National Institute of Advanced Industrial Science and Technology
AIST

Open Source Grid and Cluster Conference
Oakland, CA, May 12-16, 2008

Comparison of Gfarm- and Lustre- based Storage System

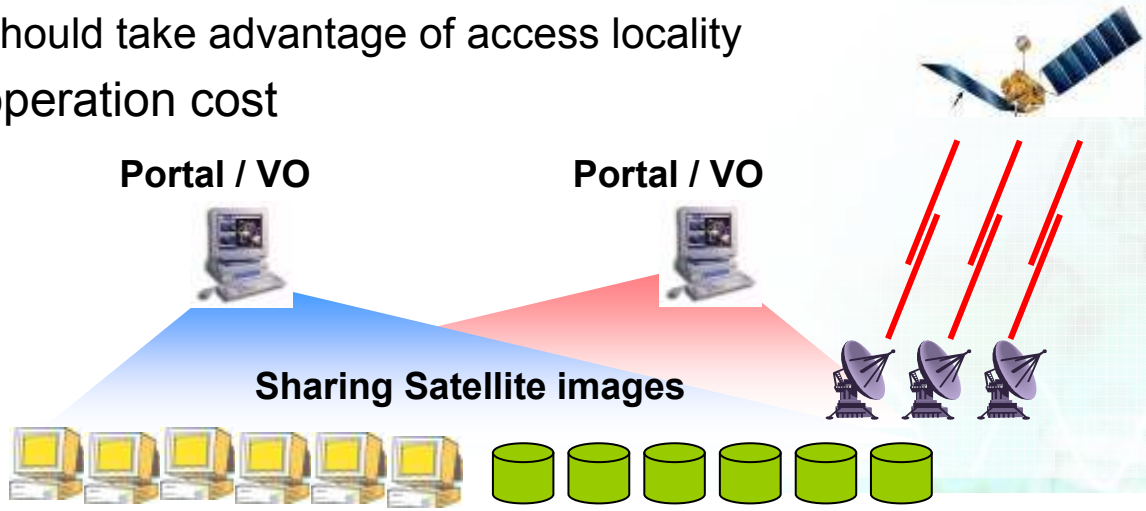
From Geosciences Applications Perspective

Yusuke Tanimura
National Institute of AIST, Japan
yusuke.tanimura@aist.go.jp

National Institute of Advanced Industrial Science and Technology
AIST

Background (1)

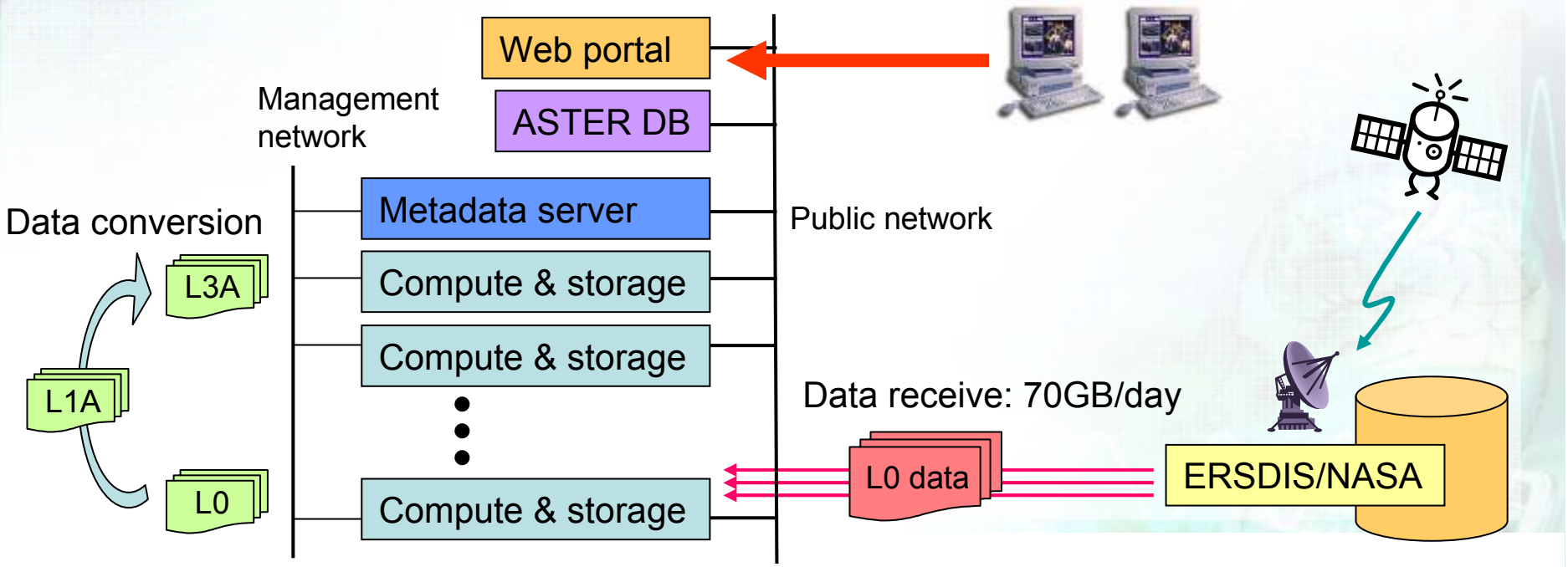
- Global earth observation in geosciences
  - Accumulate knowledge about the earth in various forms and understand the earth scientifically.
- Requirement for building a large-scale data repository:
  - Online access from anywhere at anytime
  - From Hundreds' TB to PB scale capacity
    - No data lost
    - Highly available service
  - Performance scalability to concurrent data access
    - Should take advantage of access locality
  - Low operation cost



