



SDC¹⁸

September 24-27, 2018
Santa Clara, CA

www.storagedeveloper.org

High Performance Storage for Data Intensive Workloads

Bo Zhang
SANS Digital Technology Inc.

Legal Disclaimer

The content of this presentation is confidential and intended for business review only. The author and SANS Digital Technology Inc. disclaim any liability in connection with the use of the information herein.

The information in this presentation is true and complete to the best of our knowledge, it is provided “as is” without warranty of any kind. All recommendations are made without guarantee on the part of the author or SANS Digital Technology Inc.

We do not accept any responsibility or liability for the accuracy, content, completeness, legality, or reliability of the information contained in this presentation.

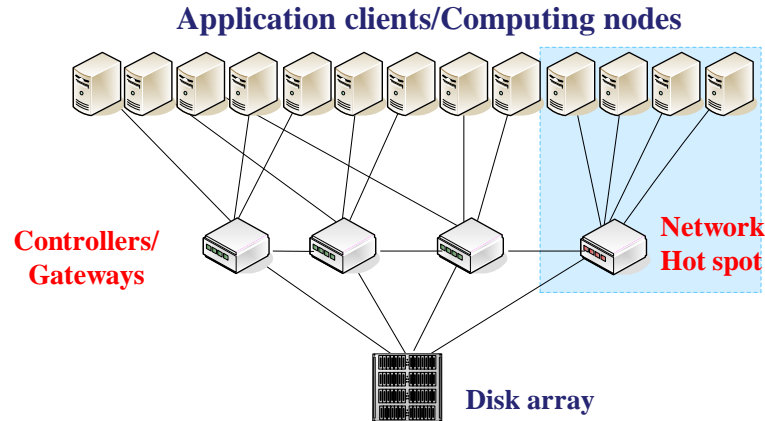
Proven Proprietary File System

- ◆ **Market proven since 2007 with over 500 PB and 700 customers***
- ◆ **No controller/gateway bottlenecks, solution guarantees to saturate hardware throughput**
- ◆ **Cluster of metadata pairs, unique setup with active-active**
- ◆ **File-based network RAID enables high aggregated bandwidth**
- ◆ **One cluster for file, block, and object storage**
- ◆ **Easy scalability with single server solution**

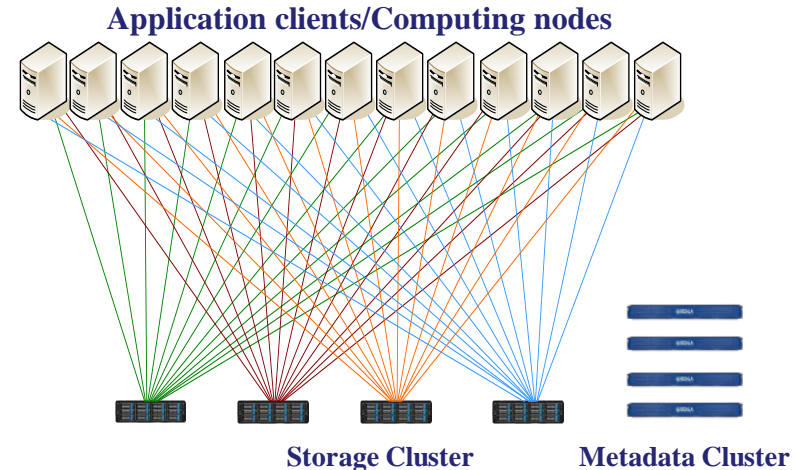
* Total aggregated numbers of capacity and customers with the underlying technology of SCALA Storage.

