Reference Guide
Hyperscalers with LeoFS

Tuesday, 28 November 2023

1 CONTENTS

2 INTRODUCTION

LeoFS is market proven parallel file system for different types of I/O intensive workloads.

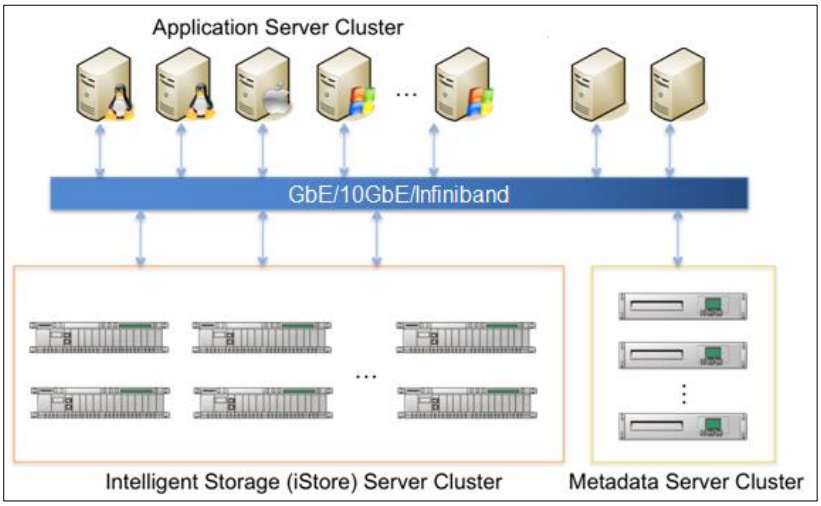
- One cluster for file, block and object-based storage
- Market proven - more than 1 EB capacity deployment
- Sustainable high performance – aggregated, no limit
- N+M erasure coding - up to 94% capacity utilization
- Easy scalability with no downtime or reboot
- More reliable vs. traditional and open source file systems
- Worry-free 24/7 customer support and management

LeoFS is fully POSIX-Compliant and compatible with all software applications, x86 based servers and IP networks. It is used in almost all industries and has hundreds of customers:

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

3 SOLUTION ARCHITECTURE

LeoFS Structure



LeoFS is a full POSIX compliant storage. It comes with native clients for Linux, Windows and macOS, all kernel modules that do not require any patches, using only commodity hardware.

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.

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.

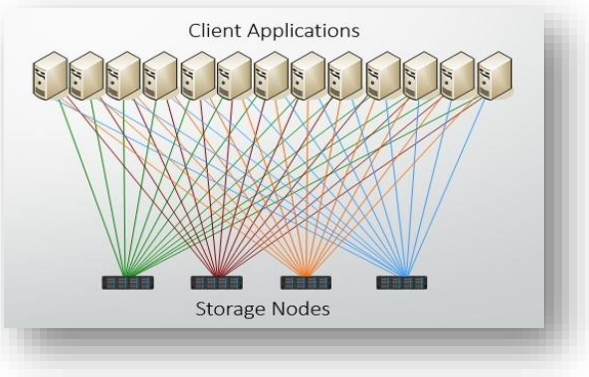
	Theoretical	Actual Deployment
Storage nodes	10,000	333
Metadata servers	256	32
System capacity	EB	95PB
Number of files	Unlimited	50 Billion

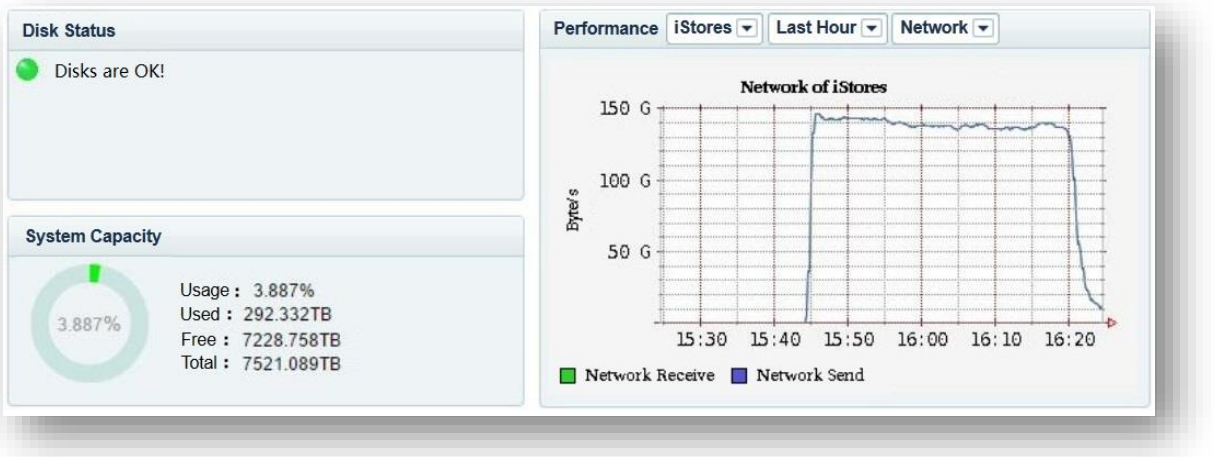
LeoFS is software defined storage system, It doesn't depand on specific hardware or devices, so all storage server of hyperscalers can be used to build LeoFS system.

