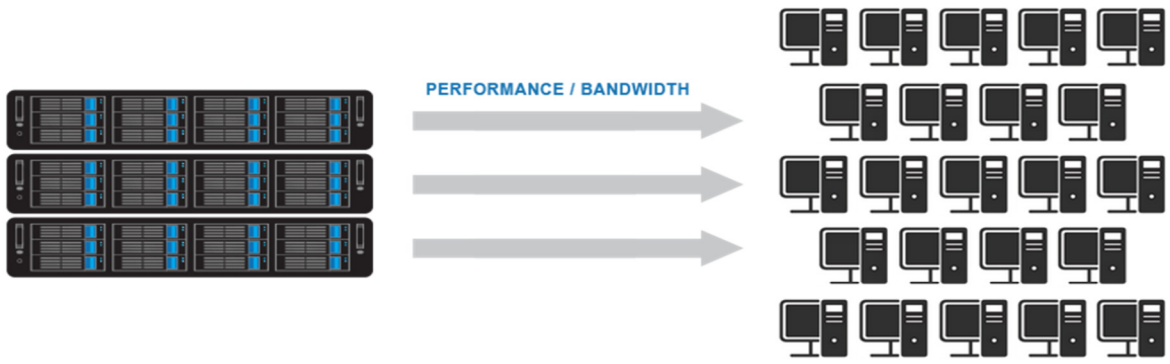
- **SCALABILITY**

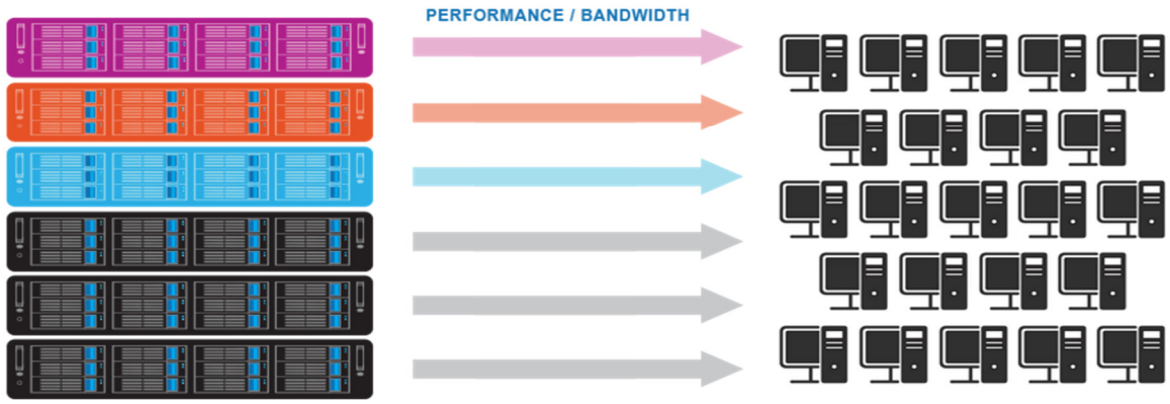
SANCluster **InfinityScale** Storage allows dynamic expansion of storage capacity without interrupting system operation or downtime. New disks or servers can be added to an existing system with just a few clicks through the centralized management GUI page. The entire process is completely transparent to application servers and there is no downtime involved. After adding additional storage node(s), the aggregated throughput of the whole cluster will increase linearly and instantly. The metadata cluster can be expanded in the same way.

While using all 4 TB SATA disks, on a storage node with 12 drives, raw capacity per node is 12*4 TB = 48 TB. On a system with 3 storage nodes, our solution can offer a raw capacity of 3*48 = 144 TB, with a capacity utilization of 50%. On a system with 6 storage nodes, our solution can offer a raw capacity of 6*48 = 288 TB, and capacity utilization can be up to 80%.

Adding one more storage node will increase system raw capacity by 48 TB when using 4 TB SATA disks. If switching 4 TB SATA disk with 6 TB SATA disk, a 12-drive storage node will then have 12*6 = 72 TB. If increasing to 36 drives, a fully-loaded node with 6 TB SATA disks will have a raw capacity of 36*6 = 216 TB.

Besides scale-out for increased capacity, performance and file retrieval efficiency, our solution enables existing systems to scale-up. Administrators can choose to replace any parts of the hardware in the system, including server CPU, memory, motherboard, hard disk, and network equipment.





