White Paper

February 2021

U.S. Service Partner for High-Performance Computing
2160 Kingston Court, Suite B
Marietta, GA 30068
Email Don Bailey at dbailey@dstonline.com

DATA IN SCIENCE TECHNOLOGIES DST

1. Introduction

LeoFS is market proven parallel file system for different types of I/O intensive workloads. It was developed by a team of high-performance computing specialists, with a simple goal – ***To provide the best combined values in data storage solutions***:

- High performance – delivers data from storage nodes in parallel to client applications
- Large and scalable capacity – leverages scale-out architecture to increase capacity
- Reliable data protection – uses file-level software RAID with N+M erasure coding
- Professional support – offers 24/7 system management with designated consultants
- Affordable cost – guarantees the best price/performance ratio

LeoFS comes with native clients for Linux, Windows and macOS, all kernel modules that do not require any patches. Using only commodity hardware, the software-defined storage offers both cluster and single server solutions.

With LeoFS, data files are transparently distributed over multiple nodes. By simply increase the number of servers and disks in the cluster, you can seamlessly scale the file system's throughput and capacity to the needed level, all being aggregated in a single namespace.

With only dual 10GbE network and inexpensive hardware, current largest cluster installed has over 300 storage nodes, 95PB and more than 200GB/s I/O throughput.

LeoFS is fully POSIX-Compliant and compatible with all software applications, x86 based servers and IP networks. The file system supports Linux kernels up to the latest version and Linux distributions including Debian/Ubuntu, SLES/OpenSuse, or RHEL/Fedora. Some advantages include:

- ✓ With POSIX interface
  - ➢ No need to modify or rewrite applications
  - ➢ Add clients and servers without downtime
- ✓ Clients and storage servers communication via
  - ➢ TCP/IP based connection
  - ➢ RDMA-capable networks
    - InfiniBand (IB)
    - Omni-Path (OPA)
    - RDMA over Converged Ethernet (RoCE)
- ✓ All-in-One solution for object, block and file-based storage

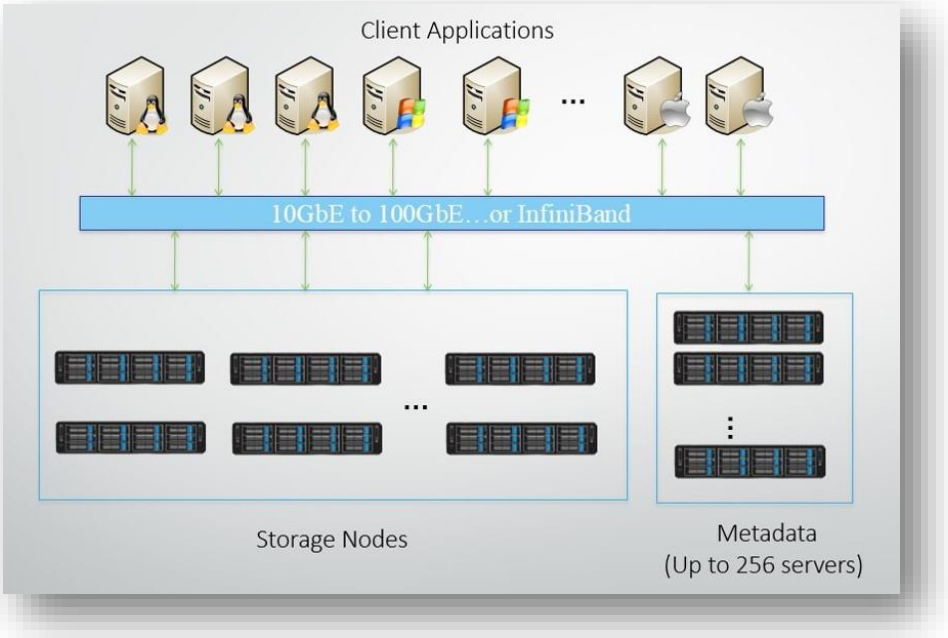
U.S. Service Partner for High-Performance Computing