4 CHARACTERISTICS

No Bottleneck

No controllers, gateways, nor distributors is required in LeoFS system and data files are transparently distributed over multiple nodes. All client applications communicate directly with all storage nodes.





Customer on-site 7.5PB LeoFS system, consists of 102 nodes of 4U 24-bay storage servers, with totally 1,880 drives of 4TB SATA HDDs. Dual 10GbE network is used and provides an aggregated I/O close to 150GB/s. Average single drive write 75MB/s.

File-level N+M Erasure Coding

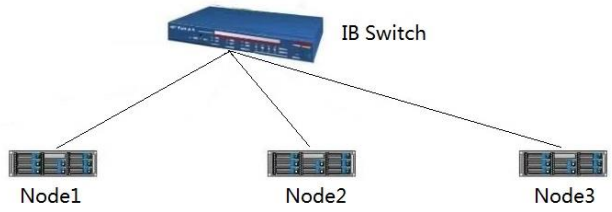
LeoFS supports N+M erasure coding data protection. Each data content is distributed on a file-level across different storage nodes. When N+2 is applied, cluster can sustain operation up to two simultaneous failures. User can set optimum data protection plans for different files. The capacity utilization ratio will be over 94% with EC 16+1. LeoFS provides self-monitoring and self-healing mechanism to ensure high data availability. LeoFS rebuilds only the files that are affected when hardware failure occurs and it uses all disk devices in the entire cluster to rebuild. It takes usually less than 20 minutes to rebuild one TB data. No downtime nor reboot required.

Enterprise Features

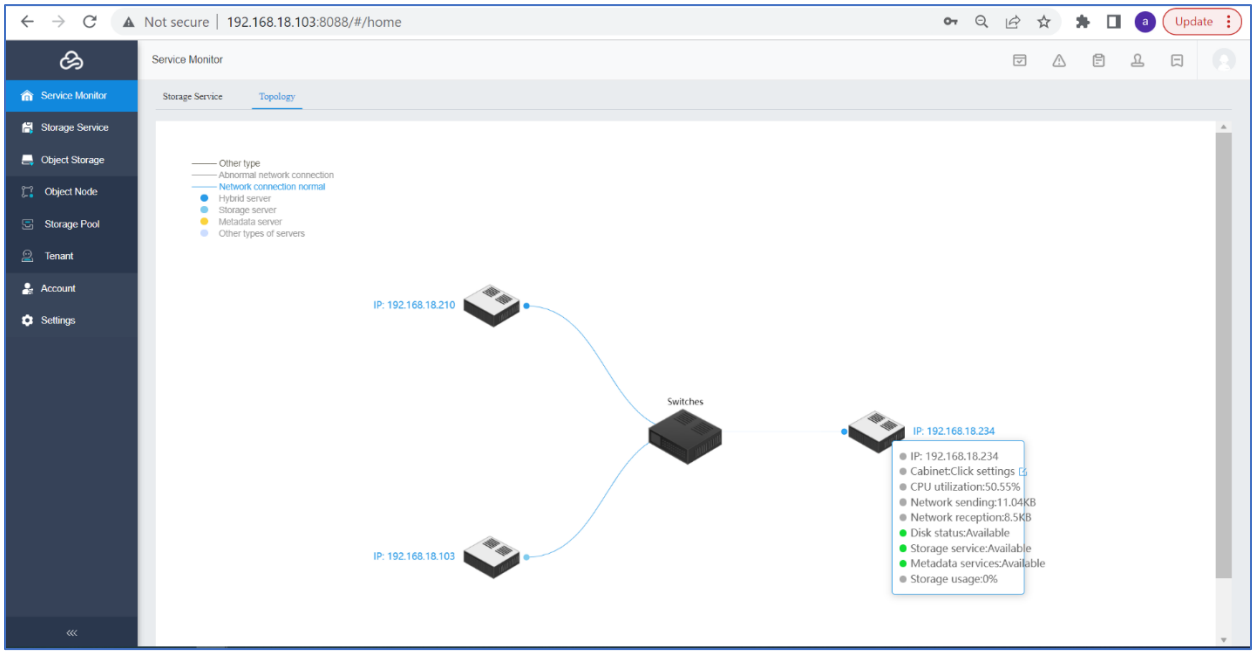
- Load balance switch, hardware evenly share system workload
- Runs on platforms such as x86, OpenPOWER, ARM, and Xeon Phi
- Provides access interfaces like Samba, NFS, FTP, HTTP, iSCSI, S3 etc.
- Support for group/user ACLs and quota
- Fully active network with automatic failure detection
- Supports Infiniband, GigE, multiple subnet and bonding
- Cold data sanity check, automatic repair, no downtime
- WORM directory, avoid modification of saved data

