Introduction to LeoFS

A Market Proven Parallel File System for Data Intensive Storage

A Supreme Parallel File System

- One cluster for file, block and object-based storage
- Market proven - more than 1 EB capacity deployment
- Sustainable high performance - always saturate hardware throughput
- Reliable N+M erasure coding - up to 90% capacity utilization
- Easy scalability with no downtime or reboot
- Worry-free 24/7 customer support and management
- Great cost savings vs. traditional and open source file systems

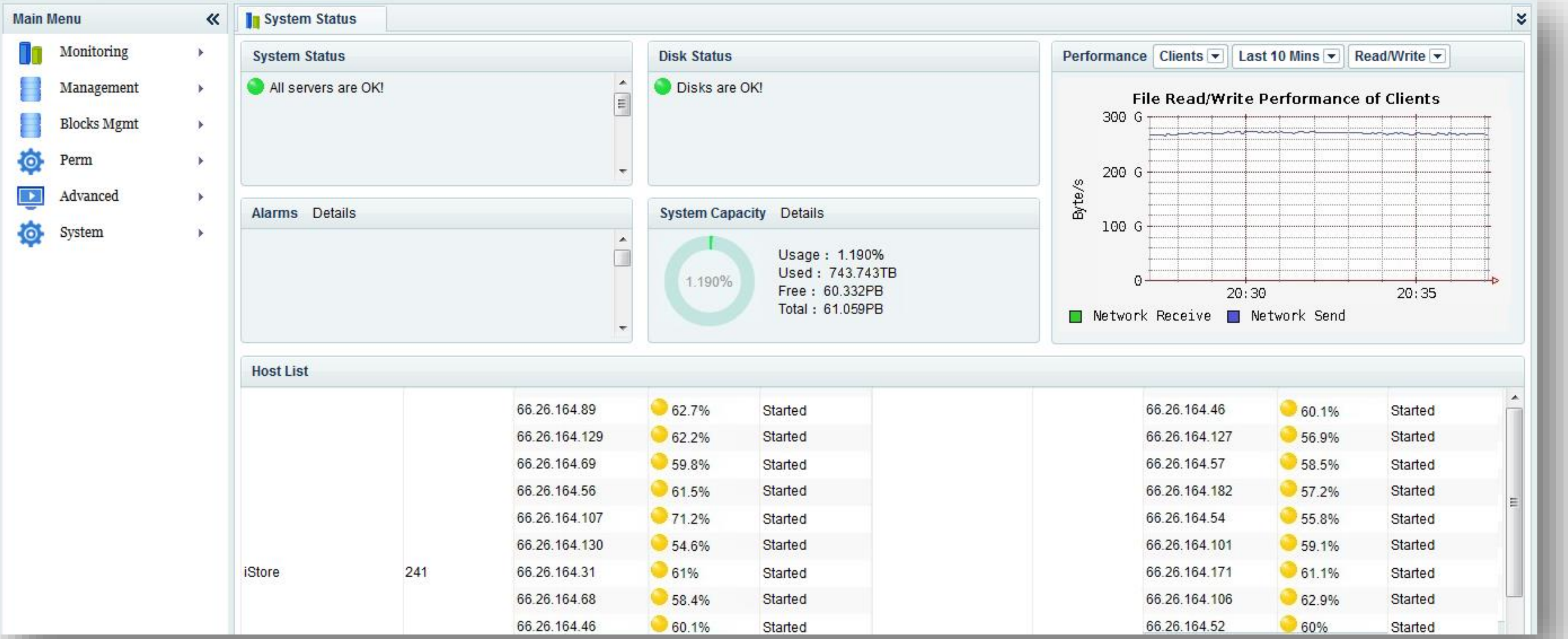
Current largest single cluster installed has 333 storage nodes, 95PB and over 200GB/s I/O throughput.

Customer Satisfaction Guaranteed

Successful deployment in Asia since 2009

Computational Storage

- Single server and cluster solutions, from tens of TB to hundreds of PB
- Both metadata and computing can be embedded in storage nodes
- Customer on-site 60PB cluster
    - 241 commodity servers: 4U 36-drive, dual Intel E5-2620v4, 64 GB RAM
    - 8 metadata + storage nodes: 2*480 GB SSDs, 34*8 TB SATA HDDs
    - 233 compute + storage nodes: 36*8TB SATA HDDs



System Status

● All servers are OK!

Disk Status

● Disks are OK!

Alarms Details

System Capacity Details

Usage : 1.190%
Used : 743.743TB
Free : 60.332PB
Total : 61.059PB

Host List

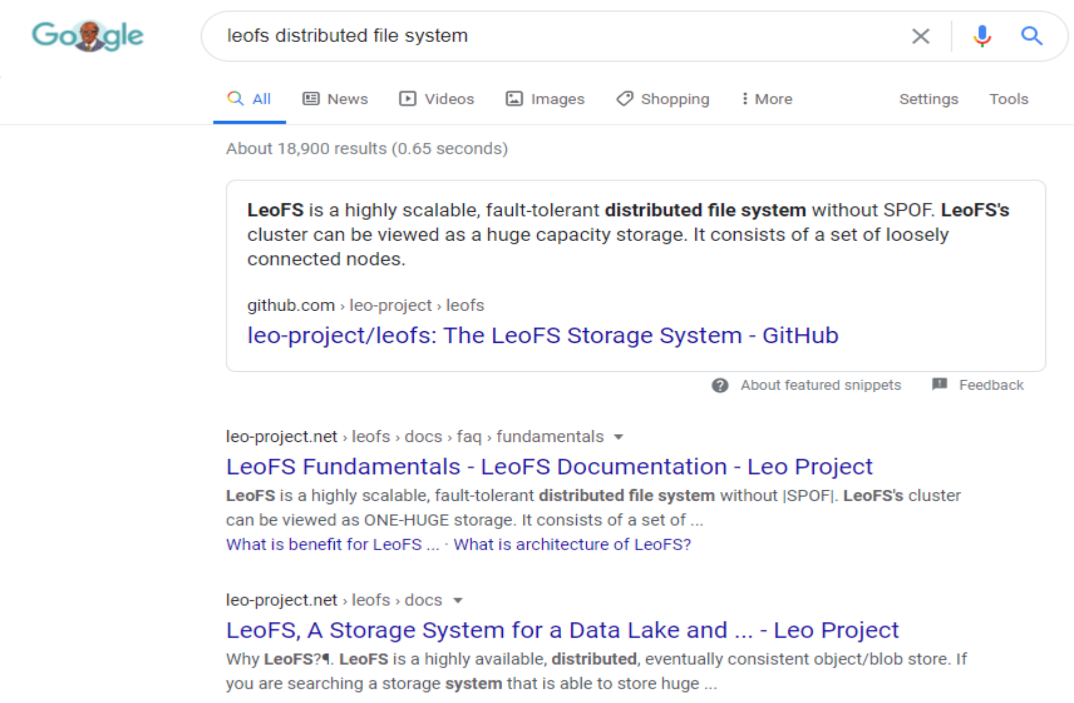
iStore	241	66.26.164.89	● 62.7%	Started			66.26.164.46	● 60.1%	Started
		66.26.164.129	● 62.2%	Started			66.26.164.127	● 56.9%	Started
		66.26.164.69	● 59.8%	Started			66.26.164.57	● 58.5%	Started
		66.26.164.56	● 61.5%	Started			66.26.164.182	● 57.2%	Started
		66.26.164.107	● 71.2%	Started			66.26.164.54	● 55.8%	Started
		66.26.164.130	● 54.6%	Started			66.26.164.101	● 59.1%	Started
		66.26.164.31	● 61%	Started			66.26.164.171	● 61.1%	Started
		66.26.164.68	● 58.4%	Started			66.26.164.106	● 62.9%	Started
		66.26.164.46	● 60.1%	Started			66.26.164.52	● 60%	Started