2160 Kingston Court, Suite B
Marietta, GA 30068
Email Don Bailey at: dbailey@dstonline.com

2. System Architecture

LeoFS architecture has three key components:

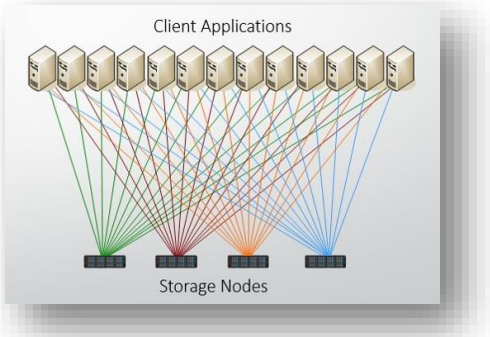
- Client applications: enable application/computing nodes to access the stored data
- Metadata: manage file information, striping, path and access permissions
- Storage nodes: stores user's distributed data contents



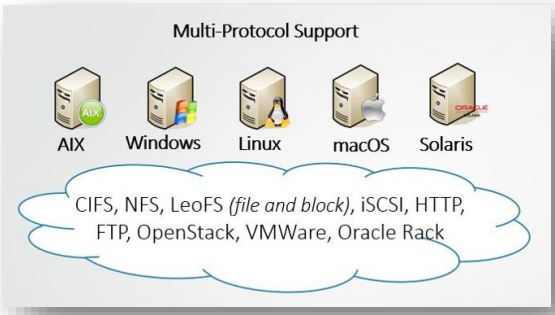
LeoFS was written to break away limitations from traditional storage solutions. It does not include concepts from legacy algorithms or other open source coding. The file system provides high performance for all workloads including big and small files, intensive reads and writes (random or sequential) and metadata heavy. Below is single cluster threshold.

	Theoretical	Actual Deployment
Storage nodes	4,096	333
Metadata servers	256	32
System capacity	EB	95PB
Number of files	Unlimited	50 Billion

Without Controllers nor Gateways
LeoFS enables concurrent access between all client applications and storage nodes in the cluster. Thereby, eliminating I/O throughput bottleneck.



Multi-Protocol Support
The system supports Hadoop, Oracle/SQL, VMware, KVM, and Xen. LeoFS works with LeoSAN, CIFS, NFS, HTTP, S3, Swift, iSCSI, HDFS, and Cinder.



2.1 Client Applications

Native clients
LeoFS provides native clients that can be installed on Linux, Windows and macOS hosts. All kernel modules that do not require any patches:

- Linux: all versions from kernel 2.6 and up
- Windows: XP, Windows Server 2003 and up
- macOS: 10.5 and up

When a client is loaded, it will mount the file system defined in LeoFS mount configuration file instead of the usual Linux approach based on /etc/fstab. Such approach makes the starting of LeoFS client like any other Linux service through a service start scrip. It makes client upgrades much more convenient.

Access also possible through NFS, CIFS, HTTP, FTP, iSCSI or from Hadoop. For example, you can re-export LeoFS mount point through NFSv4 or Samba or to use LeoFS as a drop in replacement for Hadoop's HDFS.

To achieve maximum performance, clients should be used on all hosts.
The largest number of clients currently installed is over 10,000.

2.2 Metadata

The metadata stores information about the file, such as directory information, file and directory ownership and the location of user file contents on storage nodes. It provides information about the location ("stripe pattern") for an individual user file to a client when the client opens the file, but afterwards the metadata is not involved in data access (file read and write) until the file is closed.

LeoFS metadata comes in pair and was designed to be scalable in the initial software development stage. One metadata pair with 960GB usable capacity is good for 2 billion files. As standalone metadata server, key configurations include:

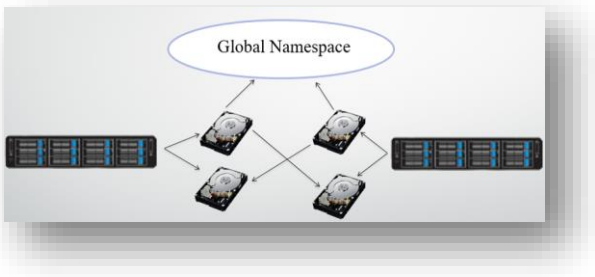
- CPU: two Intel E5-2620 v4 (8 cores each)
- Network: dual 10G GbE
- 64GB RAM, one SSD OS drive
- Two 480GB enterprise SSD