5 TEST ENVIRONMENT

Test Environment Diagram



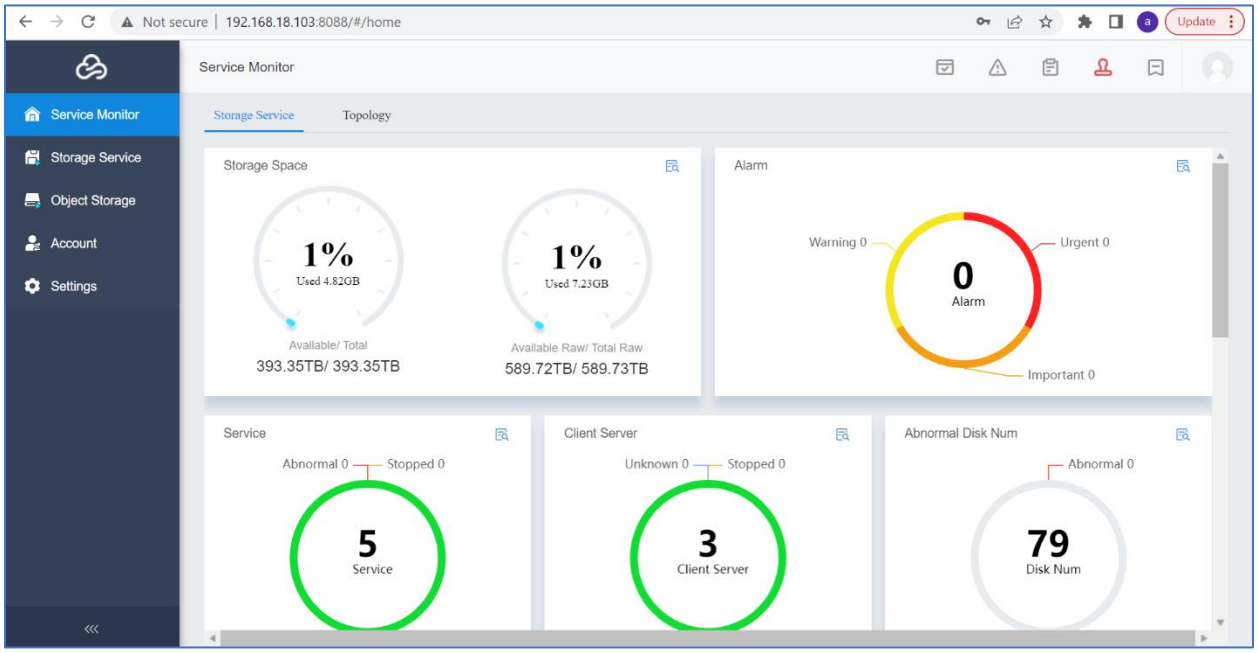
Network Topology



Hardware Configuration

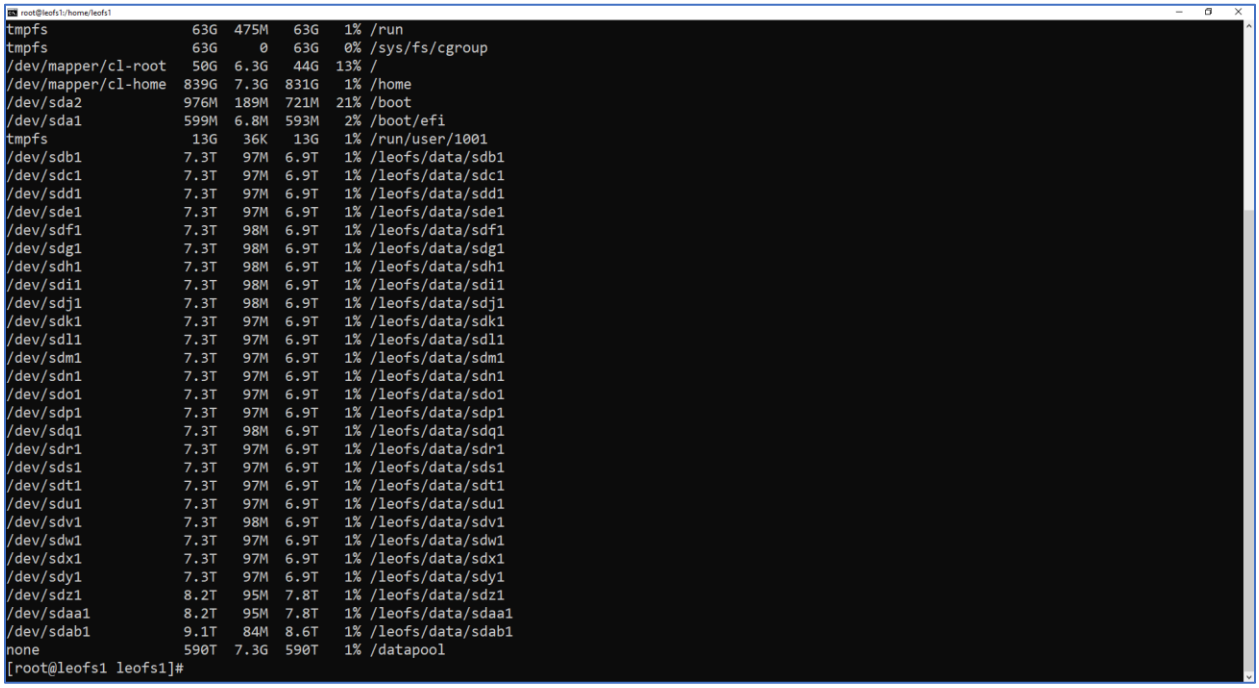
Server Type	Configuration		Qty.
Storage Server+Metadata Server (QuantaPlex T22P-4U)	CPU	2x Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz	3
	Mem	RAM – 4x Samsung 32 GB 3200MT/s M393A4G40AB3-CWE (128GB)	
	Network	1x OCP Mezz 40Gb QSFP 1 port Quanta ON 40GbE CX3PRO ConnectX-3 Pro VPI PCI-E X 8 Gen 3 35S2BMA0020 1 PORT 40G QSFP+	
	OS SSD drive	1x SATA SAMSUNG_MZ7LH960HAJR-00005 960GB (Centos 8.x)	
	Disk drive	22*SEAGATE ST4000NM0023 4TB NVMe SSD drives – 1x M.2 960GB Samsung EVO plus 1x PCIe HHHL 1.9TB Samsung PCIe	