No Bottleneck from Controller/Gateway

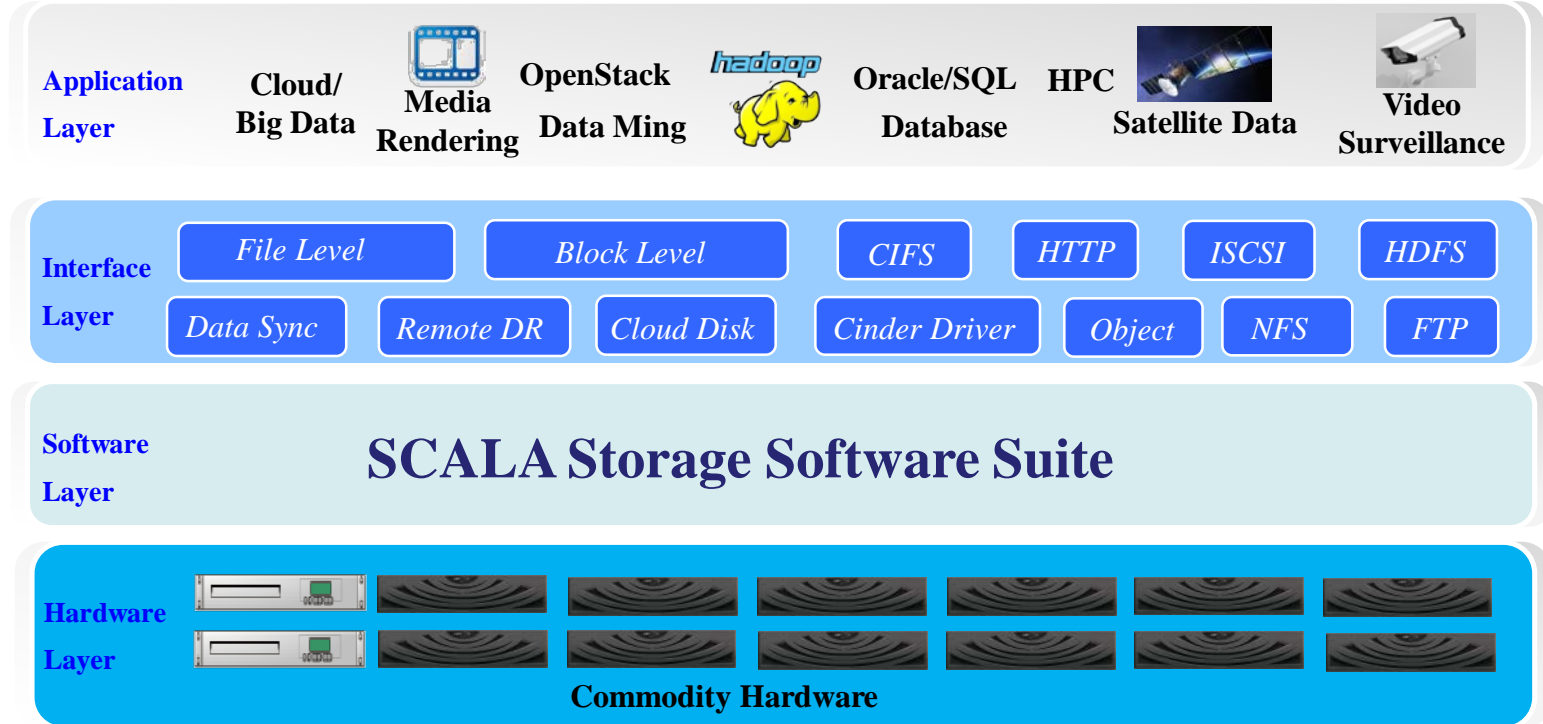
Typical SAN or NAS solutions
Either controllers or network can easily create
Throughput bottleneck



SCALA not having controllers or gateways, all
applications clients/computing nodes communicate
with all storage nodes in the cluster