Hundreds of Customers

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

Selected References

- We take pride that most of our customers in PB usage started with only a few hundreds of TB
  - First Oil & Gas customer in 2009, I/O throughput 2x greater than StorNEXT
  - Higher Education: University of Florida, Georgia Southern University
  - Scientific Computing: Direct Electron (San Diego) on electron detection for biological molecules
  - Video Surveillance: Dante Security (New York)



leofs distributed file system

About 18,900 results (0.65 seconds)

LeoFS is a highly scalable, fault-tolerant **distributed file system** without SPOF. **LeoFS's** cluster can be viewed as a huge capacity storage. It consists of a set of loosely connected nodes.

github.com › leo-project › leofs

leo-project/leofs: The LeoFS Storage System - GitHub

About featured snippets · Feedback

leo-project.net › leofs › docs › faq › fundamentals

LeoFS Fundamentals - LeoFS Documentation - Leo Project

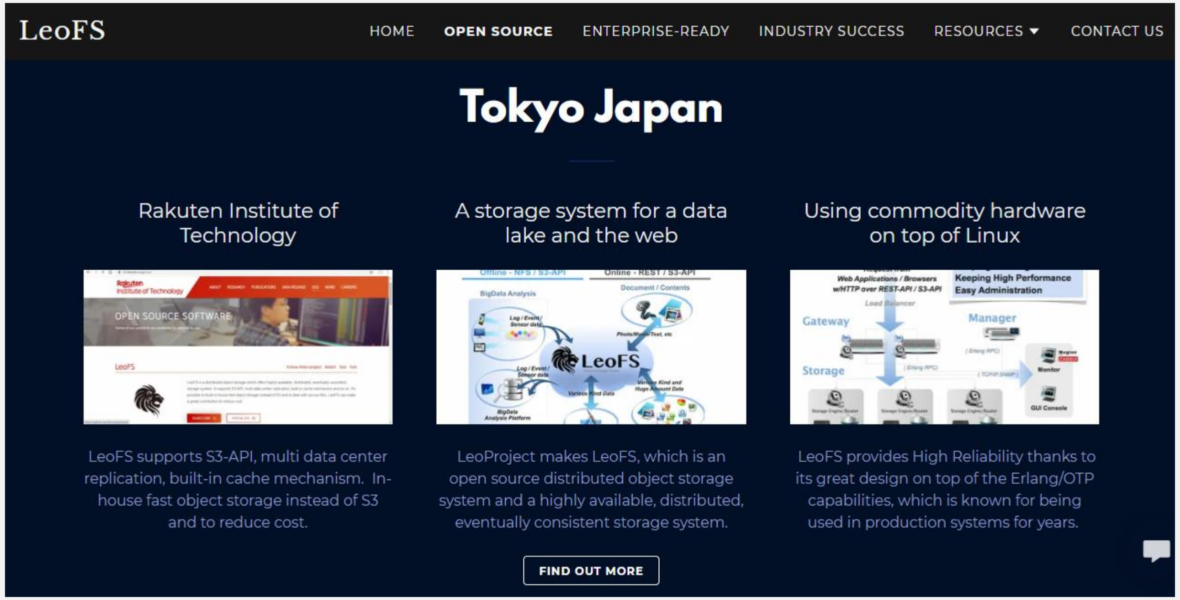
LeoFS is a highly scalable, fault-tolerant **distributed file system** without |SPOF|. **LeoFS's** cluster can be viewed as ONE-HUGE storage. It consists of a set of ...

What is benefit for LeoFS ... · What is architecture of LeoFS?

leo-project.net › leofs › docs

LeoFS, A Storage System for a Data Lake and ... - Leo Project

Why **LeoFS**?¶. **LeoFS** is a highly available, **distributed**, eventually consistent object/blob store. If you are searching a storage **system** that is able to store huge ...



Tokyo Japan

Rakuten Institute of Technology

A storage system for a data lake and the web

Using commodity hardware on top of Linux

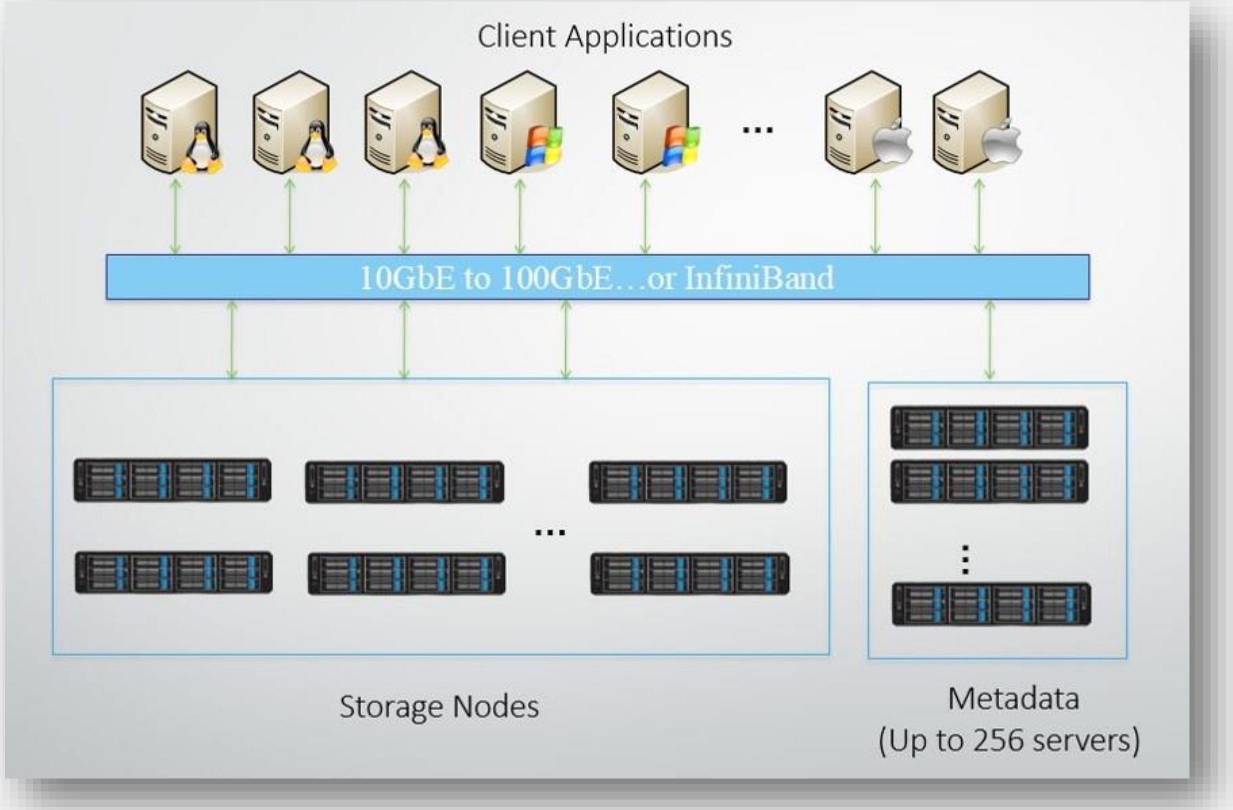
LeoFS supports S3-API, multi data center replication, built-in cache mechanism. In-house fast object storage instead of S3 and to reduce cost.