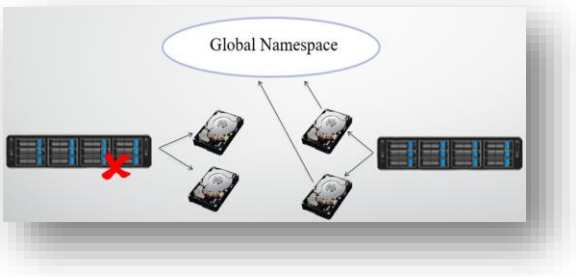
Having faster CPU cores will improve system latency and overall I/O throughput. The performance of LeoFS metadata can be improved quosi linearly when more pairs are put into use (up to 128 pairs). The metadata cluster can easily fulfill high performance requirements on extreme IOPS or file open rate.

*In a testbed with 8 storage servers (each with 24*2TB 7200r SATA), 160 clients and dual 10GbE network, one metadata pair delivers 20,000 creates per second on 8 million files (4KB each).*

Active-Active
One metadata pair requires four SSD drives. Each metadata in the pair has two drives: one is active for managing its exclusive fraction of the global namespace, while the other one backs up its pairing metadata's active drive.

Automatic Failover
When a metadata failed, the pairing metadata will take over automatically. Hence, no service interruption or operation downtime may occur. This eliminates single point of failure from metadata management.





Option to be embedded with storage nodes
While standalone metadata servers can provide the highest level of throughput, to leverage cluster hardware, metadata can be put with storage in the same physical server. For example, using 4U 24-

bay server, in a typical LeoFS system with dual 10GbE network, key configurations of such metadata plus storage node are:

- CPU: two Intel E5-2620 v4 (8 cores each)
- Metadata: two 480GB SSD
- 64GB RAM, one SSD OS drive
- Storage: twenty-two SATA HDD

For large systems, we recommend running metadata service on standalone servers, which allows independent scaling of metadata performance and capacity. Dedicated machines will cost more but ensure the optimal levels of system IOPS and file open rate.

2.3 Storage Nodes

The storage nodes in LeoFS manage the service to store striped file contents, also known as data chunk files. Using x86 commodity servers, option to choose cluster or single server solution:

- ✓ Storage server with 12-bay, 16-bay, 24-bay, or 36-bay drives
- ✓ Cluster starts with 3 nodes, and up to thousands

Like metadata management, our storage node is based on a scale-out design. You can have one or multiple storage nodes per LeoFS file system. A storage node can have one or multiple storage targets. A storage target typically is a HDD that formatted with ext4 local file system. Each additional storage target or storage node adds both capacity and performance to the file system.

The storage service works with any local Linux POSIX file system. To distribute the used space and to aggregate the throughput of multiple nodes, LeoFS uses striping, which means data file gets split up into chunks of fixed size and those chunks are distributed across multiple storage targets in multiple nodes.

The chunk size and number of targets per file is decided by the responsible metadata service when a file gets created. This information is called the stripe pattern. The stripe pattern can be configured per directory or for individual files (e.g. by using the LeoFS layout command line tool or on the system's web-based management page).

When 4+1 file-level RAID is being used, a file chunk will be sliced into four pieces, and one parity piece will be generated. All these five data pieces will be written to five different HDDs on different storage nodes. Hence, performance won't be limited by a single storage target.

LeoFS system is especially good in supporting high concurrent read/write.

Computational Storage

Besides embedding metadata management, LeoFS storage node can also run client applications so

the server become shared storage plus data processing unit.

After installment of the LeoFS client and mounting with the file system, a storage node can support client applications while providing storage service. Since some data files are directly accessed with the local storage targets on the computational storage node, overall system performance will be improved while network bandwidth is better utilized.

For a large cluster, computational storage is a great leverage of available hardware, providing significant cost savings as there is no need for separate application/computing nodes.