- **Dynamic load balance**

On existing hardware, with load detection on the network and server disks, metadata will evenly distribute data workload within the system. Such balanced distribution of tasks can ensure the highest efficiency possible on any given hardware.

With traditional storage, when adding new hardware, new unbalanced disk utilization will be a performance bottleneck and a management burden for administrators.

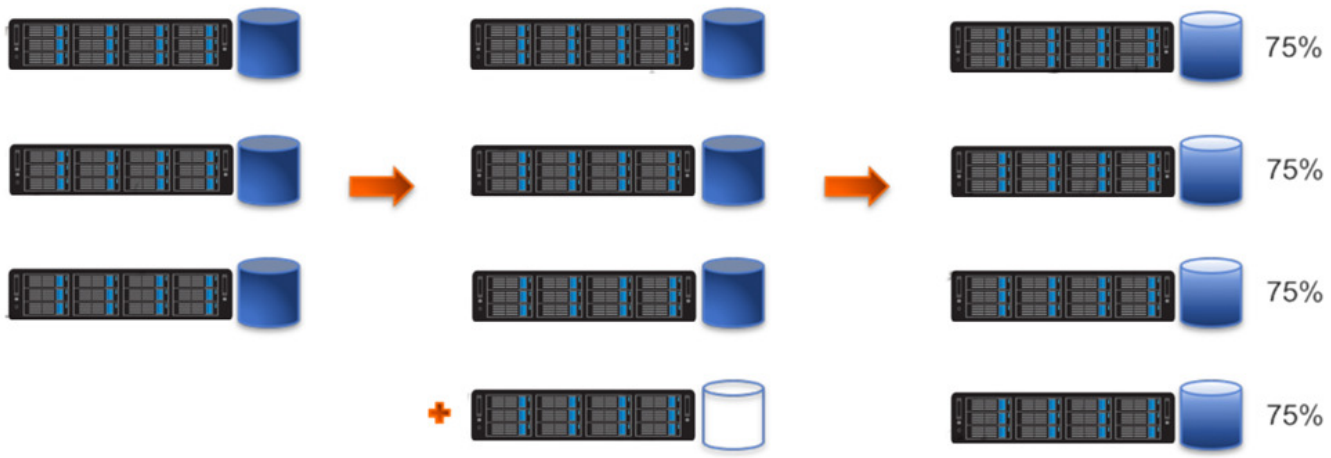
SANCluster **InfinityScale** Storage provides customers with the ability to expand storage capacity dynamically and utilize all the storage devices evenly. When adding new disks or storage nodes, there is no need for the administrators to manually copy data in order to get balanced disk utilization or specify how much load each storage server should handle.

Everything with respect to load balancing is done with a simple switch command. Administrators can schedule a time for load balancing to take place when systems are not busy.

Once the load balance command is initiated, it can be stopped at any time as needed.

When the switch is turned on again, the system will continue the load balancing process from the previous stopping point. The load balancing process automatically redistributes data across all storage nodes until disk utilization on each storage node is relatively equal.

Thereby our system can utilize all the possible resources to provide high aggregated throughput and deliver near- linear increases in performance when additional hardware is added.



Load Balancing on Demand

- **Access control**

  SANCluster **InfinityScale** Storage has built-in, byte-level granularity of the locking mechanism, to ensure data consistency and usage authority. Our system deploys two levels of access control: one from operating systems such as Linux using UNIX and Windows using ACL, the other being an enhanced one, managed by metadata servers.The enhanced control from our system covers the permissions to read, write, delete, rename, list, and additional write. The added permissions are managed and carried in the storage operation side. Control setting from the operating system will not change or override these permissions, not even with super users. The enhanced control is set at the application side on computing nodes and can be applied to any levels of the directories within the storage, even with fine-grained permission split. Without proper permissions, users including even root users will not be able to access or make changes to the protected directories.