LeoProject makes LeoFS, which is an open source distributed object storage system and a highly available, distributed, eventually consistent storage system.

LeoFS provides High Reliability thanks to its great design on top of the Erlang/OTP capabilities, which is known for being used in production systems for years.

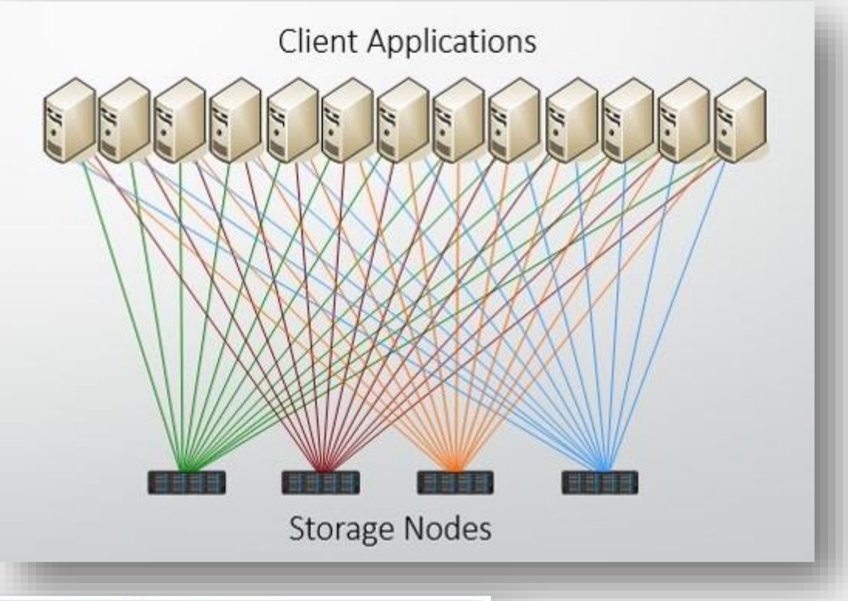
Solution Architecture

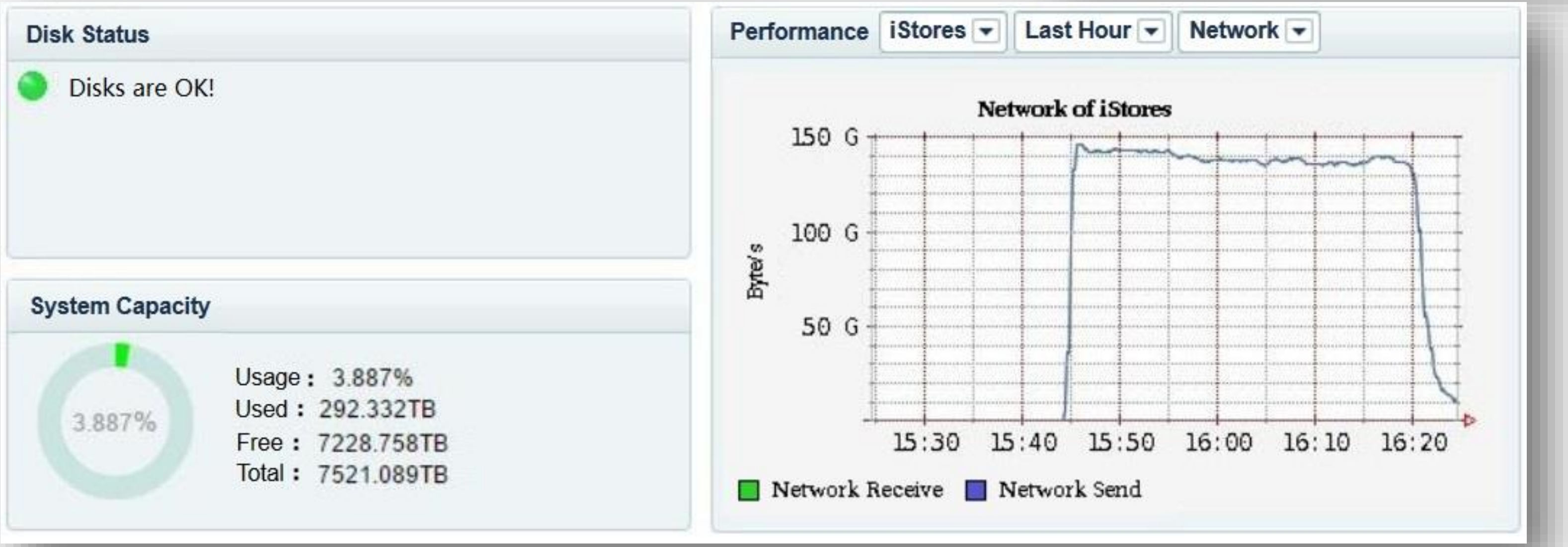
- Fully-POSIX compliant
- Native clients, all kernel modules that do not require any patches
- Best choice for combined values
  - High performance
  - Large and scalable capacity
  - Reliable data protection
  - Professional support
  - Affordable cost



No Bottleneck

- Without controllers, gateways, nor distributors
  - Data files are transparently distributed over multiple nodes
  - All client applications communicate with all storage nodes
- Customer on-site 7.5PB
  - Dual 10GbE network
  - 102 nodes of 4U 24-bay storage servers
  - Total of 1,880 drives with 4TB SATA HDDs
  - Aggregated I/O close to 150GB/s
  - Average single drive write 75MB/s





Disk Status

● Disks are OK!