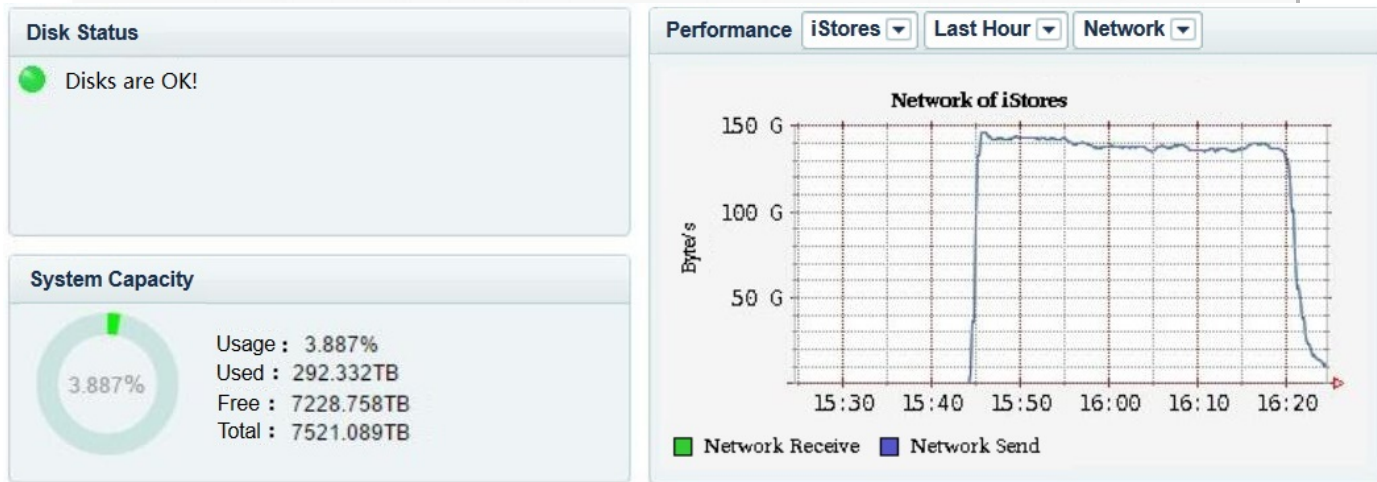


Unique Solution



Solution Clusters: Clients, Storage, and Metadata

- High aggregated throughput using SATA or SAS HDDs
- Customer on-site 7.5 PB, close to 150 GB/s aggregated throughput

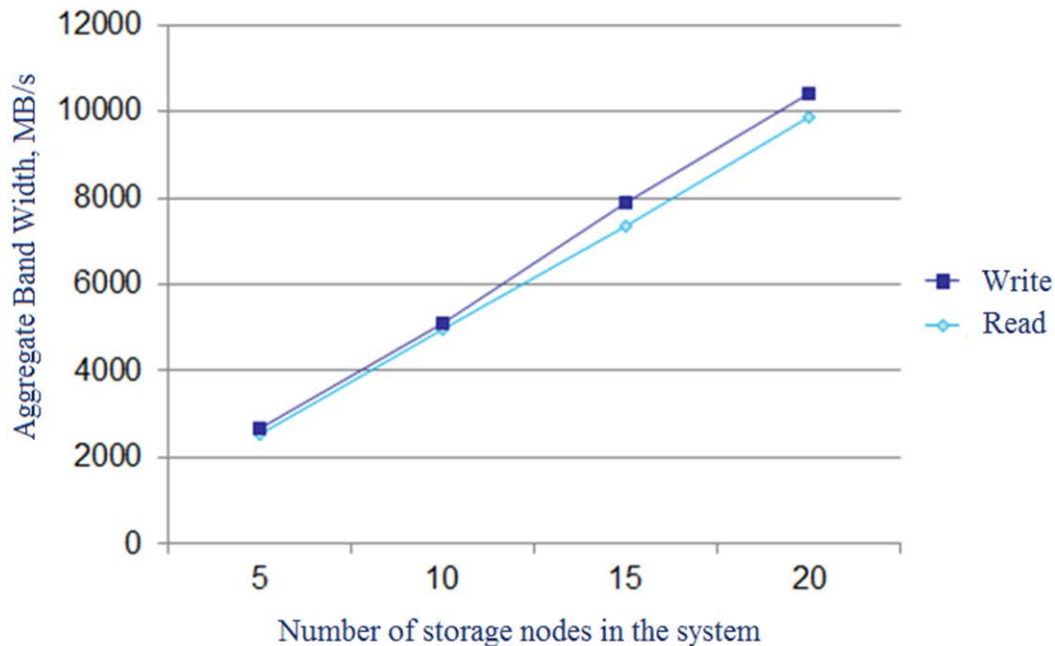


- ❖ 102 nodes of 24-bay server, each has 19 or 20 x 4 TB SATA
- ❖ Total 1,880 HDDs, average write 75 MBps/drive