National Institute of Advanced Industrial Science and Technology
AIST

Background (2)

- AIST operates a storage system for ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer).
  - 153 TB (15 millions' files) data has been stored in July, 2007.
  - Gfarm-based storage system
    - 1 metadata server and 4 metadata cache servers
    - 24 compute & storage servers
      - Each node has 7 TB disk space with 16 drives by RAID-6.



National Institute of Advanced Industrial Science and Technology
AIST

Research goal

- We need a larger storage system for our near future.
  - Next generation sensor will produce more than 1 PB data.
  - How to build a larger storage system?
    - Performance?
    - Cost?
      - Prefer using free software and commodity hardware
- Goal: Reveal the best storage system for this application
  - Pick up open source parallel filesystem: Gfarm and Lustre
    - Gfarm is in use for the ASTER storage system.
    - Lustre is widely used in HPC clusters.
  - Evaluate two storage systems with real data processing.
    - Import about 100 TB data into both systems.
  - We focus on not only performance but also operation cost.

National Institute of Advanced Industrial Science and Technology
AIST

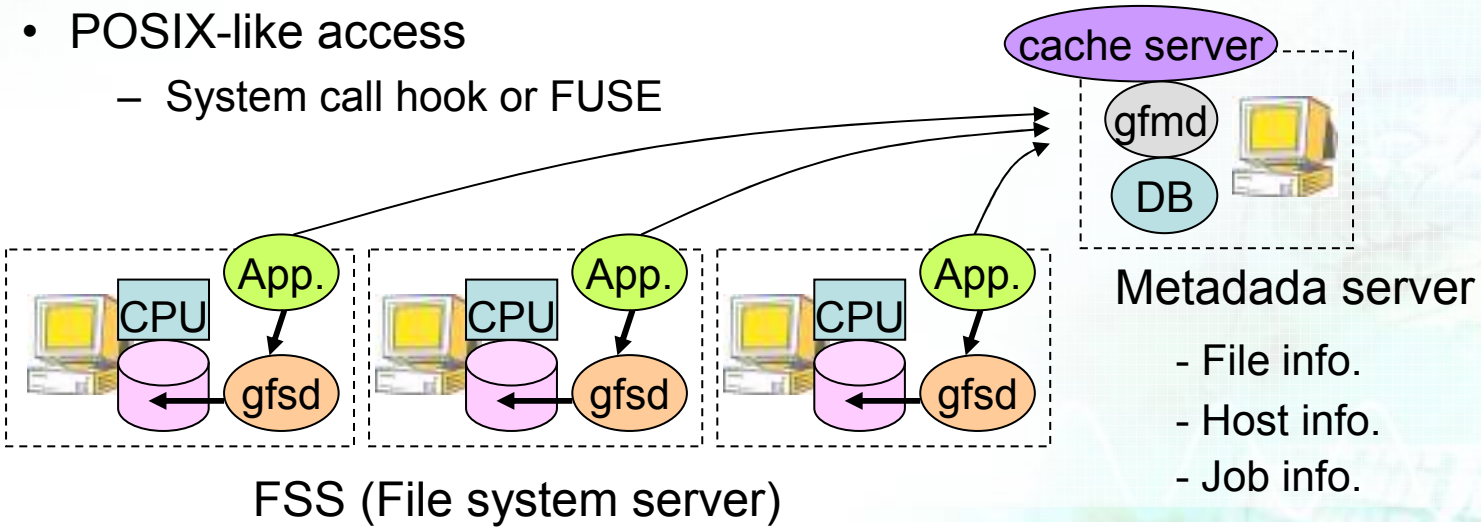
Storage components

- Storage server hardware
  - We use Sun Fire X4500 (Thumper) due to its fairly attractive architecture.
    - 24TB capacity by 48 hard disk drives
      - If 16TB is available for application data area, the total capacity will be 1PB with 64 nodes.
    - No RAID controllers but 6 SATA controllers
- Software (Parallel file system, OS, underlying file system)
  - Gfarm
    - Gfarm works with Solais and ZFS.
      - ZFS is very reliable without hardware RAID controller.
  - Lustre
    - Because Lustre does not work with Solaris and ZFS, we use Linux and Ext3 (LDISKFS).