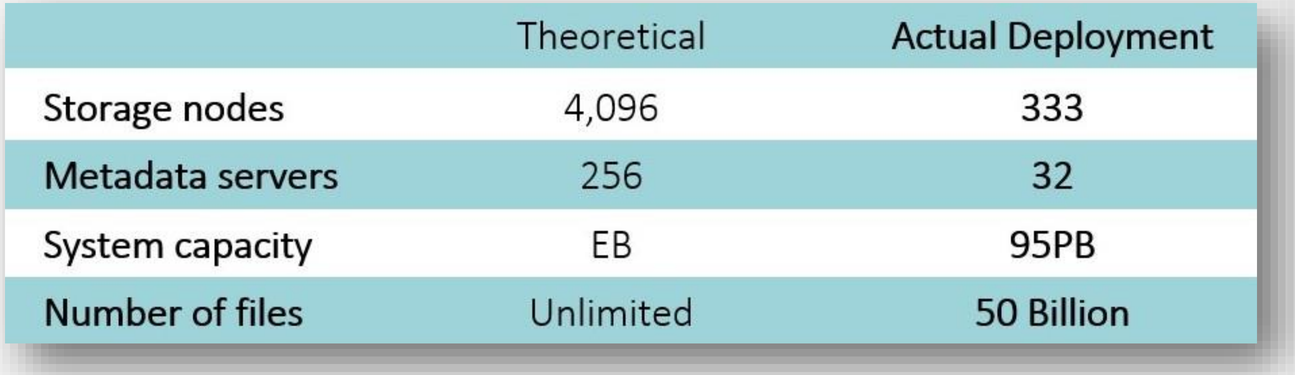
System Capacity

Usage : 3.887%
Used : 292.332TB
Free : 7228.758TB
Total : 7521.089TB

Performance iStores ▾ Last Hour ▾ Network ▾

Unlimited Capacity

- Software as a service, option to choose cluster or single server products
  - Cluster starts with 2 nodes, and up to thousands
  - Single server supports 12-bay, 16-bay, 24-bay, or 36-bay drives
- System capacity and throughput always increase with additional server or drive
- Solution threshold



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

File-level N+M Erasure Coding

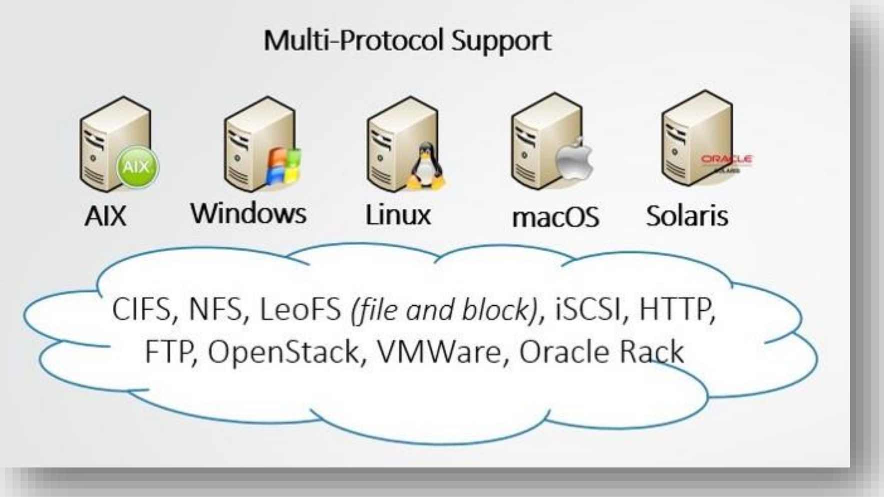
- Data content is distributed on a file-level across different storage nodes
  - When N+2 is applied, cluster can sustain operation up to two simultaneous failures
- Optimum data protection plans for different files
  - Better capacity utilization, up to 90% with 16+1
- With failed hardware, LeoFS rebuilds only the files that are affected, and it uses the entire cluster to rebuild
  - No downtime nor reboot
  - One TB data usually takes less than 20 minutes
- Highest level of data availability
  - System capable of self-monitoring and self-healing

Peace of Mind

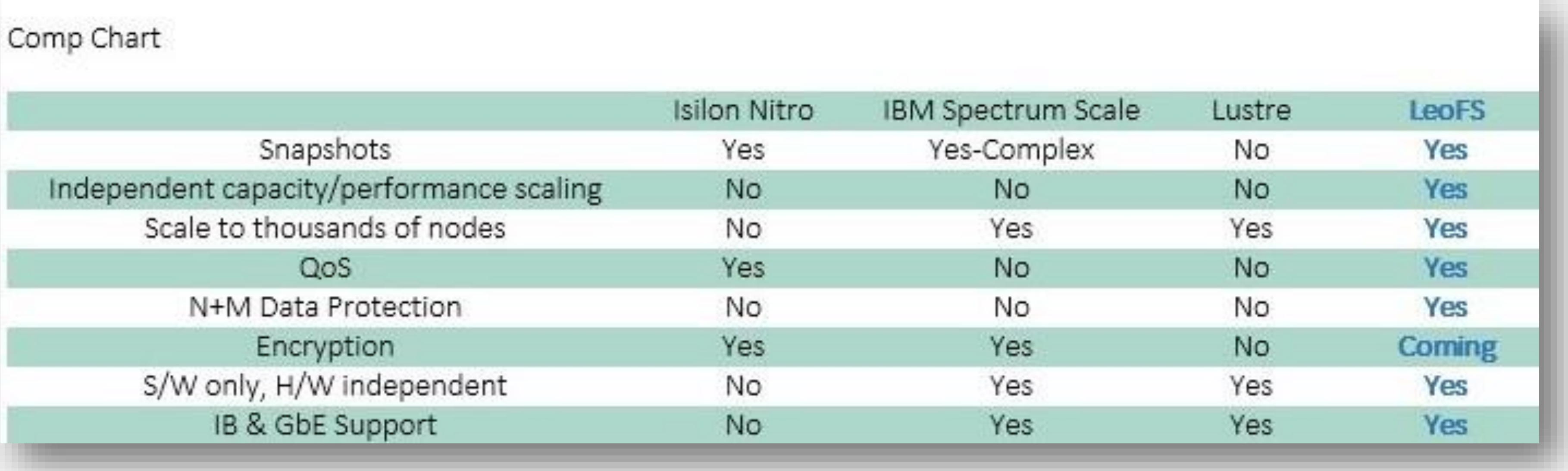
- It’s CARE, MANAGEMENT and SUPPORT
  - Cluster monitoring
  - Software maintenance and update
  - High quality hardware
- With direct access to file system developers
  - Designated consultants are available 24/7
  - Most work can be done remotely
- Options to choose
  - Next Business Day Service Level Agreement
  - Re-mote or On-Site Support Warranty
  - Advanced Hardware Replacement

Enterprise Features

- Load balance switch, 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


Multi-Protocol Support

CIFS, NFS, LeoFS *(file and block)*, iSCSI, HTTP, FTP, OpenStack, VMWare, Oracle Rack