3. Best Combined Values

LeoFS system offers the best combined values in data storage. Whether your goal is to increase productivity or have a better ROI, we guarantee usage satisfaction on all products.

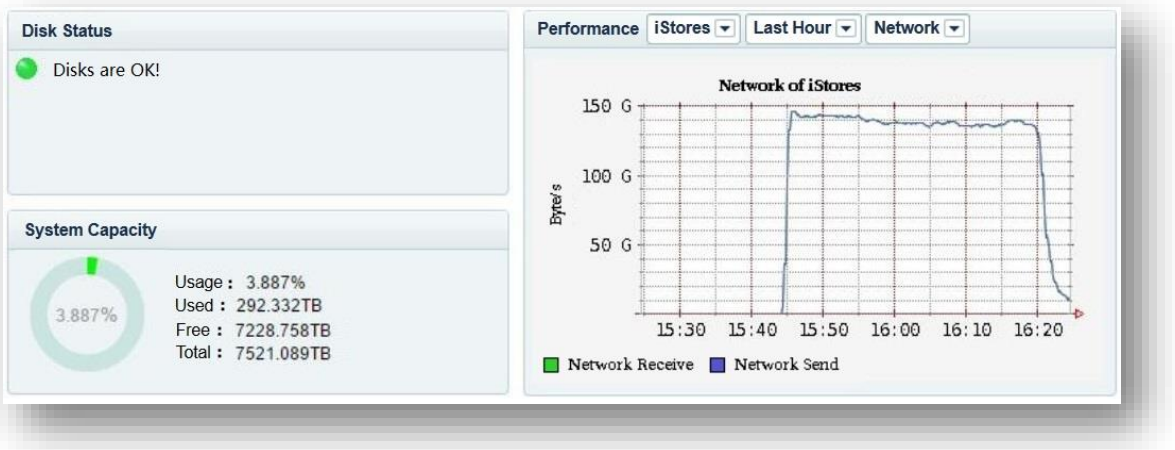
With more than 700 customers worldwide, we take pride in the fact that most of our customers in PB usage started with only a few hundreds of TB. Key industry success include:

- Scientific Computing
  - Genomics
  - Cryo-electron Microscopy
  - Satellite Imaginary and Observatory
  - Geographical Data and Mapping
  - Meteorology/Climate
- Oil and Gas
- Higher Education
- Media and Entertainment
- Telecom and Internet
- AI and Big Data
- Video Surveillance

3.1 High Performance

LeoFS delivers file data from storage nodes in parallel to client applications. Data flows between storage nodes and client applications without any intermediate controllers or gateways. Since all client applications have direct connections to all storage devices in the system, bandwidth is not just aggregated but multiplied. Unlike competition must use high-end hardware to boost performance, with only x86 commodity servers and in-expensive SATA or SAS HDDs, LeoFS system provides predictable and sustainable throughput without deterioration overtime.

Customer on-site 7.5PB, near 150GB/s throughput
(102 4U 24-bay storage nodes, dual 10GbE network, 4TB SATA)



U.S. Service Partner for High-Performance Computing
2160 Kingston Court, Suite B
Marietta, GA 30068
Email Don Bailey at: dbailey@dstonline.com

Always saturate what system hardware can offer
Our scale-out architecture increases storage capacity and system bandwidth incrementally as you add more hardware. The linear increase in performance with additional hardware applies to both metadata management and storage service.

File Fragmentation
Overtime with any data storage systems, file fragmentation will occur when groups of files that are scattered throughout a hard drive platter instead of one continuous location. Fragmentation is caused when information is deleted from a hard drive and small gaps are left behind.

LeoFS provides a file layout type that merges all the small files within a directory and write to storage targets as a big file. With this method, small gaps that generated by storing small files which size are not multiple of 4KB are avoided so that a considerable space is saved overtime. The space will not be released unless all of the file within the directory have been removed.