Dashboard



NAS File system

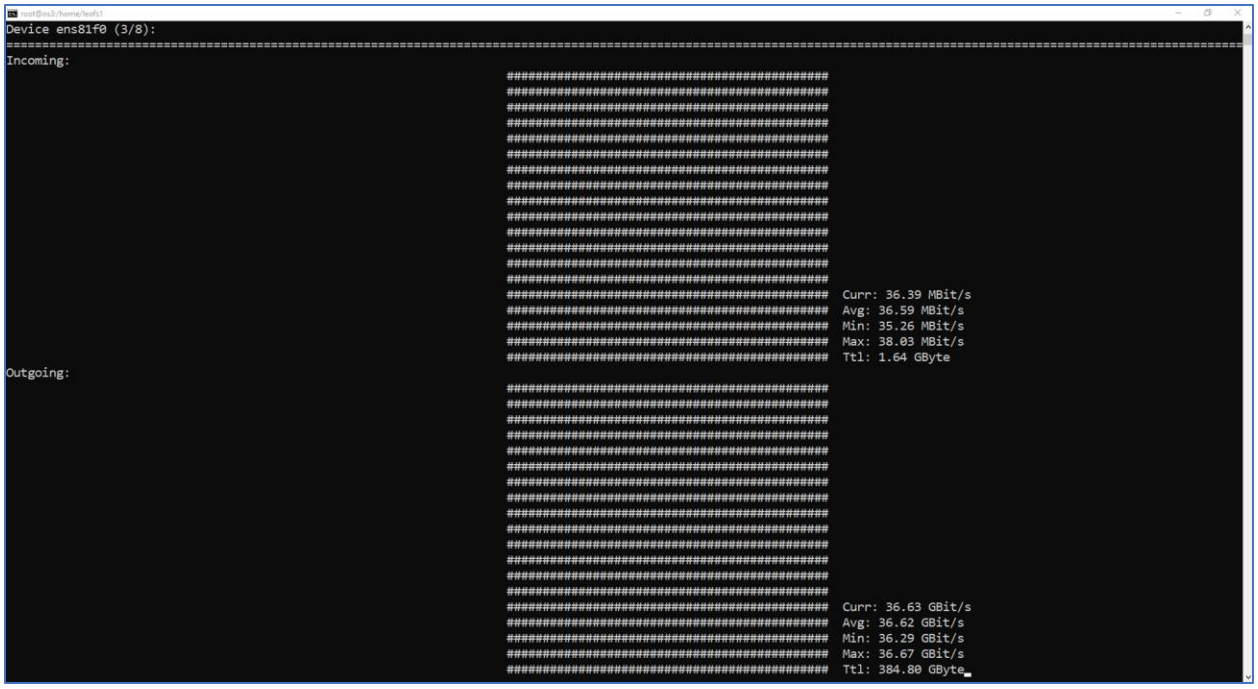
The filesystem exposed via NFS is attached to the host nodes as a Network Attached Storage (NAS) data pool to validate the performance.



```
tmpfs                63G  475M   63G   1% /run
tmpfs                63G     0   63G   0% /sys/fs/cgroup
/dev/mapper/cl-root   50G  6.3G   44G  13% /
/dev/mapper/cl-home  839G  7.3G  831G   1% /home
/dev/sda2            976M  189M  721M  21% /boot
/dev/sda1            599M  6.8M  593M   2% /boot/efi
tmpfs                 13G   36K   13G   1% /run/user/1001
/dev/sdb1            7.3T   97M  6.9T   1% /leofs/data/sdb1
/dev/sdc1            7.3T   97M  6.9T   1% /leofs/data/sdc1
/dev/sdd1            7.3T   97M  6.9T   1% /leofs/data/sdd1
/dev/sde1            7.3T   97M  6.9T   1% /leofs/data/sde1
/dev/sdf1            7.3T   98M  6.9T   1% /leofs/data/sdf1
/dev/sdg1            7.3T   98M  6.9T   1% /leofs/data/sdg1
/dev/sdh1            7.3T   98M  6.9T   1% /leofs/data/sdh1
/dev/sdi1            7.3T   98M  6.9T   1% /leofs/data/sdi1
/dev/sdj1            7.3T   98M  6.9T   1% /leofs/data/sdj1
/dev/sdk1            7.3T   97M  6.9T   1% /leofs/data/sdk1
/dev/sdl1            7.3T   97M  6.9T   1% /leofs/data/sdl1
/dev/sdm1            7.3T   97M  6.9T   1% /leofs/data/sdm1
/dev/sdn1            7.3T   97M  6.9T   1% /leofs/data/sdn1
/dev/sdo1            7.3T   97M  6.9T   1% /leofs/data/sdo1
/dev/sdp1            7.3T   97M  6.9T   1% /leofs/data/sdp1
/dev/sdq1            7.3T   98M  6.9T   1% /leofs/data/sdq1
/dev/sdr1            7.3T   97M  6.9T   1% /leofs/data/sdr1
/dev/sds1            7.3T   97M  6.9T   1% /leofs/data/sds1
/dev/sdt1            7.3T   97M  6.9T   1% /leofs/data/sdt1
/dev/sdu1            7.3T   97M  6.9T   1% /leofs/data/sdu1
/dev/sdv1            7.3T   98M  6.9T   1% /leofs/data/sdv1
/dev/sdw1            7.3T   97M  6.9T   1% /leofs/data/sdw1
/dev/sdx1            7.3T   97M  6.9T   1% /leofs/data/sdx1
/dev/sdy1            7.3T   97M  6.9T   1% /leofs/data/sdy1
/dev/sdz1            8.2T   95M  7.8T   1% /leofs/data/sdz1
/dev/sdaa1           8.2T   95M  7.8T   1% /leofs/data/sdaa1
/dev/sdab1           9.1T   84M  8.6T   1% /leofs/data/sdab1
none                 590T  7.3G  590T   1% /datapool
[root@leofs1 leofs1]#
```