National Institute of Advanced Industrial Science and Technology
AIST

Introduction of Gfarm

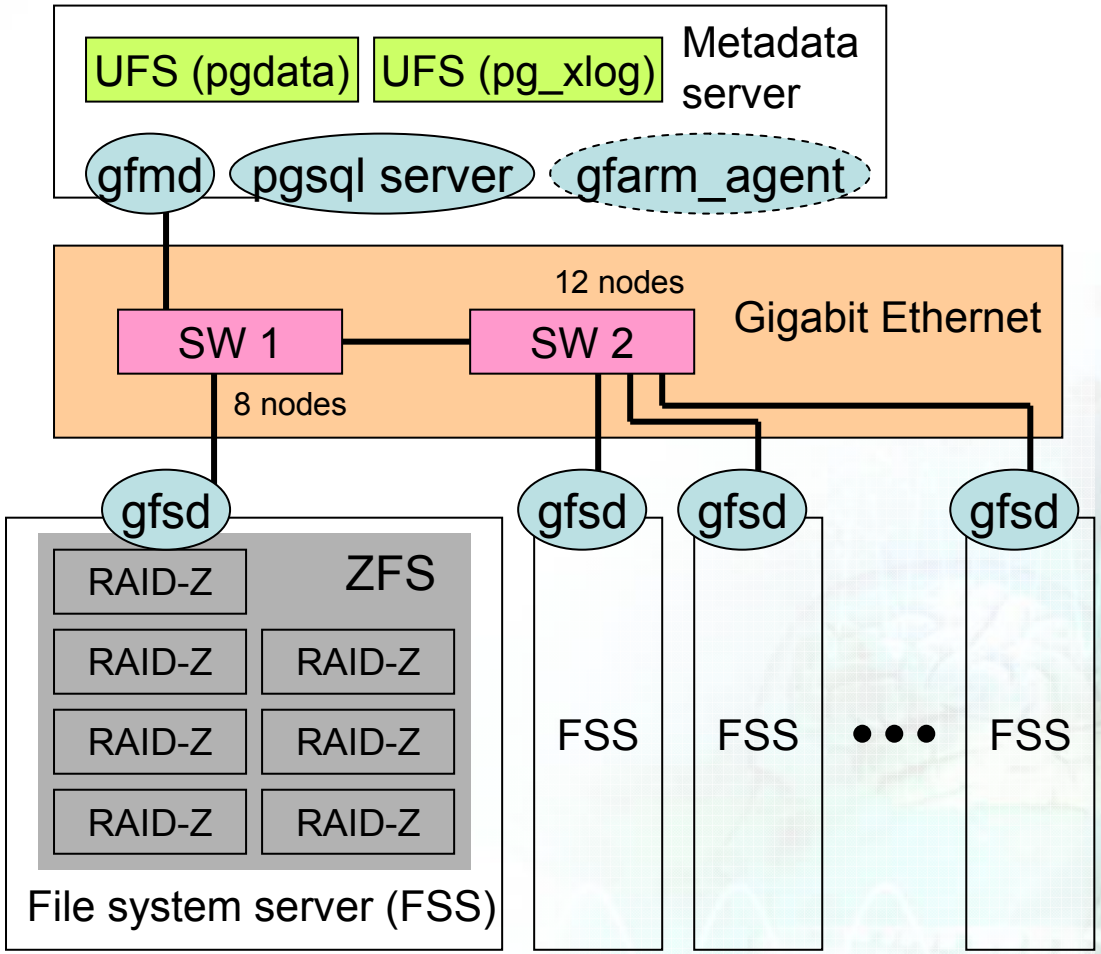
- Gfarm
  - Open-source software
    - Originally developed by AIST, and now maintained in SourceForge
  - Parallel filesystem consisting of local disks of PCs
    - Global namespace
    - File replication
    - Job scheduling based on file location
    - GSI authentication
    - POSIX-like access
      - System call hook or FUSE



National Institute of Advanced Industrial Science and Technology
AIST

Gfarm-based storage (1)

- 1 metadata server
- 19 storage servers
- 256.5 TB (13.5 x 19)
- 76 CPU (4 x 19)
- Hardware
  - X4500 x 20
  - CenterCOM GS924S
- Software
  - Solaris 10 (Update3 with Recommended patch)
  - Modified Gfarm v1.4.1
- Parameters
  - Enabled write caching of SATA disks
  - Default values in Gfarm, PostgreSQL, and ZFS