Comp Chart

	Isilon Nitro	IBM Spectrum Scale	Lustre	LeoFS
Snapshots	Yes	Yes-Complex	No	Yes
Independent capacity/performance scaling	No	No	No	Yes
Scale to thousands of nodes	No	Yes	Yes	Yes
QoS	Yes	No	No	Yes
N+M Data Protection	No	No	No	Yes
Encryption	Yes	Yes	No	Coming
S/W only, H/W independent	No	Yes	Yes	Yes
IB & GbE Support	No	Yes	Yes	Yes

Solution Sample – University of California

- UCSF Wynton HPC center:
  - ✓ 1.2PB storage, cost about $192K USD
  - ✓ Hardware: 4 nodes of 60-bay servers, 2 nodes of metadata servers
  - ✓ Software: ZFS and BeeGFS
  - ✓ https://wynton.ucsf.edu/hpc/about/pricing-storage.html Genomics
- Competitive LeoFS cluster
  - ✓ 1.4PB storage, only $100K USD
  - ✓ Throughput: read 6.8GB/s, write 10GB/s (asynchronous)
  - ✓ Hardware: 4 nodes of 36-bay servers
    - 2 Metadata + Storage nodes
      - CPU:Intel Xeon E5-2630 V4 x 2
      - Motherboard: Supermicro X10DRL-i
      - HBA: LSI SAS 9300-8I
      - System Disk Drive: 480GB SSD x 2 + 240GB SSD x 2
      - Storage Disk Drive: 10TB HDD x 34
      - RAM: 128GB
      - Network Port: 4 x 10 GbE
    - 2 Storage-only node
      - CPU:Intel Xeon E5-2630 V4 x 2
      - Motherboard: Supermicro X10DRL-i
      - HBA: LSI SAS 9300-8I
      - System Disk Drive: 240GB SSD x 2
      - Storage Disk Drive: 10TB HDD x 36
      - RAM: 64GB
      - Network Port: 4 x 10 GbE

Solution Sample – Geneva Observatory

- BeeGFS case study of Geneva Observatory
  - 4 storage nodes and 2 metadata servers, Infiniband
  - Effective 800TB, 144 drives, I/O 5-8GB/s
  - https://www.hpc-ch.org/hpc-ch-forum-yves-revaz-epfl-beegfs-the-hpc-storage-solution-adopted-at-the-geneva-observatory/
- Competitive LeoFS cluster
  - 4 nodes of 4U 36-bay storage servers, 2 metadata servers, dual 10GbE
  - Usable 1PB, 144 drives, I/O 8-11GB/s
  - No buddy mirroring, 80% capacity utilization
  - No single point of failure from drives, nodes or network
  - File-level RAID, faster data recovery
  - Better ROI

BeeGFS

The HPC Storage Solution Adopted at the
Geneva Observatory

Yves Revaz

The Geneva Observatory in a nutshell

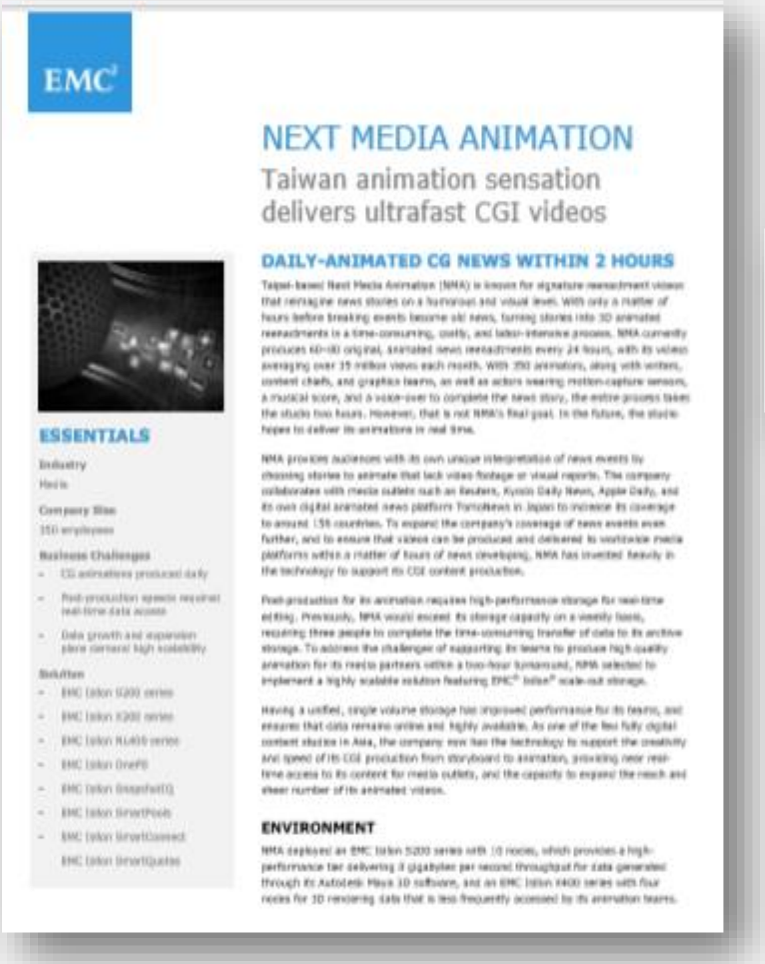
Main fields of research

- Stellar Physics
- Galaxy and Cosmology
- High energy astrophysics
- Extra-solar planets

Vs. EMC Isilon S Series

- From EMC 3D animation customer case
  - Ten nodes of Isilon S200 series
  - Raw capacity of 600 TB
  - I/O throughput 8 GB/s
- Comparable LeoFS solution
  - Commodity hardware, dual 10GbE network
  - Ten nodes of 4U 24-bay storage servers
  - Raw capacity of 960 TB
  - I/O throughput over 12 GB/s

60% more capacity and 50% higher throughput



EMC²

NEXT MEDIA ANIMATION

Taiwan animation sensation delivers ultrafast CGI videos

DAILY-ANIMATED CG NEWS WITHIN 2 HOURS

Taipei-based Next Media Animation (NMA) is known for signature reenactment videos that reimagine news stories on a humorous and visual level. With only a matter of hours before breaking events become old news, turning stories into 3D animated reenactments is a time-consuming, costly, and labor-intensive process. NMA currently produces 60-80 original, animated news reenactments every 24 hours, with its videos averaging over 15 million views each month. With 350 animators, along with writers, content chiefs, and graphics teams, as well as actors wearing motion-capture sensors, a musical score, and a voice-over to complete the news story, the entire process takes the studio two hours. However, that is not NMA's final goal. In the future, the studio hopes to deliver its animations in real time.

NMA provides audiences with its own unique interpretation of news events by choosing stories to animate that lack video footage or visual reports. The company collaborates with media outlets such as Reuters, Kyodo Daily News, Apple Daily, and its own digital animated news platform TomoNews in Japan to increase its coverage to around 150 countries. To expand the company's coverage of news events even further, and to ensure that videos can be produced and delivered to worldwide media platforms within a matter of hours of news developing, NMA has invested heavily in the technology to support its CGI content production.