Nload Network throughput test

Nload is a real-time network traffic monitoring tool used to verify there are no bottleneck related to the network data transfer both inter and intra cluster.



Test Result

Node	write	read
Node 1	1645MB/s	2088MB/s
Node 2	1693MB/s	2111MB/s
Node 3	1682MB/s	2075MB/s

A LeoFS storage system was built with above 3 nodes of model QuantaPlex T22P-4U with totally 68 disks and deliever total write bandwidth 5020MB/s and read bandwidth 6274MB/s.

Sequential Filesystem Read Test

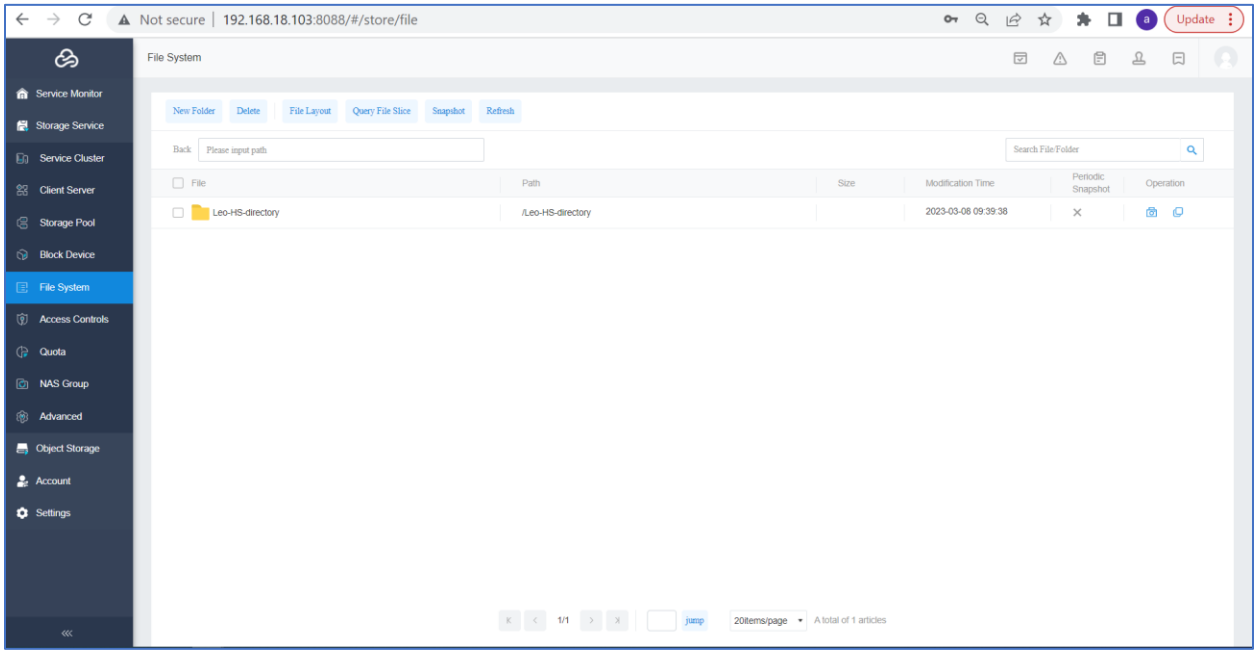
Flexible I/O tool is installed on the host and client to verify the sequential read throughput on the data pool with a block size of 512KB written parallelly with 16 threads simultaneously with an iodepth of 64 requests at a time. The filesystem read bandwidth reached 2.4GB/s which is remarkable for enterprise NAS filesystem storage on HDD with NVMe caching.

```
root@leofs1:/home/leofs1
[root@leofs1 leofs1]# sudo fio --directory=/datapool/Leo-HS-directory/ --direct=1 --rw=write --bs=512k --ioengine=libaio --iodepth=64 --runtime=12
0 --numjobs=16 --time_based --group_reporting --name=throughput-test-job --size=8MB
[root@leofs1 leofs1]# sudo fio --directory=/datapool/Leo-HS-directory/ --direct=1 --rw=read --bs=512k --ioengine=libaio --iodepth=64 --runtime=120
 --numjobs=16 --time_based --group_reporting --name=throughput-test-job --size=8MB
throughput-test-job: (g=0): rw=read, bs=(R) 512KiB-512KiB, (W) 512KiB-512KiB, (T) 512KiB-512KiB, ioengine=libaio, iodepth=64
...
fio-3.19
Starting 16 processes
Jobs: 16 (f=16): [R(16)][100.0%][r=2309MiB/s][r=4617 IOPS][eta 00m:00s]
throughput-test-job: (groupid=0, jobs=16): err= 0: pid=2720302: Wed Mar  8 13:45:31 2023
  read: IOPS=4610, BW=2305MiB/s (2417MB/s)(270GiB/120004msec)
    slat (usec): min=1228, max=208115, avg=3462.94, stdev=1076.38
    clat (usec): min=6, max=451992, avg=218414.25, stdev=26829.50
     lat (msec): min=2, max=455, avg=221.88, stdev=27.21
    clat percentiles (msec):
     |  1.00th=[  169],  5.00th=[  176], 10.00th=[  182], 20.00th=[  188],
     | 30.00th=[  203], 40.00th=[  222], 50.00th=[  228], 60.00th=[  232],
     | 70.00th=[  236], 80.00th=[  239], 90.00th=[  245], 95.00th=[  251],
     | 99.00th=[  262], 99.50th=[  268], 99.90th=[  439], 99.95th=[  439],
     | 99.99th=[  447]
   bw (  MiB/s): min= 1415, max= 2747, per=99.97%, avg=2304.56, stdev= 7.00, samples=3824
   iops        : min= 2827, max= 5493, avg=4608.77, stdev=14.01, samples=3824
  lat (usec)   : 10=0.01%, 20=0.01%
  lat (msec)   : 4=0.01%, 10=0.01%, 20=0.01%, 50=0.03%, 100=0.04%
  lat (msec)   : 250=94.67%, 500=5.25%
  cpu          : usr=0.28%, sys=1.20%, ctx=1073315, majf=0, minf=9703
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=99.8%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.1%, >=64=0.0%
     issued rwts: total=553298,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=64

Run status group 0 (all jobs):
   READ: bw=2305MiB/s (2417MB/s), 2305MiB/s-2305MiB/s (2417MB/s-2417MB/s), io=270GiB (290GB), run=120004-120004msec
```