Sizes of Files
When a file is written to a file system, it may consume slightly more disk space than the file requires. This is because the file system rounds the size up to include any unused space left over in the last disk sector used by the file. A sector is the smallest amount of space addressable by the file system. The wasted space is called slack space. Although smaller sector sizes allow for denser use of disk space, they decrease the operational efficiency of the file system.

The maximum file size a file system supports depends not only on the capacity of the file system, but also on the number of bits reserved for the storage of file size information. LeoFS is proven in handling different file sizes with no performance deterioration.

3.2 Large and Scalable Capacity

Raw capacity of single server starts with 48TB (12-bay using 4TB HDD) and up to 360TB (36-bay using 10TB HDD). Cluster solution starts with 144TB (3 nodes of 12-bay using 4TB HDD) and up to EB. On the largest system currently installed with 95PB, there are:

- 32 metadata plus storage nodes, each with 34*8TB HDDs
- 301 computational storage nodes, each with 36*8TB HDDs

LeoFS 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 need to reboot.

After hardware is added or replaced, through a load balance switch, LeoFS will automatically rebalance capacity across the system. Each added storage unit will immediately increase more performance and capacity. Metadata service can be expanded in the same way. Because there is no single point of failure, you can choose to replace any parts of the hardware in the system, including server CPU, memory, motherboard, hard disk, and network equipment.

3.3 Reliable Data Protection

LeoFS system's self-monitoring mechanism can single out failed hardware automatically, either at the disk or server level. It constantly scan all files in the background. Once inactive hardware is detected, it will be isolated for read-only operation or taken out. The system will then start a self-healing process without any operation interruption. It will automatically reconstruct the full levels of erasure-coded data protection for all files.

File-level Software RAID (N+M Erasure Coding)

Different from traditional hardware RAID, LeoFS uses file-level software RAID with N+M erasure coding. Depending on selected erasure codes, data are stripped into different segments with parities. LeoFS ensures any given segment or parity is stored on different storage node from the rest, or in single product, on different drive in the server. For example, when N+2 erasure code is applied, while the total number of drives in a single server or the total number of nodes in a cluster is more than N+2, then the system can sustain operation with up to two simultaneous failures, whether it is an individual drive or a whole node.

While traditional hardware or software RAID needs to rebuild a whole physical drive, LeoFS solution rebuilds only the files that are affected, and it uses the entire cluster to rebuild. Hence, it delivers much faster data recovery, usually in a fraction of the time traditional RAID architectures require. No downtime or reboot, recovery of one TB data usually takes less than 20 minutes.

LeoFS offers much greater capacity utilization than buddy mirroring or hardware RAID. No waste of resources as there is no spare drives needed for data recovery. By directory base, the system offers optimum data protection plans for different files. With 16+1, capacity utilization is over 90%.

No Single Point of Failure

As metadata always come in pair, when a metadata server or an active metadata drive fails, the other metadata server or failed drive's backup in the pair will become active and take over the metadata management.

When a storage target or node fails, the file system will detect it and start data recovery automatically. All available storage targets and nodes in the system will get involved, fulfilling the requirements set by our file-level RAID policy.

All subnets that configured in LeoFS system are active. If one subnet fails, all the data request will be shifted to other active ones.

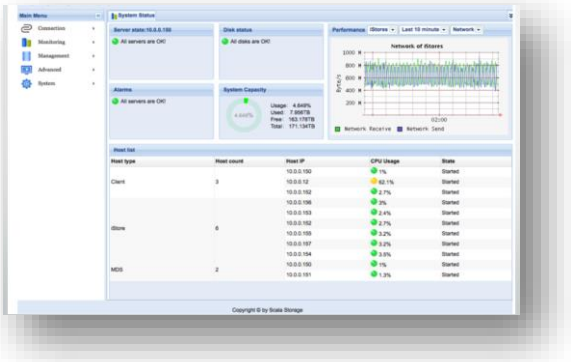
3.4 Professional Support

Peace of Mind
With direct contact to the file system developers, LeoFS offers easy management with designated consultants. Our highly-trained and well experienced engineering team is available 24/7. No matter how large is the cluster, our staff always take customer system monitoring a top priority.