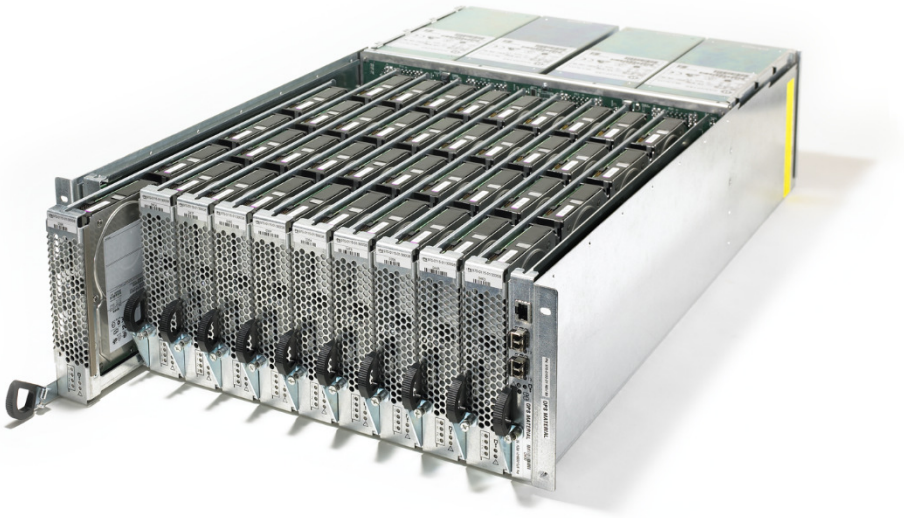
  Any changes to a file such as data creation, deletion or rename will be recorded by the system and can be searched anytime by file name.

- **Network reconfiguration**

  When there is a network glitch or during recovery from an interruption, system will automatically reconnect to the network switch.

- **Collaborative management**

  An arbitration mechanism involving all the nodes of the system. When there is a hardware failure, all the nodes will participate to determine what the damage is and where the damage has occurred. Such collaborative effort will prevent arbitration misjudgment.



Scale out Hardware

➢ System Core

- Performance

It is our goal to provide customers with the best performance-to-cost systems, leveraging only commodity hardware with no vendor lock-in and not to rely on expensive hardware, such as high-end servers or SSDs or Flash disks for high performance. On average, each drive of a storage node in our system can deliver 50 MBps to 60 MBps. While using all SATA disks, on a storage node with 12 drives, aggregated performance is: 12*50 = 600 MBps per node, it can be 720 MBps on the high end. On a system with three 12-drive storage nodes, our solution can offer an aggregated performance of 3*600 MBps = 1.8 GBps. On a system with five 12-drive storage nodes, the aggregated performance is 5*600 MBps = 3.0 GBps.

If increased to 36 drives, the aggregated performance will be: 36*50 = 1.8 GBps per storage node. Hence, a system with five 36-drive storage nodes will reach aggregated performance of 5*1.8 = 9 GBps, and it does not matter what size SATA disks the system uses, supporting 1 TB to 6 TB.

Using SSD disks with dual 40 GbE network cards or 100 Gb Infiniband card, A cluster of 4 of such storage nodes can provide 192 TB in capacity and more than 156 GB/s write performance.

- **Aggregated bandwidth and concurrency control**

  By striping files across multiple storage nodes, our system enables read/write operations from/to multiple storage nodes. This is done in a parallel process with the utilization of aggregated bandwidth of all available storage nodes in the system. When an application reads one file, it accesses different stripes of that file concurrently from multiple storage nodes thereby improving read performance. The parallel read/write process eliminates the bottleneck formed by single data path in traditional systems, and meets the bandwidth requirement of concurrent access from multiple application servers.

- **Metadata management**

  In a typical system, there is a least one pair of metadata servers. Both servers are active with information being fully backed up by each other. One metadata server can efficiently handle up to 20,000 files per second (complete fopen operations). To increase file lookup performance, simply expand the cluster to the desired level, up to 128 servers.

- **Index management**

  Our system provides global namespace, which allows all application servers to access the same directories and folders. Traditional solutions can only host a limited number of files under a single namespace or a single directory. When the number of files in a traditional system reaches a certain limit, the system performance will drop dramatically.

  SANCluster **InfinityScale** directory support can break the limit on the number of files under a single directory. It outperforms other competing file systems, especially when dealing with a massive amount of small-size files. In one of our production clusters, our solution manages over 40 billion files, the majority of which are KB sized files.