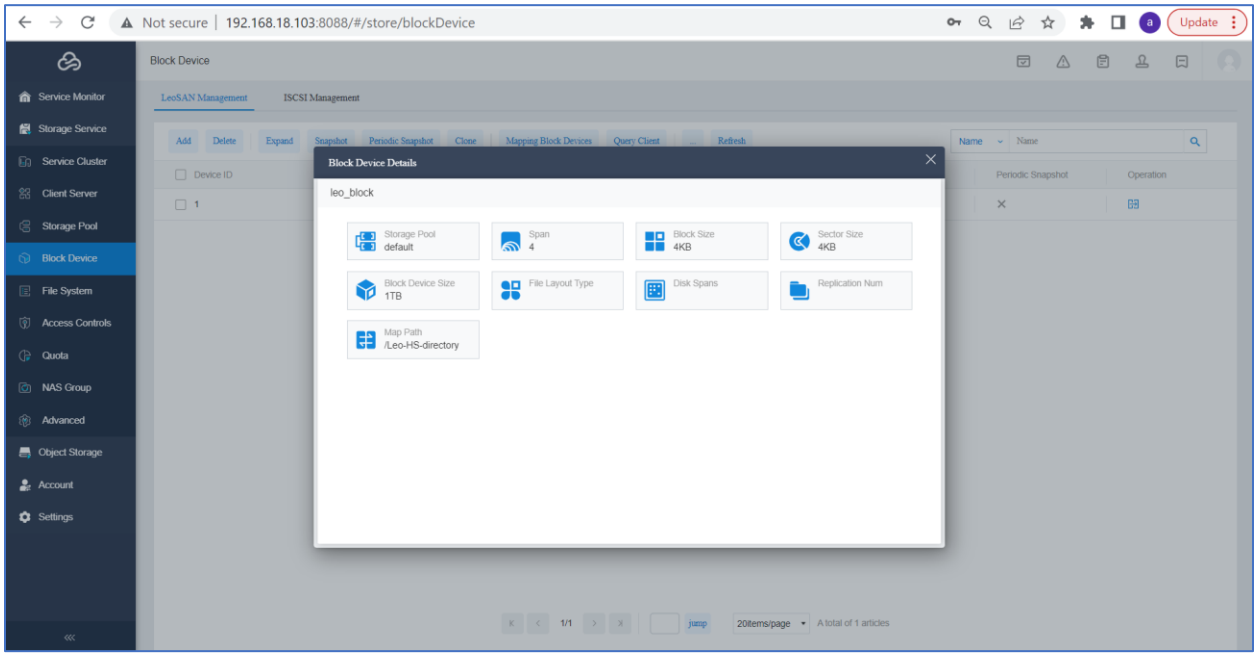
Filesystem Management

The client attached filesystems can be management from a single dashboard providing a seamless operation and management of an enterprise NAS solution.



Block SAN ISCSI

The client storage needs for block device (SAN) exposed via iSCSI is an additional feature provided for enterprise customers requiring data-intensive, mission-critical VM storage and backups.