Uncompromised Reliability
LeoFS uses a single management GUI to configure, manage and monitor the storage system. When hardware is ready, software installation and update can be done in one hour. Adding servers for performance or capacity increase requires no downtime nor the need to reboot. Even with our large deployments, all data are within a single namespace. Through the centralized management with graphical monitoring and administration, most cluster problems can be fixed remotely without operation interruption.

The Web-based GUI Interface
Daily performance and statistics are recorded and monitored in a dashboard:

- ✓ System health information can be exported to different formats for review.
- ✓ When a failure occurs, system will go into self-healing process automatically with alert sending at the same time.



Options to choose Next Business Day Service Level Agreement, Re-mote or On-Site Support Warranty, Advanced Hardware Replacement.

- ❖ Cluster Monitoring: Free support access via emails, phone and live chat. Our consultants can remotely access the system and run diagnostics to ensure cluster condition. On-site support is also available to keep customer business seamless.

- ❖ Software Maintenance and Update: Once installed, enjoy free software upgrades and access to a vast suite of enterprise features, such as snapshot, clone and WORM.

- ❖ High Quality Hardware: All hardware including replacements must go through pre-

configuration testing so installation and repair will done in a time sensitive manner.

3.5 Affordable cost with Software as a Service

Annual service fee

- No capacity limit, $5,000/storage node, any application servers and any metadata servers
- Current special: 1 PB free usage if replacing competitive file systems such as Lustre, GPFS, or BeeGFS

Option to choose cluster or single server products

- Cluster: Starts with 2 nodes, and up to thousands
  - CPU: Intel Xeon E5-2630 V4 * 2
  - Mother Board: Supermicro X10DRL-i
  - HBA: LSI SAS 9300-8I
  - System Disk: 480GB SSD * 2 + 240GB SSD * 2
  - Storage Disk: 2TB to 10TB HDD, SATA or SAS
  - RAM: 128GB
  - Network Port: 2 * 10 GbE or 40 GbE or Infiniband
- Single Server: from 12-bay to 36-bay drives
  - CPU: Intel Xeon E5-2620 V4 * 2
  - Mother Board: Supermicro X10DRL-i
  - HBA: LSI SAS 9300-8I
  - System Disk: 480GB SSD * 2 + 240GB SSD * 2
  - Storage Disk: 2TB to 10TB HDD, SATA or SAS
  - RAM: 64GB
  - Network Port: 2 * 1GbE or 10 GbE

LeoFS solution guarantees the best price/performance ratio in the industry. After the initial purchasing term, customers have the perpetual right to use our software. For optional maintenance and updates, the annual service cost is only 15% of the original software cost.

Using only commodity hardware, our system offers the best cost-of-ownership options. For higher capacity, customers can choose 4U 36-bay servers with 10TB HDDs. If I/O throughput is the top consideration, 4U 24-bay nodes with 4TB HDDs are recommended.

4. Key Features

Cluster or single server product, LeoFS is ideal for High-performance Computing, AI, and Big Data Analytics. Selected benefits include:

- Dynamic load balance: hardware evenly share system workload
- Runs on platforms such as x86, OpenPOWER, ARM, and Xeon Phi
- Re-export through Samba, NFS, FTP, HTTP, LeoSAN or iSCSI
- Support for group/user ACLs and quota
- Fully active network with automatic failure detection
- Supports Infiniband, GigE, multiple subnet and bonding
- Cold data sanity check, automatic repair, no downtime
- WORM directory, avoid modification of saved data

Data Security

LeoFS 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.

Snapshot and Clone

LeoFS has snapshot and clone capabilities for datasets, single file, or file-based block devices in the file system. It uses redirect-on-write as snapshot/clone creation method. Snapshots and clones won't take any storage space if no changes made on the snapshot source or all clone entities, they share as most data blocks as possible with the source entities. If a block needs modification, the storage system merely redirects the pointer for that block to another block and writes the data there. LeoFS snapshot and clone can be easily created and maintained on the web based management system.

WORM

WORM directory is a filesystem function that can be used if modifications or deletions of saved data are not allowed. You can create or manage a WORM directory on the web based management system. Files in a WORM directory can be modified or removed within a period of time from their creation. When the file enters a protected period, during which the file can't be modified or removed. After the protection time, the file can be removed but not modified.