National Institute of Advanced Industrial Science and Technology
AIST

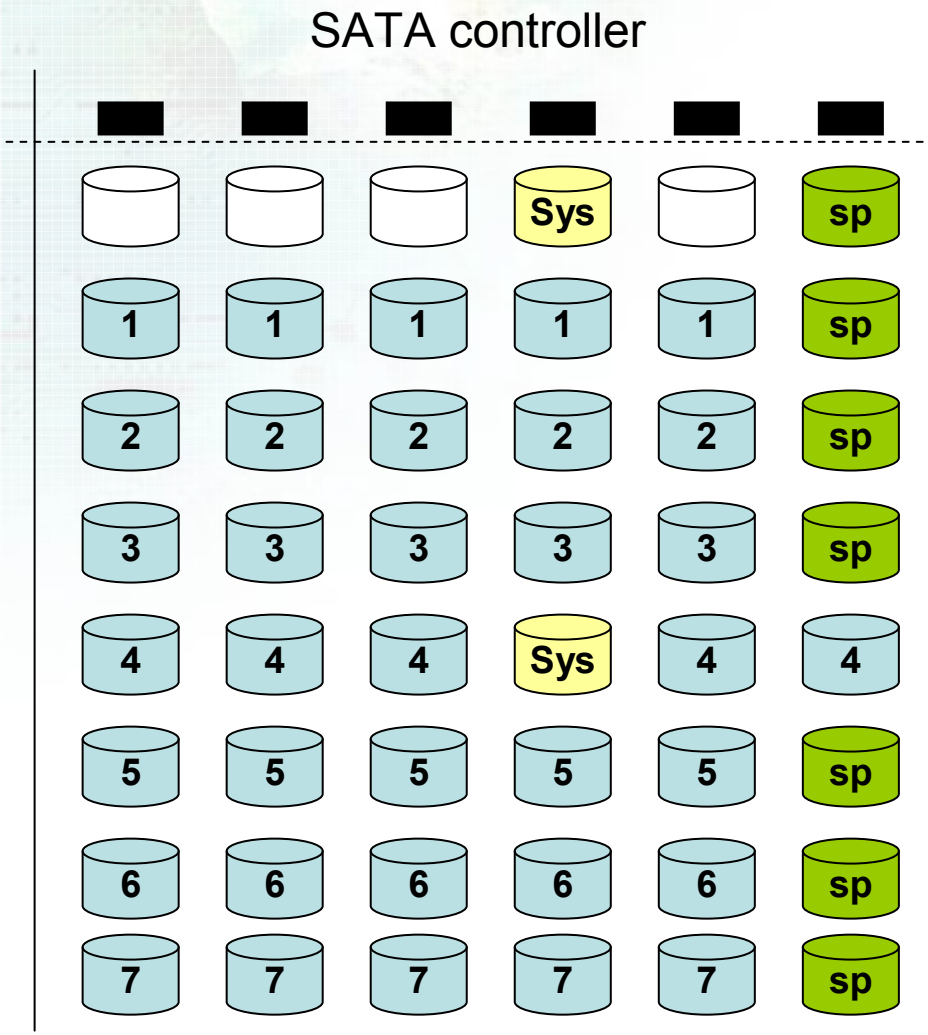
Gfarm-based storage (2)

- Configuration of metadata server
  - Gfarm uses PostgreSQL as a backend database.
    - Use PostgreSQL v8.1.9
    - Put pg_xlog on a dedicated disk
    - Put PGDATA without pg_xlog on a dedicated disk
    - Use UFS with nonforcedirectio
  - Put PGDATA on UFS or ZFS?
    - Tried to tune ZFS according to ZFS Best Practice Guide
      - Limit the ARC (Adaptive Repleacement Cache) size
      - Set ZFS recordsize=8K, and etc.
    - UFS showed better metadata ops. performance than ZFS.

	pgbench [tps]	metadata ops. [sec]
UFS	392, 751, 826	166
ZFS	91.6	869
Tuned ZFS	106	838

National Institute of Advanced Industrial Science and Technology
AIST

Gfarm-based storage (3)