➢ Data security

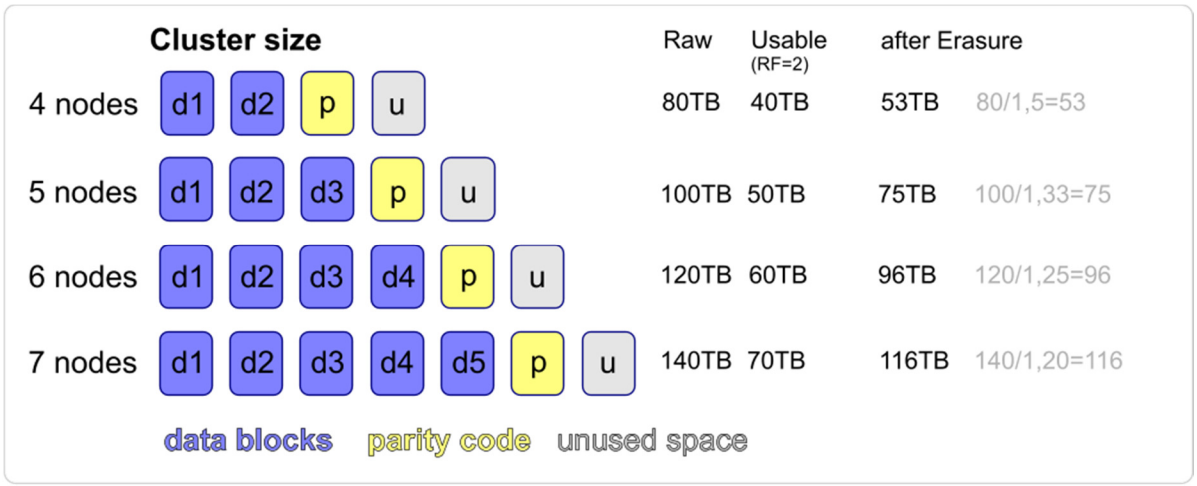
The SANCluster **InfinityScale** Storage solution has built-in data encryption and a block-based algorithm for data assembling. Depending on data types, the system will automatically slice up the data and encrypt it with the client-side software. Only sliced and encrypted data will go through the system network and be stored in the storage server cluster. Data assembly and decryption can only be done by application servers with the uniquely assigned clients, which are controlled by the drivers of the client-side software. No other application servers or computing nodes can decrypt or assemble the data through network interception or stolen hardware.

Our system has no single point of failure. The built-in automated self-monitoring mechanism can single out inactive hardware, either at the disk or server level, to provide high data availability, eliminating system downtime. Many systems of our PB level customers have been running without downtime for over 5 years.

➢ Data (file) copies

Redundancy is an efficient way to protect data, but it will take up storage capacity. With SANCluster **InfinityScale** Storage, users can select any number of file replicas based on directory importance. While each of the files in one important directory has 3 copies, each of the files in a less important directory can have 2 copies. By replicating data across multiple storage nodes, the system will prevent downtime from failures of disks, servers or network. In addition, availability of multiple file copies can enhance the performance of concurrent usage.

When storing a file or its copy (copies), the file or its copy (copies) will be sliced or grouped up into different segments. Each of the segments or copies of the segments will be stored in different storage nodes in the storage server cluster to ensure protection in case of disk or server failure. The more copies of the stored files, the higher hardware tolerance the system can offer, but capacity utilization will be reduced according.