(Up to 128 servers)

Performance Increases When Adding Storage Nodes

Storage node: 2U 12-bay commodity server; 10 GbE; SATA disks; 45~75 MBps/drive



Throughput vs. Lustre and GPFS

- Lustre/GPFS*: six LUNs, each chassis with 30 x 4TB disks, total of 180 HDDs
 - **Limitation seen on 8 clients, 1 stream/client**
 - Write saturated controllers (~1.2 GB/s per controller), 4-5 GBps
 - Read limited by network interfaces, 5-7 GBps
- SCALA: 8 nodes of 24-bay storage servers, total of 192 HDDs
 - **8 clients, dual 10 GbE, 20 streams/client**
 - Write: 11 GBps
 - Read: 7 GBps



* Lustre and GPFS numbers are from CERN's presentation on High Performance Storage in Science, SDC 2017

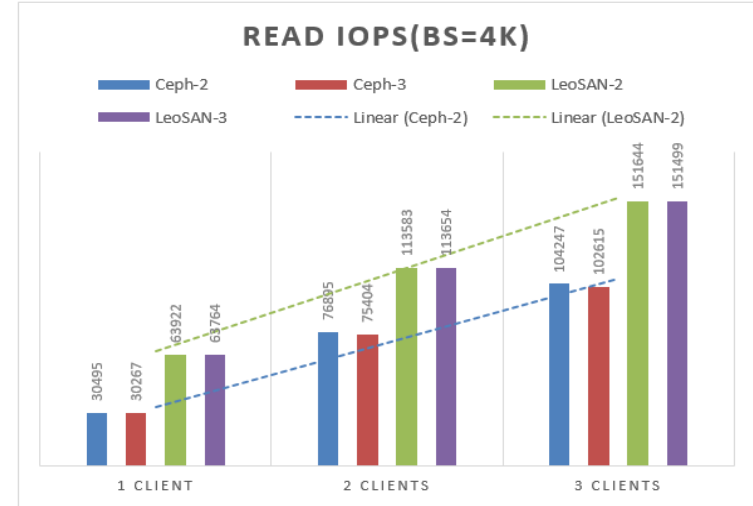
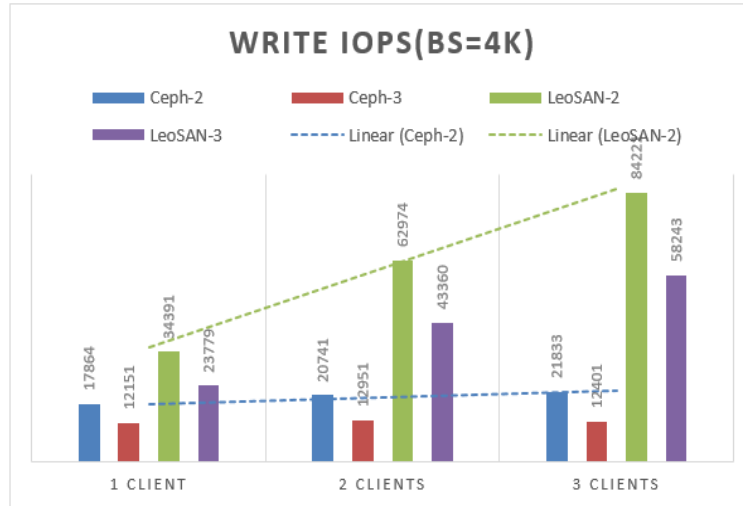
Seismic Data Processing Test vs. StorNEXT*

	FC SAN	SCALA
Hardware	32 x 8 GB FC Controllers 512 x 300 GB FC disks	17 storage nodes, 2 metadata 267 x 750 GB SATA HDDs
File system	StorNEXT FS	SCALA FS
Network	1GbE	1GbE
Clients	100 Linux Blade Servers	100 Linux Blade Servers
File sizes	32KB / 256KB / 1MB	32KB / 256KB / 1MB
IOzone concurrent number	180	180
Read	2.4 GB/s	4.2 GB/s
Write	1.9 GB/s	4.7 GB/s

* Customer on-site test done in 2009. Since then, customer has scaled from 200TB to over 3 PB, changed 1 GbE to 10 GbE.