Post-production for its animation requires high-performance storage for real-time editing. Previously, NMA would exceed its storage capacity on a weekly basis, requiring three people to complete the time-consuming transfer of data to its archive storage. To address the challenges of supporting its teams to produce high-quality animation for its media partners within a two-hour turnaround, NMA selected to implement a highly scalable solution featuring EMC® Isilon® scale-out storage.

Having a unified, single volume storage has improved performance for its teams, and ensures that data remains online and highly available. As one of the first fully digital content studios in Asia, the company now has the technology to support the creativity and speed of its CGI production from storyboard to animation, providing near real-time access to its content for media outlets, and the capacity to expand the reach and sheer number of its animated videos.

ENVIRONMENT

NMA deployed an EMC Isilon S200 series with 10 nodes, which provides a high-performance tier delivering 8 gigabytes per second throughput for data generated through its Autodesk Maya 3D software, and an EMC Isilon X400 series with four nodes for 3D rendering data that is less frequently accessed by its animation teams.

ESSENTIALS

Industry
Media

Company Size
350 employees

Business Challenges
- CG animations produced daily
- Post-production speeds required real-time data access
- Data growth and expansion plans demand high scalability

Solution
- EMC Isilon S200 series
- EMC Isilon X200 series
- EMC Isilon NL400 series
- EMC Isilon OneFS
- EMC Isilon SnapshotIQ
- EMC Isilon SmartPools
- EMC Isilon SmartConnect
- EMC Isilon SmartQuotas

Vs. DDN ES7K

- World's fastest Lustre appliance
  - SSD cache, SSD, performance SAS, and capacity SAS
  - 8U 144 drives, I/O up to 12 GB/s
- Comparable LeoFS solution
  - Six nodes of 4U 24-bay storage servers, 144 drives
  - Using only SATA HDDs for storage, 4 SSD drives for metadata
  - Sustainable I/O without deterioration over time

A better balance of Price, Performance and Capacity.

DDN® STORAGE

PRODUCT BROCHURE

SCALER APPLIANCES

ES7K™: WORLD'S FASTEST ENTRY-LEVEL LUSTRE APPLIANCE

World's fastest entry-level Lustre® solution delivers easy migration to parallel file systems

Highest Performance and Density - Intel® Enterprise Edition Lustre

Organizations with high performance workflows find their productivity is capped by how quickly they can access data. Breakthroughs in research, oil and gas, financial services, manufacturing and big data applications require massive throughput and highly parallel access that are challenging traditional SAN and NAS storage technologies.

Ideally, these challenges would be addressed with a data infrastructure solution that:

- Is simple to install, configure, manage and, scales linearly
- Delivers massive storage performance density at scale
- Automatically unifies data across different storage types
- Offers the flexibility of open source with the security of a standard supported product

INTRODUCING THE WORLD'S FASTEST ENTRY LEVEL LUSTRE APPLIANCE

Parallel file systems offer a higher performance alternative to traditional SAN and NAS architectures, but organizations attempting to deploy white box storage cobbled together with open source file systems are running into deployment, management, performance and scaling challenges.

Ideally, parallel file system solutions would have the extreme performance of a reliable and proven open source parallel file system with exclusive performance and stability features, and support combined in a high throughput all-in-one appliance enabling high performance and density at scale.

DDN, sells, installs and supports more Lustre than any other storage vendor, and is the #1 reseller of Intel® Enterprise Edition for Lustre (IEEL). DDN's EXAScaler® family of massively scalable Lustre appliances deliver enterprise grade reliability and support combined with extreme-scale open source innovations.

EXAScaler combines Lustre performance and scalability with DDN-unique features to deliver: ease of install and configuration, drive-to-file system monitoring, DDN-only performance features for accelerated metadata, and unified data across different storage types.

The ES7K is the world's fastest and densest entry-level Lustre appliance. Delivering 1.5 GB/s per rack unit* and 124 TB per rack unit*, the ES7K is the perfect system for moving your most demanding applications to a parallel file system infrastructure. Not only does the ES7K outperform its closest competitors by almost 3X, it also has all the usability features and support of the mature EXAScaler product family (ES7K and ES12K).

* - industry standard 42U rack

The World's Fastest Entry-Level Lustre Appliance

Start Small

- Base building block configuration of up to 144 drives in 8U
- Up to 12GB/s in 8U - higher performance at smaller footprint than closest competitor
- Proven, easy to deploy, converged Lustre solution

Scale Up and Out

- Up to 5 PB per rack, capable of spanning multiple racks
- Up to 60 GB/s per rack, capable of spanning multiple racks

ES7K ADVANTAGES

Open source and fully supported parallel file system appliance

Start small and scale non-disruptively by adding ES7K systems as building blocks

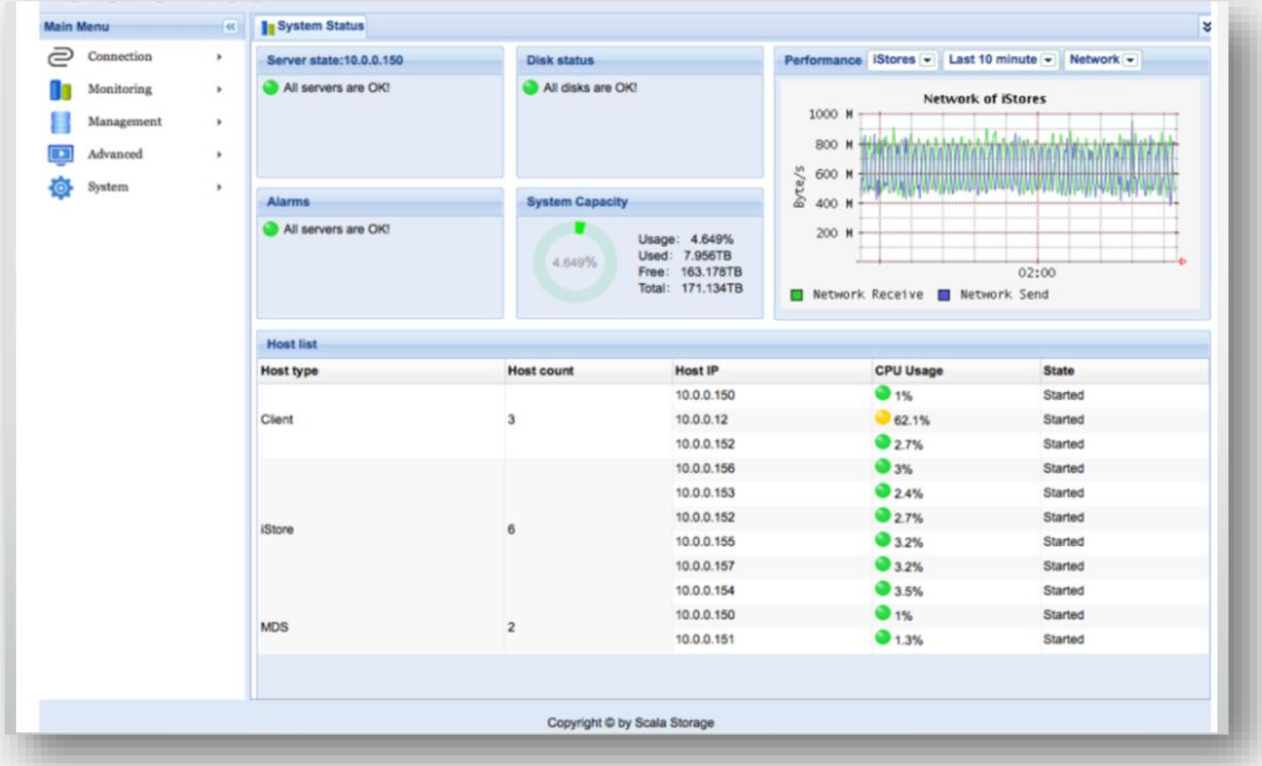
Simplify management across cloud, and enable multisite collaboration