File allocation different methods

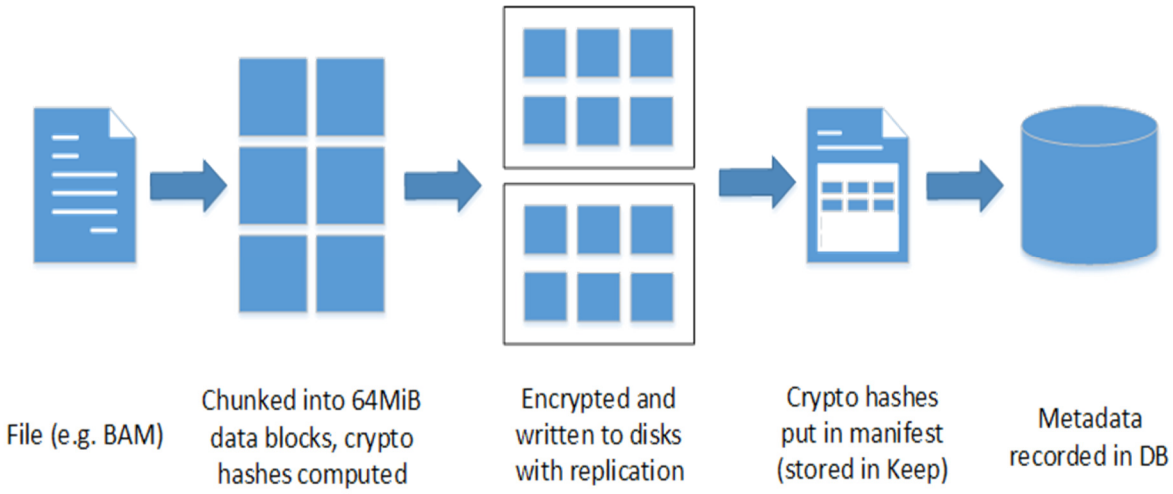
File-level RAID

According to a built-in data-granularity setup, the system will allow using either 2+1 or 4+1 protection methods on file storage. While using 2+1, the data will be sliced up to two main files plus a parity file. Each of the three files will be stored in three different storage nodes in the system. In case any of the two main files is not usable, the system will use the other main file and the parity file to replicate the storing data thereby ensuring data availability.

While using 4+1, the data will be sliced up into four main files plus a parity file. Each of the five files will be stored in five different storage nodes in the system. In case any of the four main files is not usable, the system will use the other three main files and the parity file to replicate the stored data thereby ensuring data availability.

Byte lock

The system provides access control with byte-level granularity to ensure data consistency and usage authority.



File Level RAID and data replacement

➢ Hardware failure tolerance

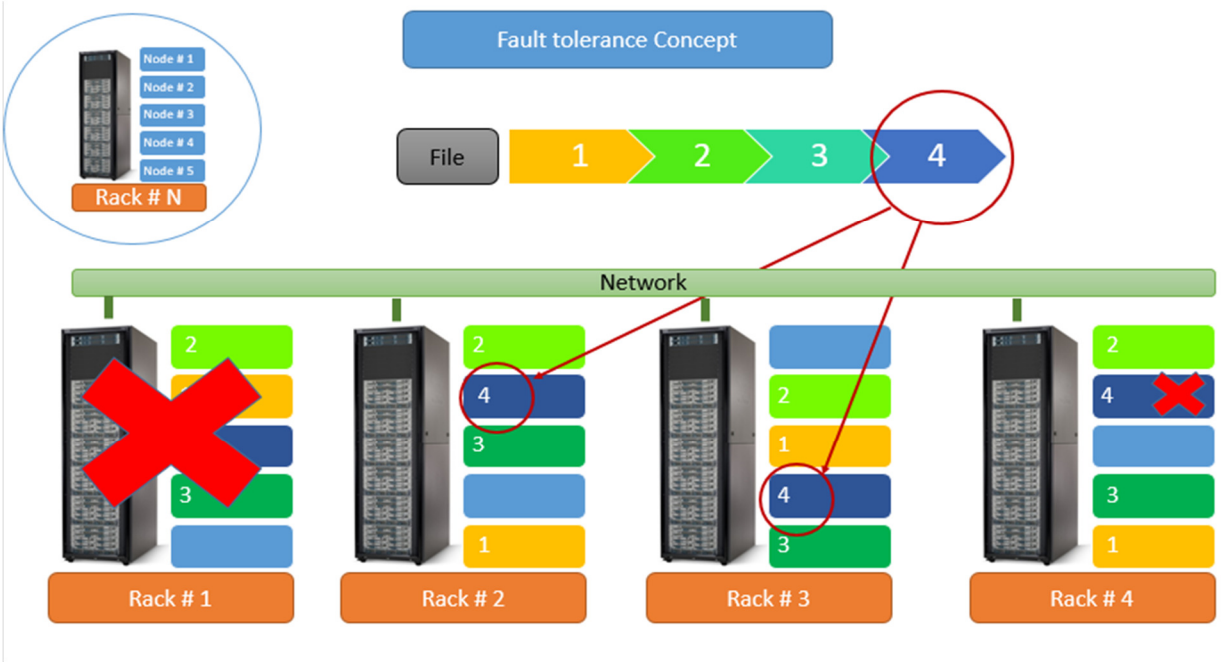
It depends on the size of the cluster and what method customer use on data protection with redundancy. Usually a cluster can sustain normal operation with one storage server failure. Therefore, if a cluster is equipped with 16-drive nodes, a total of 16 disks can fail without cluster downtime. Similarly, with 36-drive nodes, the cluster will still be up even when 36 disks fail.

We reduce system downtime risk by using a Self-Healing / Self-Monitoring mechanism. The build-in feature detects drive performance periodically, and if the performance does not match what is expected, the system will mark the drive to be read only, will move data out, and ask for new drives for replacement. This is preemptive protection.