Access Control

To ensure data consistency and usage authority, LeoFS provides built-in byte-level granularity of a locking mechanism. Our system deploys two levels of access control:

- One from operating systems
  - Linux using Unix
  - Windows using ACL
- An enhanced one managed by metadata servers

The enhanced control covers the permissions to read, write, delete, rename, ln, 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.

The enhanced control is set at the client application side and can be applied to any levels of the directories within the storage, even with fine-grained permission split. Without proper permissions, not even root users will 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.

Data Base (Structured Data) Support

LeoSAN is file-based block device solution we have developed. The client server using LeoSAN is configured as LeoFS client and mount with the storage system. A LeoSAN block device can be created and mapped to client server on web-based management system. The block device is displayed with command fdisk -l. You can create partition on it, format it and mount it as a usual block device.

LeoSAN fulfills 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 VNX5100. The testing report is available upon request.

Virtualization

LeoFS works with KVM seamlessly on Linux platform. You can create virtual machines on the local mounted LeoFS storage. LeoSAN can also be attached to the virtual machines as local storage. Any data operations on the virtual machines are all directly done locally. This provides a client-like high performance for KVM usage.

With VMware and Xen, as clients can't be installed, LeoFS provides iSCSI target to work with the virtualization systems. Although these types of virtualization platforms won't achieve client-like high performance, they will have other benefits using LeoFS storage, such as high data availability through file-level software RAID and automatic system self-monitoring and healing.

Dynamic Load Balance and Switch

After adding new hardware or making a replacement, traditional storage is difficult to balance data workload automatically. This is a management burden for administrators, and if not dealt with, most likely it will cause hot spot problems and performance bottleneck.

Through metadata management service, LeoFS provides the ability to expand or replace storage capacity dynamically. It will evenly distribute data workload in the system. The file system ensures each storage node in the cluster and each disk in the node shares a close-to-equal amount of workload (within 5% differentiation rate). Such balanced distribution of data and workload can provide the highest efficiency possible on any given hardware. It ensures the linear increase in system performance when new hardware is added.

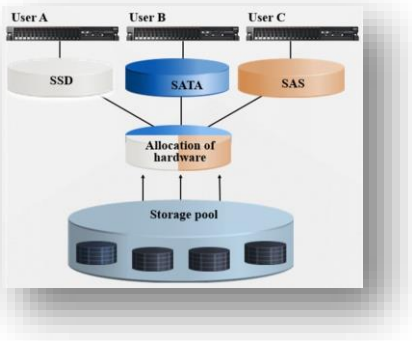
To avoid operation interruption, load balancing is done with a simple command switch. It can be turned on or off at any given time. Administrators can schedule load balance to take place when operations are not busy. When the switch is turned on, the system will always continue the load balancing process from the previous stopping point.

Hierarchical Storage Management (HSM)

LeoFS allows automated or semi-automated movement of data to multiple tiers of storage. The mechanism of our 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.

Our system 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.



U.S. Service Partner for High-Performance Computing

2160 Kingston Court, Suite B
Marietta, GA 30068
Email Don Bailey at: dbailey@dstonline.com

Cache Management
Built-in algorithm to identify and place frequently accessed data in cache. In our system, cache can be created in all the nodes, including client applications, storage nodes and metadata servers.

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

Capacity Constraint
While there is no limitation on the number of files under certain directory, LeoFS system can set capacity constraint based on directories. Once reaching the upper limit of capacity, no additional data can be written onto the directory.

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.

Geo-replication
Under the broader LeoFS software suite, we have developed LeoSync, which provides easy geo-replication of any given data among different data centers. Key values of include:

- ✓ Data transfer among multiple locations, commends such as: Start, Stop, Pause and Resume
- ✓ Three synchronization thread pools: traversing, metadata and data
- ✓ Multiple synchronization modes: manual, scheduled or periodic synchronization
- ✓ Efficient handling of either large size or small size files
- ✓ Full set of data replication or incremental data synchronization
- ✓ Automatic bandwidth checkup to avoid excessive usage of network resource
- ✓ Error handling options: either skip and continue or stop the task
- ✓ Breakpoint continuation

5. Sample Case – University Research

Background and Purpose

The researchers at one university have no central storage cluster. Some research data is stored in HPC storage system, some are being stored in personal desktops/laptops/personal servers/department servers, etc.