6 RECOMMENDED HARDWARE

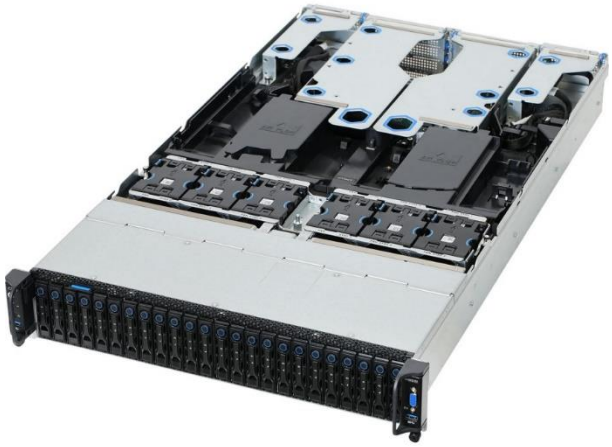
S6P 3.5" | S24P-5U

Support up to 1.8 Petabytes with 84x PMR HDD within 5U form factor. Thermal and RVI optimized mechanical design, future proof for 50TB+ HAMR generation. Flexible option for single node / dual node MB to meet different workload needs. Optimized topology with dual SAS card to achieve enhanced performance. Serviceability built in design for painless data center deployment and maintenance.

S5XQ 2.5" | D53XQ-2U

2 CPU Sockets for up to 80 cores using Intel® Xeon® Platinum 8380 Processor 40cores each. 32 Memory slots for up to 8TB DIMM or Up to 12TB DIMM+PMEM (200 series. Storage Option A - 24 Front Storage drive bays 2.5" hot-plug SATA/SAS. Storage Option B - 16 x SATA or SAS and 8 x U.2 NVMe. Storage Option C - 24 x U.2 NVMe. 2 Rear Storage drive bays 2.5" hot-plug NVMe/SATA/SAS drives (optional). 10 PCIe 4.0 Expansion Slots for Network Interface Cards. 2 Dual width accelerators (GPU or FPGA)

https://www.hyperscalers.com/storage/storage-servers/hyperscalers-S5XQ-D53XQ-2U-ice-lake-densest-hyperscale-server-nvme-drives-buy



S6Q | D54Q-2U

Powered by 4th Gen Intel® Xeon® Scalable processors with up to 350W TDP. PCIe 5 ready, supports up to 400GbE networking bandwidth. DDR5 platform. Up to 2x DW accelerators in a 2U chassis for AI inference workloads. Enhanced serviceability with tool-less, hot-swap designs. NEBS compliant for Telco/5G datacenter deployment. Liquid cooling supported.

https://www.hyperscalers.com/storage/storage-servers/S6Q-D54Q-2U-intel-gen4-sapphire-rapids-pcie5-densest-hyperscale-server-nvme-drives-buy

S5P | T22P-4U

Available with either 1 or 2 server nodes(*1). Extremely high density up to 78 HDDs. Optional 2 hot-swappable NVMe for caching. 1 M.2 on board per node for OS redundancy. Dual SAS path design to release the bottle neck of SAS card.

https://www.hyperscalers.com/storage/storage-servers/S5P-T22P-4U

S2PL | D51PL-4U

102 x 3.5" HDD Storage Capacity. Extreme Computing Performance. Screw-less HDD Tray for Efficient Deployment. Flexible and Versatile I/O Scalability. Proprietary 7mm SSD for Fast Booting.

https://www.hyperscalers.com/storage/storage-servers/hyper-converged-D51PL-4U-densest-storage-server-qct-hyper-scale-buy-hyperscale-quanta

S2P | T21P-4U

Single or Dual Server Node Design. HDD/SSD Individually power off. Lower Operating Costs Screwless HDD/SSD. Record Storage Capacity up to 2.34 PB IN 4U. Flexible networking options up to 12x25G Ports

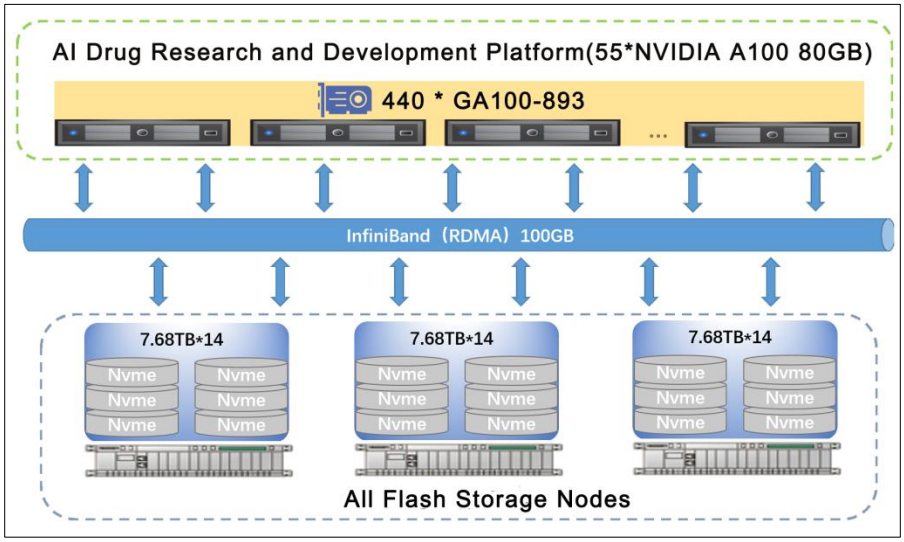
https://www.hyperscalers.com/storage/storage-servers/Quanta-QCT-storage-server-QuantaPlex-T21P-4U-T21P4U-buy-distribution-usa-s2p-tc-hp-apollo-4510-gen10-dell-XA90-super-storage-6047r-compare

LeoFS can be built with almost all storage servers on Hyperscalers website. For interested on high density storage, you can just chose those with as many disk slots as possible. Please note that LeoFS system should be built with all hot swappable disks as online system for production often will have disk defects and disk replacements. The system

must be stopped if disk defects occurred in a storage server with non-hot-swappable data disks or metadata disks. This is something we need to avoid. In addition, no JBOD is recommended.