Risk-free, tried, tested and proven open source Lustre based solution

Vs. Lustre

- Easy installation and management
  - No problems having a large number of files in a single directory
  - No problems accessing huge amount of small files
  - No worries using ls –l
- Metadata cluster up to 256 servers, scalable performance
  - Each pair 20,000 file creates per second
  - Standalone or within storage nodes

Single Log-on GUI



Server state:10.0.0.150

● All servers are OK!

Disk status

● All disks are OK!

Alarms

● All servers are OK!

System Capacity

Usage: 4.649%
Used: 7.956TB
Free: 163.178TB
Total: 171.134TB

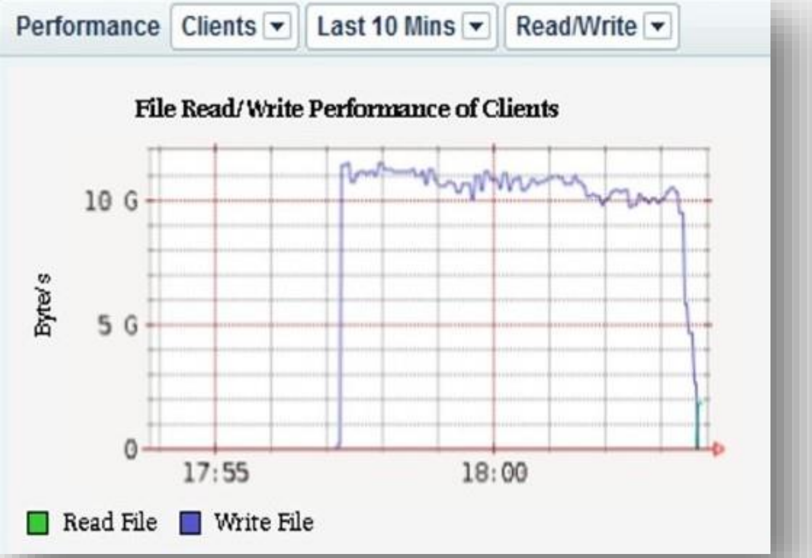
Host list

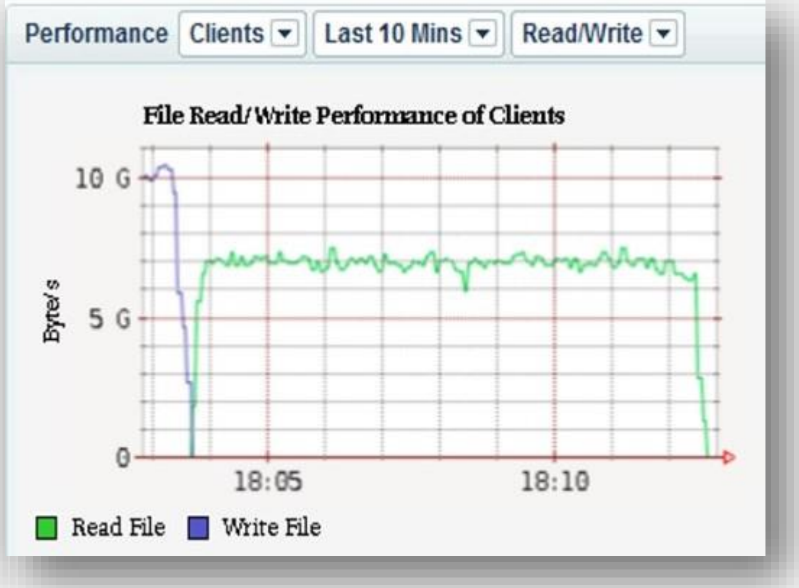
Host type	Host count	Host IP	CPU Usage	State
Client	3	10.0.0.150	● 1%	Started
		10.0.0.12	● 62.1%	Started
		10.0.0.152	● 2.7%	Started
iStore	6	10.0.0.156	● 3%	Started
		10.0.0.153	● 2.4%	Started
		10.0.0.152	● 2.7%	Started
		10.0.0.155	● 3.2%	Started
		10.0.0.157	● 3.2%	Started
		10.0.0.154	● 3.5%	Started
MDS	2	10.0.0.150	● 1%	Started
		10.0.0.151	● 1.3%	Started

Copyright © by Scala Storage

Throughput vs. Lustre and GPFS

- Lustre/GPFS*: no data on X clients and X streams
  - Six LUNs, each chassis with 30*4TB disks, total of 180 HDDs
  - Because of controller, limitation seen on 8 clients, 1 stream
    - Write: 4-5 GB/s, Read: 5-7 GB/s
- LeoFS without controller limitation
  - Eight nodes of 4U 24-bay storage servers, total of 192 HDDs
  - Dual 10 GbE network, 8 clients, 20 streams, 4TB testing data
    - Each client 500GB (greater than 64GB RAM)
    - File size 1MB, redundancy 8+3
    - Write 11GB/s, Read 7GB/s