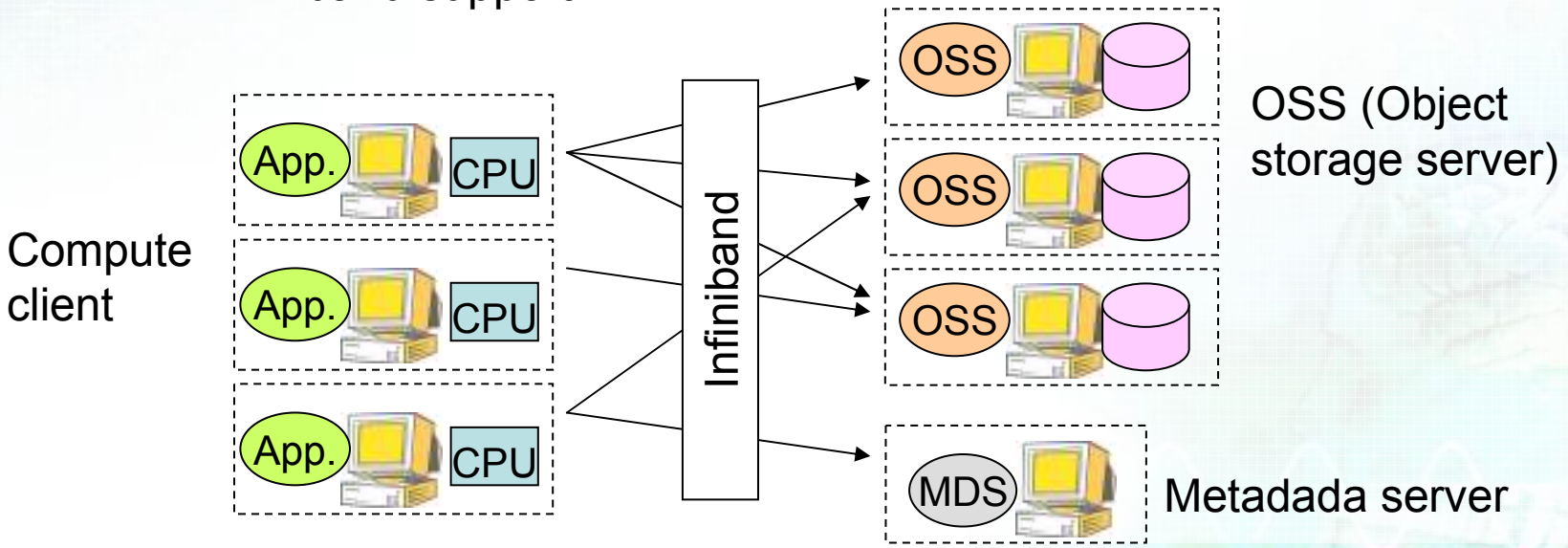
Disk Layout of X4500

- System area: 2 disks with mirroring
- Data area (2TBx7): 7 partitions
  - Each partition consists of 5 disks. One is for parity.
- Spare: 7 disks (green)
- No use: 4 disks (white)
- Gfarm and Lustre take similar configuration.
  - Gfarm:
    - RAID-Z by ZFS
    - 7 partitions are integrated to 1 ZFS pool.
  - Lustre: Software RAID-5

National Institute of Advanced Industrial Science and Technology
AIST

Introduction of Lustre

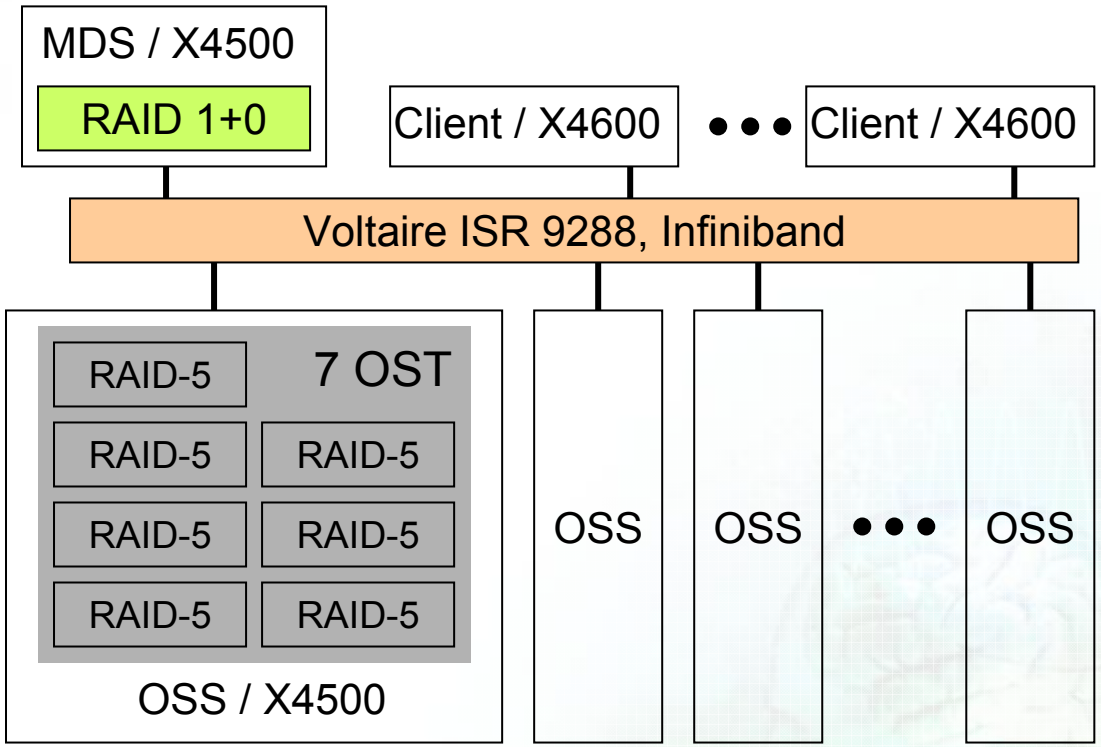
- Lustre
  - Open-source software
    - Developed and maintained by Sun Microsystems
  - High performance cluster file system
    - POSIX-compliant
    - Object storage
    - Infiniband support



National Institute of Advanced Industrial Science and Technology
AIST

Lustre-based Storage

- 1 metadata server
- 10 storage servers
  - 135 TB (13.5 x 10)
- 16 compute servers
  - 256 CPU (16 x 16)
- Hardware
  - X4500 x 11
  - X4600 x 16
  - Voltaire ISR 9288
- Software
  - Linux (RedHat, SuSE)
  - Lustre v1.6.2 release
- Parameters
  - Disabled write caching of SATA disks
  - Use striping: Count is 3 and size is 2MB.