When using replica method, one file can be replicated up to 4 replicas. When having 2 replicas, system will still be up with 1 server failure in the cluster. When having 3 replicas, system will still be up with 2 server failure in the cluster. When having 4 replicas, system will still be up with 3 server failure in the cluster of minimum of 5 storage servers.

Besides the replica method, customer can chose our File-level RAID to protect the cluster for better capacity utilization. When using 2+1 File-level RAID, capacity utilization can be up to 67%. When using 4+1 File-level RAID, capacity utilization can be up to 80%. In either case, the cluster can sustain normal operation with 1 server failure.

In addition, our system offers fast rebuild times; much faster than any RAID controller. The failing drive's data will be moved to a safe location before the next failure. Separately, replication protection provides a complete copy of data and this is used to provide additional protection. Our system supports up to 4



Fast data recovery

Our special File-level RAID enables up to 80% capacity utilization and fast data recovery in the case of hardware failure. When there is a hardware failure, because data have file copies across different storage nodes, only the unavailable data files will be replicated in the recovering process to enable data availability. This is different from the traditional SAN or NAS solutions where entire data set on failed hardware needs to be recovered for data availability.

The whole recovering process is transparent to applications and will not interrupt the user operations. Moreover, the recovery itself is a parallel process, so all the storage nodes in the system will participate. Typical data recovery time is only 1/5 the time required with traditional RAID. The larger the cluster, the faster the system can recover from failed hardware. In a 500 TB system, recovering data from a failed disk of 3 TB will only take about 30 minutes.

POSIX module interface

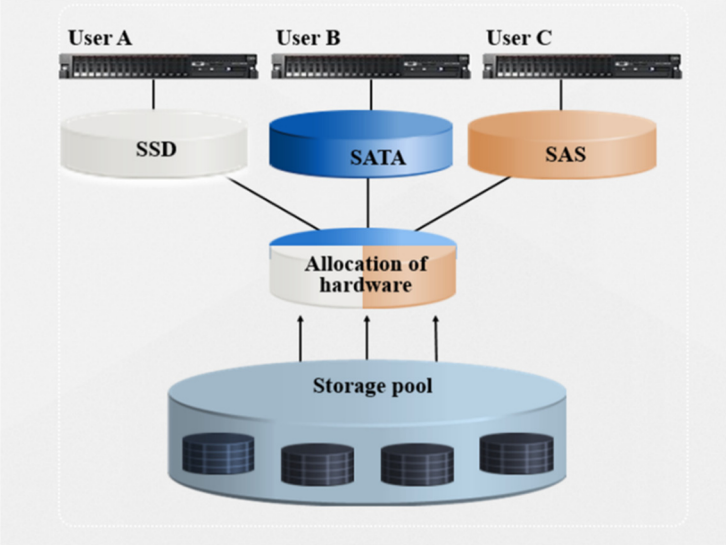
SANCluster Storage provides POSIX interface and multi-protocol support including NFS, CIFS, and iSCSI. Our system can easily work with virtualization platforms such as VMware, Xen, and KVM.

Data base support (structured data)

SANCluster **InfinityScale** Storage supports structured data at the block level. Our system can fulfill requirements of Oracle RAC, and MySQL. In a testing case of running an Oracle database, when having the same number of disks, the performance of our system with inexpensive commodity hardware is similar to EMC VNX Series.

Chapter 4: Hierarchical Storage Management (HSM)

Hierarchical storage management (HSM) allows automated or semi-automated movement of data to multiple tiers of storage. The mechanism of SANCluster **InfinityScale** HSM is based on the concept of an added volume in clustered storage, which allows different disk groups be assigned to different data directories. After grouping a variety of hardware with data directories, different applications can be assigned to various hardware types to derive the best access efficiency, with the support of hot and cold data migration.



According to different disk performance in the storage node, our system can consolidate the same performance ones into various disk groups, which will then be assigned to support specific data directories. The system allows authorized users to make changes to the disk-group configuration through a built-in graphical interface for dynamic disk expansion or deletion. This function is applicable to all disks in the system across all storage nodes.

As an add-on feature in HSM, SANCluster Storage can consolidate all the disks from storage nodes into one large virtual drive. During such virtualization process, the system will assign a unique identity to each of the disks in all the storage nodes. Disk identities will be stored in the metadata server and when there is a data storage need, according to data size, file granularity and disk load, metadata will instruct the system using the least-loaded disks for the task. Thereby our solution can automatically provide the optimal usage of available disks in the system.