Block Device SSD IOPS vs. Ceph

- System: 3 nodes of 12-bay storage servers, each with 12 x 240 GB SSD
- Dual 10 GbE, 3 clients
- Replications: 2 and 3, block size 4 KB



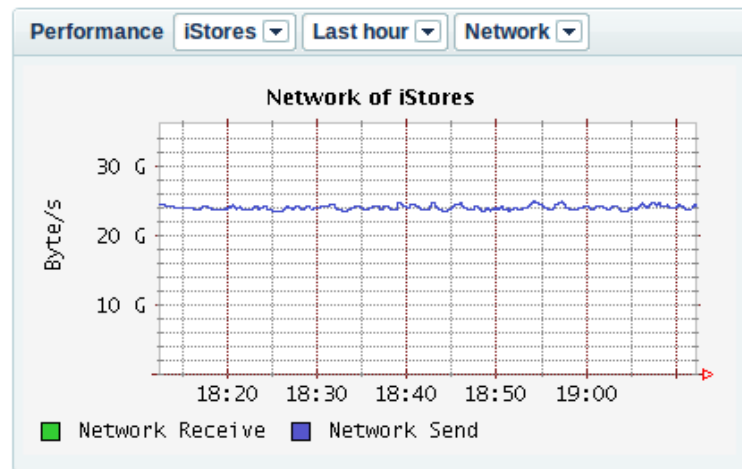
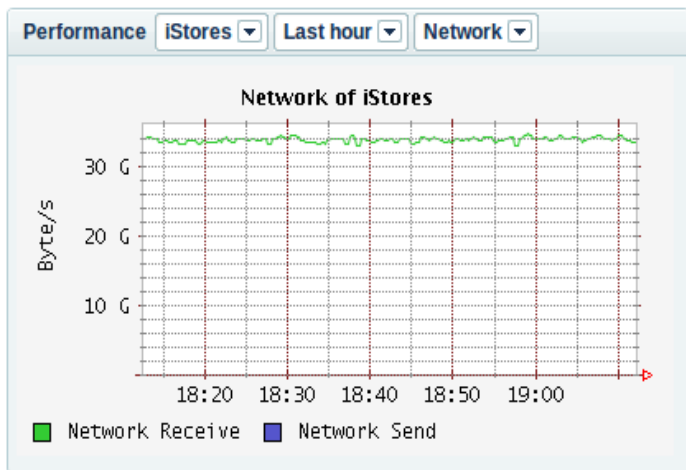
File Accessed Performance

- System with 3 storage nodes and 2 metadata servers, 1 application client
- Creation of 20 million small-size files, each 10 KB
- Test the system response time on 20 million files with random accesses

Response time (ms)	File accessed	% of Total
0-0.3	19,135,123	91.3%
0.3-0.6	625,588	3.1%
0.6-1.0	489,139	2.3%
1.0-5.0	619,896	3.0%
5.0-10.0	72,625	0.3%

Genomic Customer On-site Performance

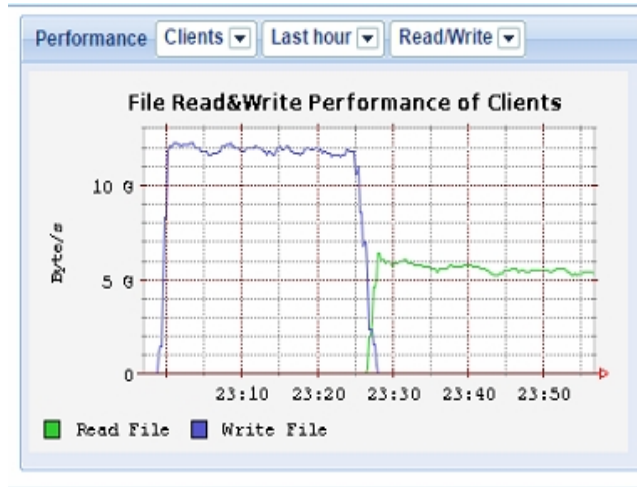
- System with 31 storage nodes of 3U/16 drives and 4 metadata servers, 10 GbE
- With IOzone, clients consist of 190 blade servers, simulate 200 TB data
- Aggregated write of 30 GBps, aggregated read of 24 GBps