The proposed solution for a central storage pool will have to provide a central storage system that is accessible from within university network. This system will ensure high-availability data access, and will scale up as demand grow without performance penalties. The research computing group intends to serve approximately 3,000 researchers with this system.

The goal is to select a system that would provide best price/performance ratio to the research community and must be expandable and flexible.

Storage System Deliverables

The university requires a storage system to provide 500TB of distributed and parallel storage solution, which must be stable to operate and sustainable for a large-scale deployment, up to 10PB.

Storage system must have SFTP, NFS, SMB or its own proprietary export services for clients access from Windows, macOS and Linux operating systems. Storage system native client software (to access storage system from Windows, macOS) is highly desired.

Storage system must support quotas at user level, group level and file-set level. Quota management must be done natively by the storage management software, not by any underlying *NIX operating systems.

System software features expected

- Virtual disk pools
- Thin provisioning
- Rapid drive rebuilds
- Degraded disk detection
- Automated storage provisioning tool
- Centralized management interface
- Disk background scrub
- Drive spin down
- Volume copy snapshot
- Clone snapshot
- File level N+M redundancy data protection
- Automatic load balance

Storage system performance should not be impacted with horizontal capacity expansions. As for read/write throughput - must support approximately 3,000 simultaneous logins, and at least 500 active SMB/SFTP sessions with the following characteristics for the whole system:

- ✓ Single client steaming read 2 GB/s
- ✓ Single client streaming write 2 GB/s
- ✓ Multi-client streaming reads (different files) 10 GB/s
- ✓ Multi-client streaming writes (different files) 12 GB/s
- ✓ Multi-client small file reads (different files) 500MB/s
- ✓ Multi-client small file writes (different files) 300MB/s

Moreover, storage system is expected to sustain up to 500K-2M IOPS without a significant level of performance penalty. A hyper-converged storage design is desirable to help improve performance, potentially reduce cost, and improve efficiency in data center. Moreover, system health and performance data should self-reporting, measurable and verifiable.

LeoFS Solution

For each data center, we suggest the following system: six 4U 36-bay servers as storage nodes (12 Gbps backplane), each with dual 10GbE network:

- ❖ CPU: 2*E5-2620V4
- ❖ Motherboard: Super Micro X10DRL-I
- ❖ RAM: 64GB
- ❖ System disk: one Intel 3520 enterprise SSD
- ❖ Data disks: 32*4TB Seagate enterprise SATA
- ❖ Expansion card LSI: 9300-8I (12G)

For metadata management, instead of standalone servers, use four 480GB Intel 3520 enterprise SSD, and embedded them into two of the six storage.

If network changes to quad 10GbE, average write performance of a single drive is over 100 MB/s (asynchronous write), the system can deliver:

- ✓ Single client steaming read 4 GB/s
- ✓ Single client streaming write 4 GB/s
- ✓ Multi-client streaming writes (different files) 14 GB/s

6. Getting Started

To get started with LeoFS, you can run initial evaluation with only a single Linux machine or minimum cluster of three servers. We also have laboratory machines available for a technology demo or remote testing.

Single server testing setup:

- CPU: e5-2620, two CPUs required to support 12 or more HDDs
- RAM: 64GB
- Motherboard: SuperMicro X10DRL-I
- Enterprise HDD, 4TB to 10TB, 7,200 rot.
- OS and Metadata: 2*480GB SSD (hardware raid for redundancy, 100GB for OS, others for metadata)
- Network Interface Card: 2*10GbE

Basic cluster testing needs three servers, same configurations as above except having separate OS and metadata SSDs:

- OS Disk: 240GB SSD
- Metadata: 4*480GB SSD (embedded in two storage nodes, two disks per node)