7 SUCCESSFUL CASES

Successful Case 1 Biopharmaceutics



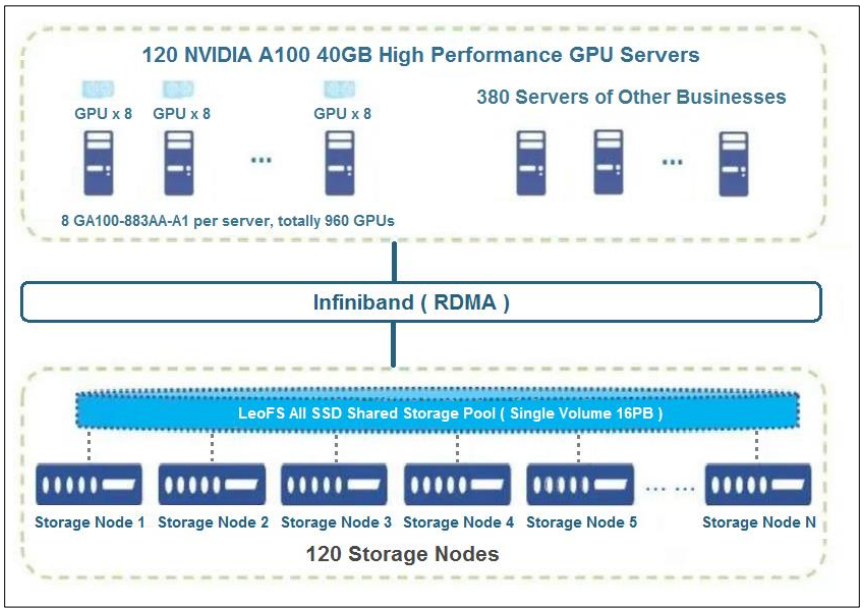
In this project, 3 storage servers are used, each equipped with 14*7.68TB NVMe disks as the unified shared storage system of the AI drug research and development platform, supporting the front-end 55 NVIDIA A100 80GB computing servers with a total of 440 GPU cards with model GA100-893 for deep training and new drug development.

LeoFS Value

- By using RDMA technology and small file aggregation optimization, each NVMe SSD disk in the LeoFS system reaches 4GB/s+ IO bandwidth.
- LeoFS ensures that the performance of a single storage node with only 5 NVMe SSDs to provide an aggregation bandwidth of 20GB/s.
- Supports fast access to tens of millions of small files in the current system, accelerates the process of new drug development.

Successful Case 2 AI Research

System Diagram

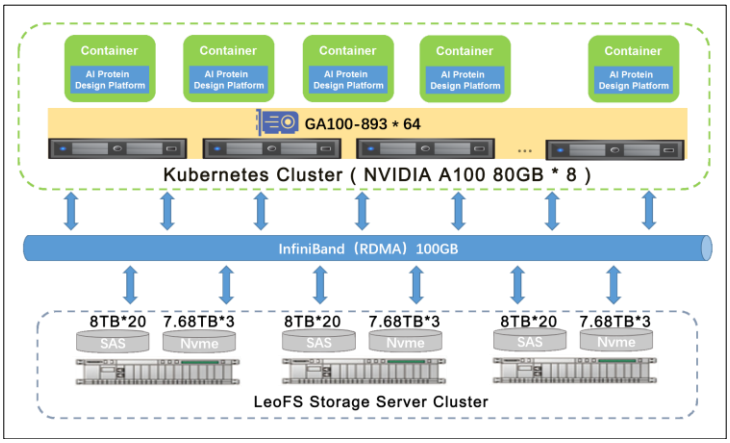


LeoFS Value

A LeoFS with all SSD was built to provide a high-performance data storage base for the AI research institute. The system provides more than 200W IOPS, with a bandwidth of over 100 GB/s, and stores nearly 10 billion files. It supports 500 front-end computing servers, including 120 NVIDIA A100 40GB high-performance GPUs servers with total 960 GA100-883AA-A1 online services.

Successful Case 3 AI Protein Design

System Diagram



LeoFS Value

standard X86 servers, each with 3*7.68TB NVMe + 20*8TB SAS disk drives, are used to build a set of LeoFS system to support Kubernetes cluster and stores tens of billions of small files of tens of kilobytes and provides concurrent access support for deep learning of the AI protein design platform, meets the high bandwidth and low latency small file access characteristics of GPU or CPU computing clusters in AI training. In this project, when

computing resources on 64 GA100-893 GPUs are maximized, LeoFS can still provide stable data access performance even with billions of files, effectively improving training efficiency.

8 REFERENCES

Company Name: Hyperscalers Pty., Ltd.

Address: 10 of 65 Tennant Street Fyshwick ACT 2609 Australia.

Tel: +61 1300 113 112

www.hyperscalers.com

Solutions (hyperscalers.com)

Company Name: LoongStore Technology(Beijing) Co., Ltd.

Address: Room 502, Satellite Building, No 63 Zhichun Road, Haidian District, Beijing China.

Tel: 400-803-6006

LeoFS_Installation_Guide.docx

LeoFS_WhitePaper.docx

www.leofs.info

https://leofs.snazzydocs.com/

http://www.loongstore.com.cn

Index