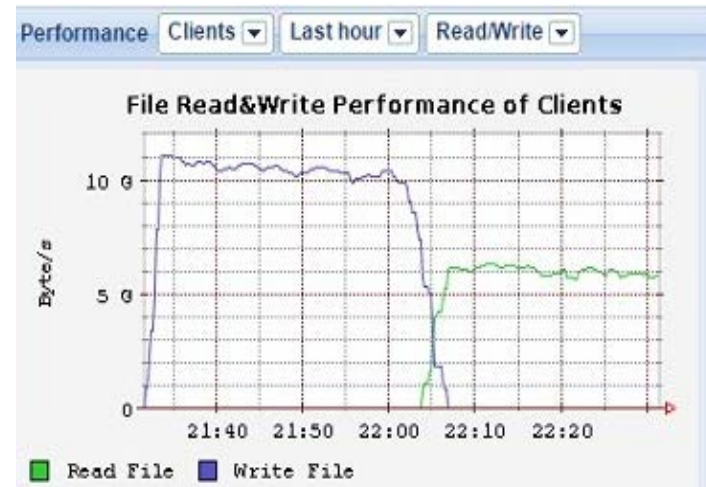
Video Surveillance 960TB System

- 10 nodes of 2U 12-bay storage servers, each has 12 x 8 TB SATA disks
- 2 metadata servers, 10 GbE network
- Aggregated throughput: write 12 GB/s, read 6 GB/s

4+1 replication performance:



2+1 replication performance:



Throughput Variables

- ◆ **Use case applications**
 - High-performance Computing: bandwidth/throughput
 - Data base and VMs: IOPS
 - Media and Entertainment: latency
- ◆ **Cluster hardware**
 - Storage servers: 12-bay, 16-bay, 24-bay, or 36-bay
 - Hard drives: SATA, performance SAS, or SSD
 - Network: 10 GbE, 100 GbE, or Infiniband
- ◆ **Constraints**
 - Budget
 - Technical and management expertise

Our Learnings

- ◆ **Using HDDs in storage nodes, 10 GbE network is sufficient**
 - A 12-bay server, one 10 GbE
 - A 24-bay or 36-bay server, dual 10 GbE
 - Single drive performance ≤ 80 MBps
- ◆ **Using SSDs only in metadata nodes**
 - Metadata disk operations per second: read 20,000 write 10,000
 - To increase IOPS, simply increase the number of metadata disks
- ◆ **CPU, RAM, Latency and Redundancy**
 - Depending on disk number of storage node, Intel E5-2620 or higher
 - RAM: 12-bay 16 GB, 16-bay 24 GB, 24-bay 32 GB, 36-bay 48 GB
 - Latency: consider using SSDs in storage nodes (4K video rendering)
 - Redundancy can be optimized for capacity and performance

Sample Hardware Specifications

◆ Storage nodes

- Dual Intel E5-2620 v4
- Dual 10 GbE (40 GbE, 100 GbE or IB)
- 64 GB RAM, 1 SSD OS drive
- 2U 12-bay or 3U 16-bay or 4U 24-bay or 4U 36-bay
- 2 TB to 10 TB 7200 enterprise SATA or SAS HDDs

◆ Metadata nodes

- Dual Intel E5-2620 v4
- Dual 10 GbE
- 64 GB RAM, 1 SSD OS drive
- Dual 480 GB enterprise SSD
- Option to be put into storage nodes

Thank You!

Welcome to Q & A

bozhang@scalastorage.com