National Institute of Advanced Industrial Science and Technology
AIST

Benchmarks

- What we measured?
  - Performance of concurrent access from multiple clients
- Basic benchmarks
  - I/O intensive benchmark
    - Write/read a large file (More than 10 GB)
    - It shows throughput (MB/sec).
  - Metadata intensive benchmark
    - O_CREAT+close(), O_RDONLY+close(), and unlink() operations
    - It shows metadata operations speed (ops/sec).
- Practical application benchmark
  - Use real application program (DTMSOFT) with real data sets.
  - It shows execution time (sec).

National Institute of Advanced Industrial Science and Technology
AIST

Throughput of a single node

- First, we measured basic performance of the disk resources without Gfarm and Lustre.
- Result:
  - ZFS/RAID-Z achieved significant performance benefit by ZFS dynamic striping.
  - LDISKFS/RAID-5 was affected by overhead of Linux's Software RAID-5.

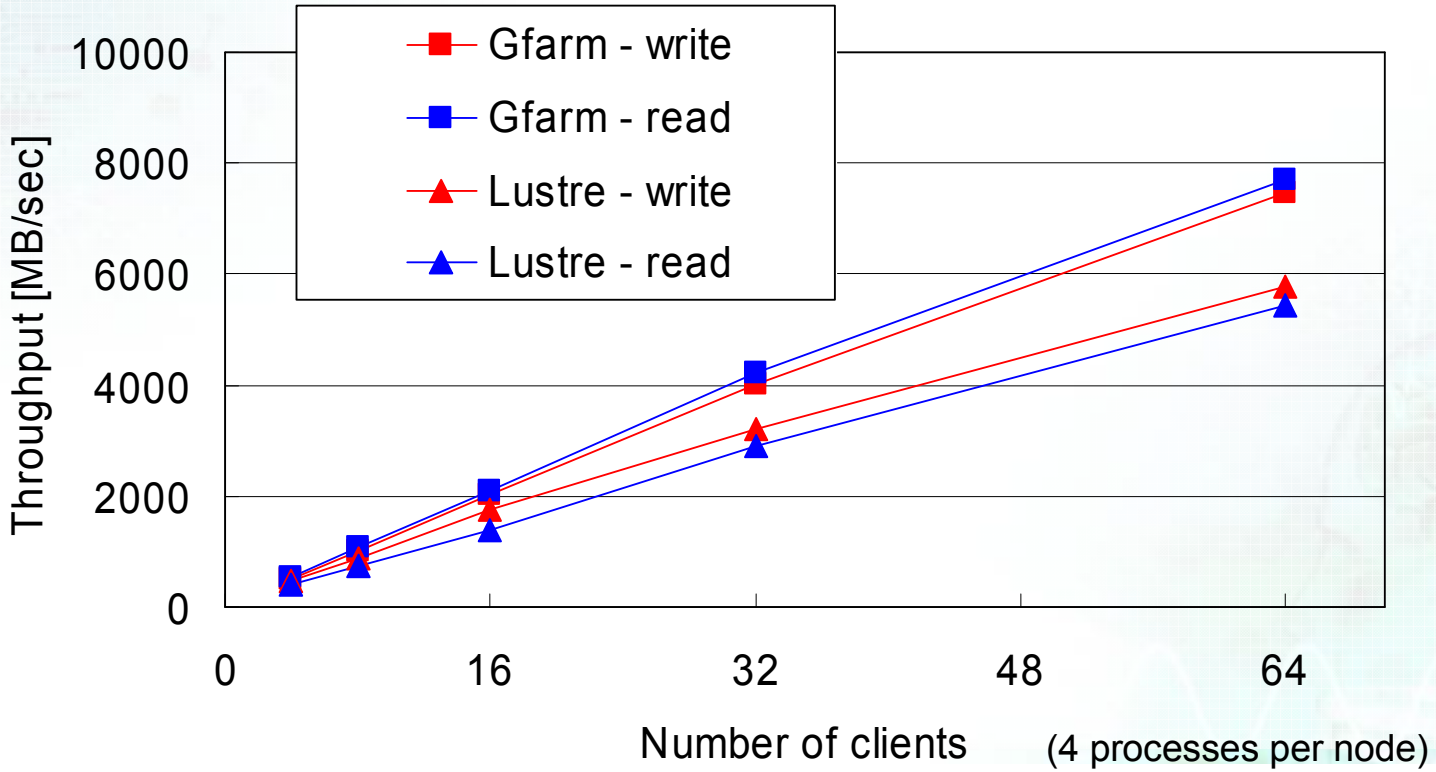
# clients	ZFS/RAID-Z		LDISKFS/RAID-5	
	Write	Read	Write	Read
1	451	701	97	188
2	602	849	184	337
4	593	723	338	605
7	664	777	448	680