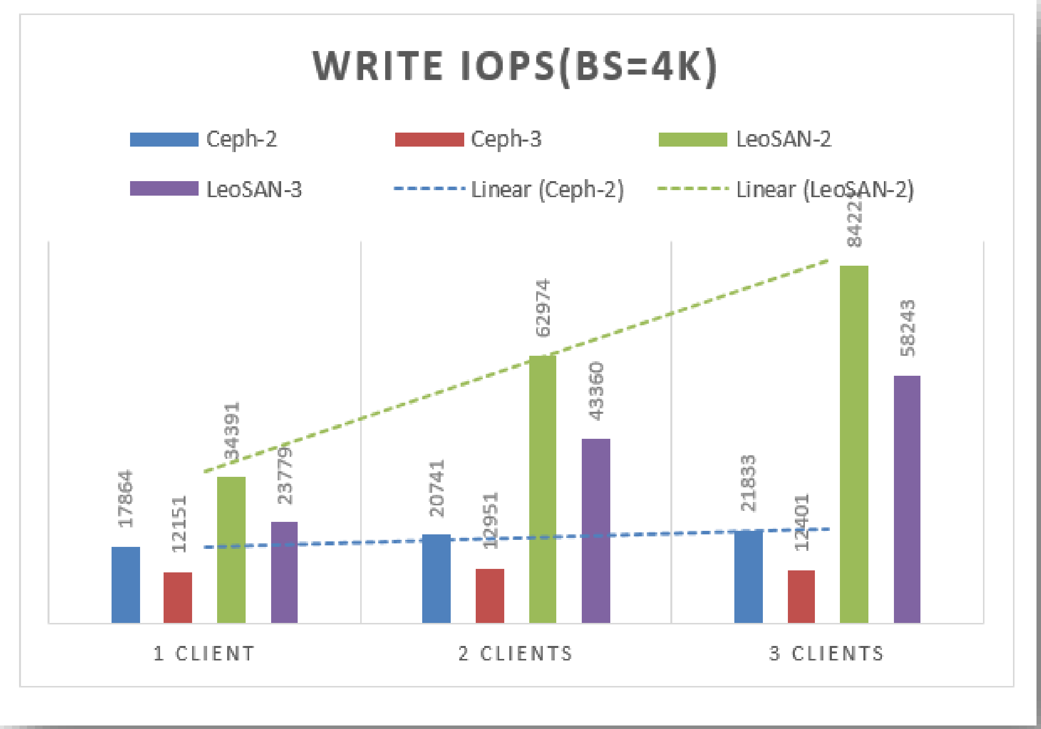


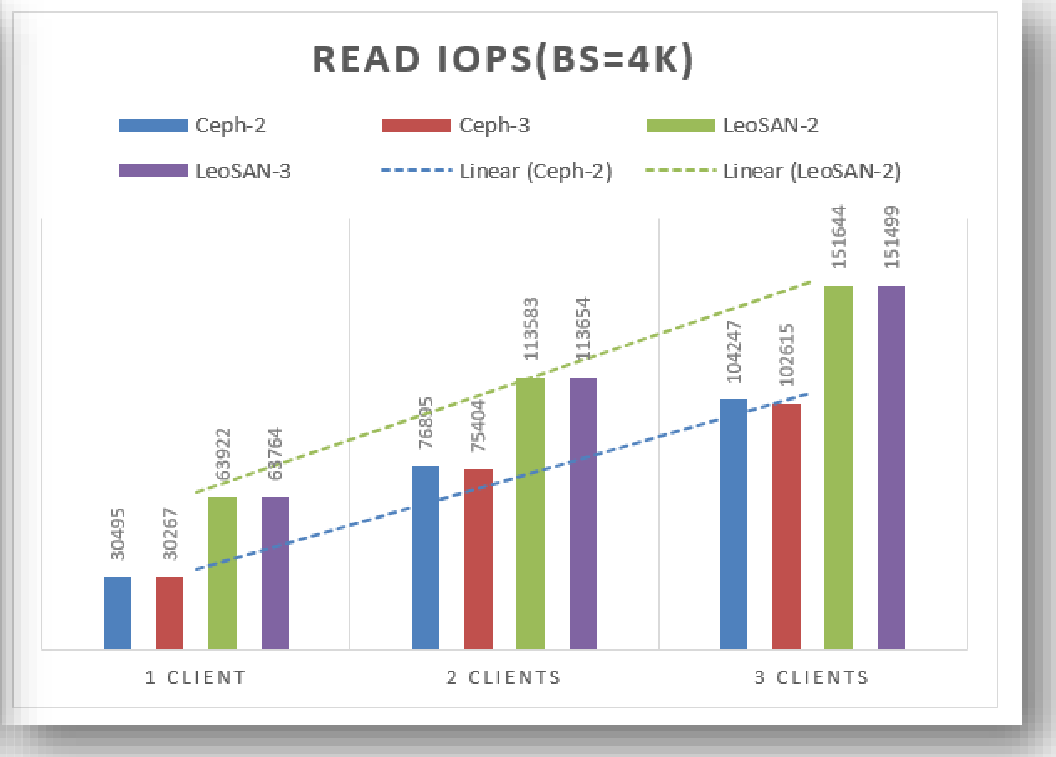


** CERN's presentation on High Performance Storage in Science, SDC 2017*

Vs. Ceph

- Block device SSD IOPS
- Testing hardware
  - 3 nodes of 12-bay storage servers, each with 12*240GB SSD
  - Dual 10GbE, 3 clients, block size 4KB
  - Replications: 2 and 3





Products – Cluster

- Starting minimum 2 nodes
  - New system or add to existing LeoFS installation
  - Option to build separate instances of LeoFS
  - Hyper-convergence option
  - Integrated virtualization technology
  - Automated self-monitoring and self-healing
  - Potentially geographically redundant storage (LeoSync)
  - Supports CIFS, NFS, FTP, HTTP, ISCSI, OpenStack and Hadoop
  - LeoSAN and LeoFS private block and file interfaces
  - Fine-grained access rights
- 2U 12 bay to 4U 36 bay chassis
  - CPU: Intel Xeon E5-2630 V4 * 2
  - Mother Board: Supermicro X10DRL-i
  - HBA: LSI SAS 9300-8I
  - System Disk Drive: 480GB SSD * 2 + 240GB SSD * 2
  - Storage Disk Drive: 2TB to 10TB HDD, SATA or SAS
  - RAM: 128GB
  - Network Port: 2 * 10 GbE or 40 GbE or Infiniband

Products – Single Server

- Benefits over typical NAS or SAN products
  - Plug and use, fully-POSIX complaint, supports NFS or ISCSI
  - Inexpensive commodity hardware, HDDs, 1 GbE or 10 GbE
  - Single directory up to 10 billion files
  - High performance, throughput > 2.5 GB/s, IOPS > 30,000
  - File-level RAID, better fault tolerance than RAID 6
  - Dynamic scalability with load balance switch
  - Faster online disk rebuild or replacement, no operation interruption
- 2U 12 bay to 4U 36 bay chassis
  - CPU: Intel Xeon E5-2620 V4 * 2
  - Mother Board: Supermicro X10DRL-i
  - HBA: LSI SAS 9300-8I
  - System Disk Drive: 480GB SSD * 2 + 240GB SSD * 2
  - Storage Disk Drive: 2TB to 10TB HDD, SATA or SAS
  - RAM: 64GB
  - Network Port: 2 * 1GbE or 10 GbE

Thank you, let’s get in touch

- No licensing fees
  - Free 1PB case study usage if replacing other parallel file system
  - Monthly service subscription with optional 24/7 support
  - Unbeatable pricing, contact us for your free solution estimate

www.leofs.info