Unit: MB/sec

National Institute of Advanced Industrial Science and Technology
AIST

Aggregated throughput

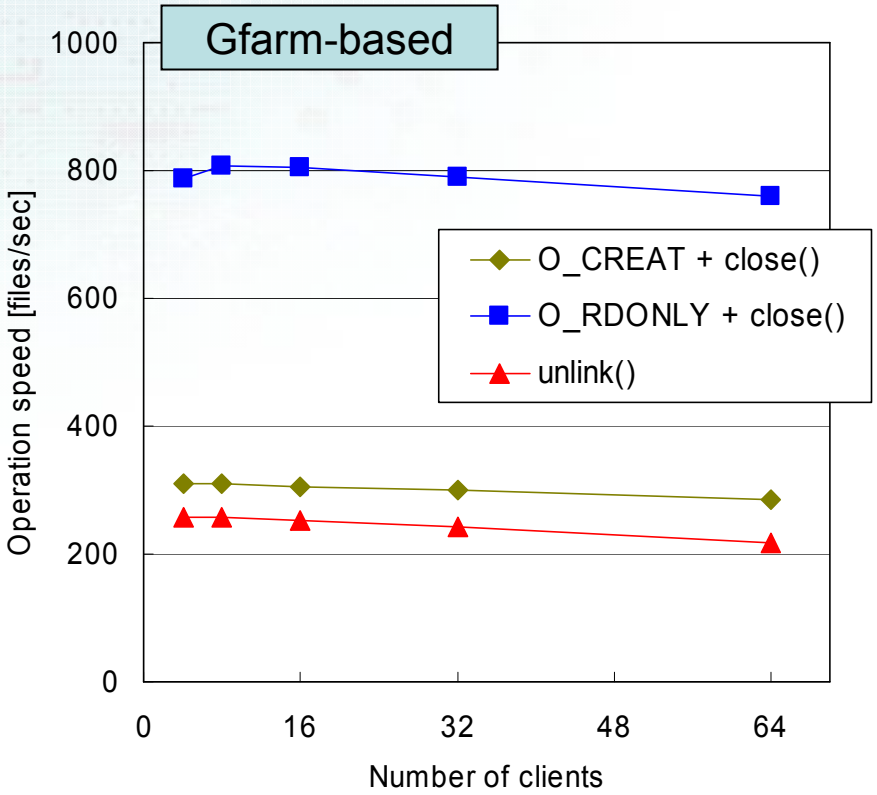
- Next, we measured aggregated throughput of two storage systems.
- Result:
  - Both storage systems achieved scalable performance.
  - Gfarm-based showed slightly better performance than Lustre-based.

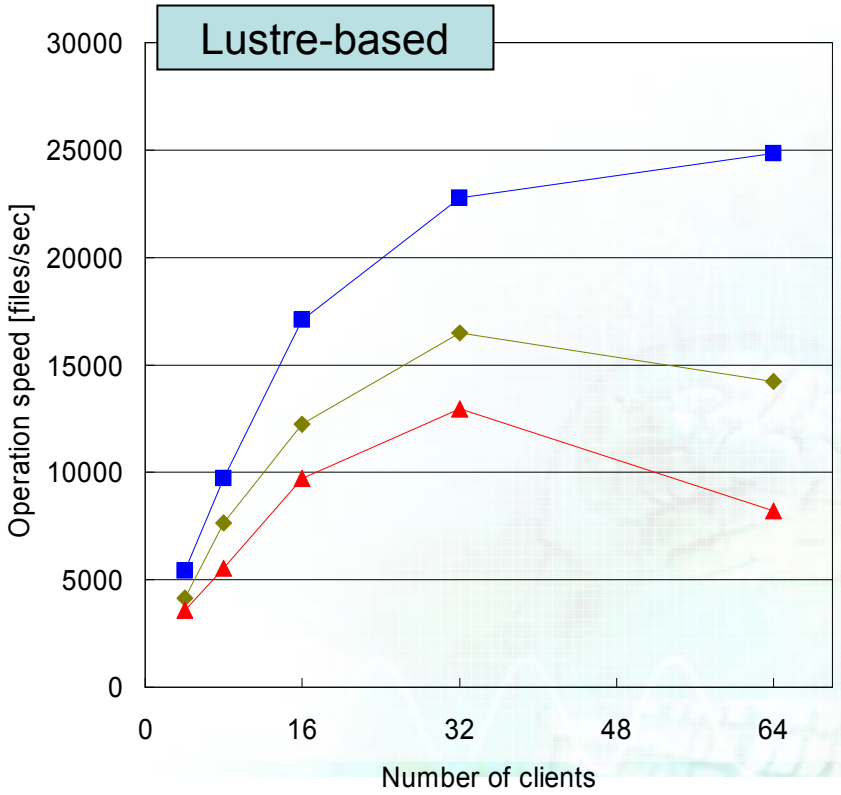


National Institute of Advanced Industrial Science and Technology
AIST

Aggregated metadata ops.

- Next, we measured aggregated metadata operations speed.
- Result:
  - Metadata operations speed is very slow in Gfarm-based.
  - Lustre-based's performance increases until 32 clients.





National Institute of Advanced Industrial Science and Technology AIST

Application benchmark (1)

- DTMSOFT
  - Most frequently submitted jobs in the ASTER storage system
  - We concerned performance of concurrent metadata operations.
    - Each job copies the data from the Gfarm-based to a local disk, run DTMSOFT, and return outputs to the Gfarm-based.
    - DTMSOFT directly access the data on the Lustre-based.
- Experiment
  - Measured execution time of processing 3137 data (311 GB) by DTMSOFT, when about 75% of the storage capacity is full.
  - Our experiment environment:
    - In Gfarm-based storage system, 73% capacity is used.
      - 173 TB by 18 millions' files are stored.
    - In Lustre-based storage system, 78% capacity is used.
      - 93 TB by 9.5 millions' files are stored.

National Institute of Advanced Industrial Science and Technology
AIST

Application benchmark (2)

- Result:
  - The Lustre-based showed performance scalability.
  - By minimizing metadata operations, the Gfarm-based also showed similar performance to the Lustre-based.
  - Note:
    - Speed-up was calculated from sequential execution to process 30 data.
    - Speed-up of Lustre-based is high due to the AMD PowerNow effect.

	(#pe x #node)	Exec. time	Speed-up
Gfarm	4x16	40310 [sec]	1.01
Lustre	4x16	39579	1.15
Lustre	16x 4	33295	1.37
Lustre	16x16	9330	1.22

National Institute of Advanced Industrial Science and Technology
AIST

Operation cost

- Installation
  - Purchase cost
    - We do not want to discuss about it here but software is free.
  - Work cost for system setup (and test)
- Operation
  - Work cost (Possibly employing system engineers)
    - Work for faults
    - Daily maintenance
    - Work for software update / change of configuration
  - Purchase cost of replacing broken parts

National Institute of Advanced Industrial Science and Technology
AIST

High Availability

- High availability is highly expected by geosciences user-side.
  - New data comes everyday.
- Fault tolerance & automatic recovery are necessary.
  - File replication
    - Set preferable replication level to each file.
    - Not suitable to frequently updated data
  - Failover
    - Combination with fault detection tool such as Heartbeat
    - Need to prepare both active and inactive resources.

Possible configuration for fault tolerance

Fault items	Gfarm	Lustre
Storage node	File replication	Failover
Disk on storage node	File replication, RAID	RAID
Metadata server node	PgPool for PostgreSQL	Failover

National Institute of Advanced Industrial Science and Technology
AIST

Maintenance work issues

- Similar functionalities
  - For scheduled maintenance
    - By notifying shutdown of the storage node to the metadata server, we do not have to stop the entire system.
  - For data migration at addition/exchange of storage nodes
    - Need manual operation by a set of commands.
- Differences
  - Need performance tuning for the Gfarm metadata server.
  - Need kernel patch for the Luster nodes.
  - Flexibility of the system enhancement
    - Luster: Storage nodes and compute nodes (clients) are separated.
    - Gfarm: Storage nodes and compute nodes are same.

National Institute of Advanced Industrial Science and Technology
AIST

Discussion

- A factor of choosing either storage system is not performance but operation cost.
  - By minimizing metadata access, the Gfarm-based could achieve similar performance to the Luster-based.
  - Many differences in operation between Gfarm and Luster
    - It comes from design concept and software maturity.
- Note that the following was excluded in this comparison.
  - Local optimization is sometimes not easy to apply.
  - Infiniband is not cheap.

National Institute of Advanced Industrial Science and Technology
AIST

Summary

- Introduction of a geosciences application
  - A large storage system is required for satellite data archiving, analyzing, and publishing.
    - Want to build it with free software and commodity hardware.
- Comparison between the Gfarm-based and the Lustre-based storage system
  - Used real application and real data set.
    - Stored about 100TB data in each system.
    - Examined concurrent access by 64 clients.
  - Performance scalability is ensured at the scale of 93-173 TB data and concurrent access from 64-256 clients.
  - Operation cost is different in several points.

National Institute of Advanced Industrial Science and Technology
AIST

More info. about Gfarm

- Grid Data Farm project:
  - http://datafarm.apgrid.org/
- Gfarm v2 in SourceForge:
  - http://sourceforge.net/projects/gfarm/
- Gfarm v1 roll for Rocks v4.2.1:
  - Please email to yusuke.tanimura@aist.go.jp.

National Institute of Advanced Industrial Science and Technology
AIST

Acknowledgement

This work was performed in collaboration between National Institute of AIST and Tokyo Institute of Technology.

- National Institute of AIST
  - Yusuke Tanimura
  - Naotaka Yamamoto
  - Yoshio Tanaka
  - Satoshi Sekiguchi
- Tokyo Institute of Technology
  - Takeshi Nishikawa
  - Takeshi Yamanashi
  - Satoshi Matsuoka
- Sun Microsystems K.K.
  - Shuuichi Ihara
  - Junichi Koike
- SOUM corporation
  - Takuya Ishibashi