Approach to provide supercomputer storage I/O information toward users

Tsuyoshi　NAKAGAWA
JAMSTEC (Japan Agency for Marine-Earth Science and Technology)

JAMSTEC
http://www.jamstec.go.jp/

28th WSSP Stuttgart Oct. 9-10 2018

Motivation

- Parallel supercomputer system become more and more complicated
- The needs from users are diverse, too

  (not only simulation but data analytics)

- Difficulty to promote the high level operation for all (node, network, storage)
- For users side, the information for advanced use is not always enough
- Necessary collaborate to other supercomputer center;
  - Exchange HW, SW and performance information
  - Statistic and technical support for advanced use
  - Operation technique

Our First step:

Open storage information to users

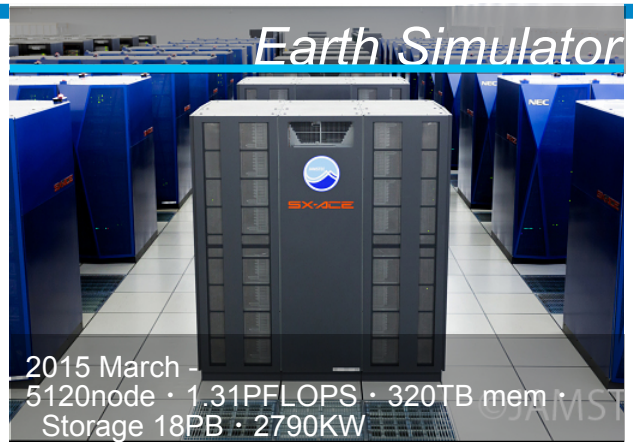
- I/O Benchmark
- I/O monitoring
- I/O profiler environment
- I/O statistics DB

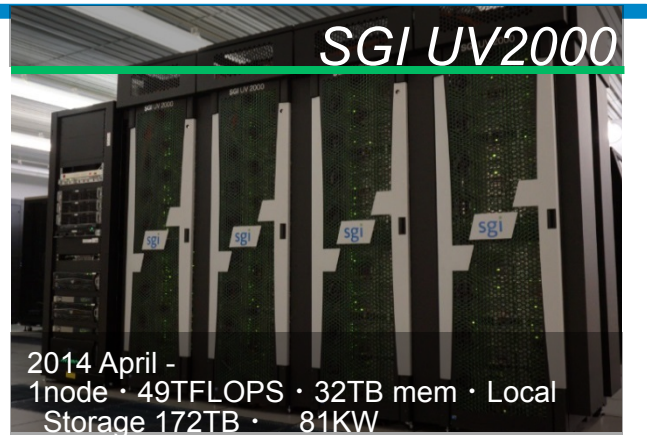
Agenda

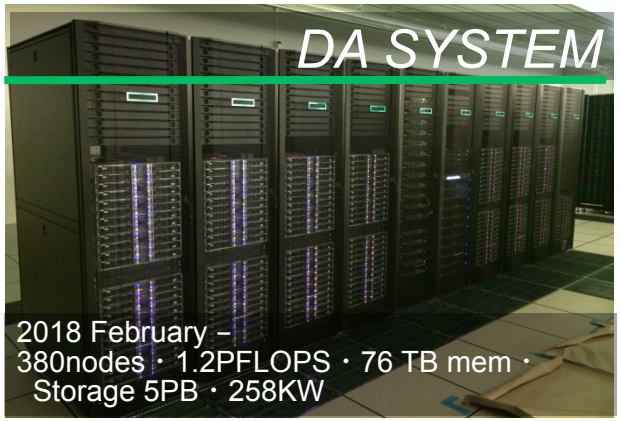
1. Introduction of JAMSTEC Storage System
2. Earth Simulator (ES)
   - New additional Storage's performance (ScaTeFS)
   - MPI-IO & POSIX performance toward users
3. New Linux cluster; "Data analyzer (DA)"
   - DA Storage's performance (Lustre)
   - DDN Lustre monitoring system
   - Further Plan
4. Application I/O survey
5. Summary

1: JAMSTEC Super Computer and Storage

Supercomputers & Storages



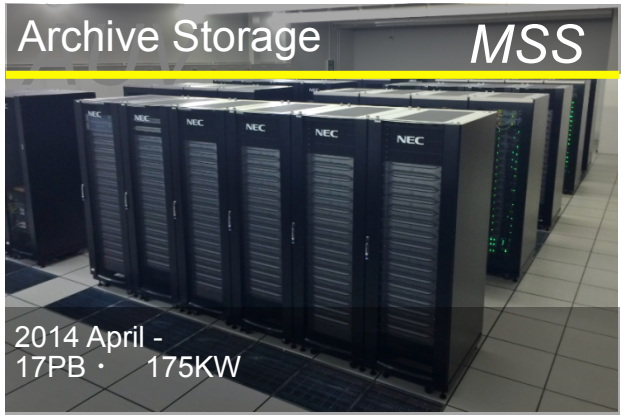


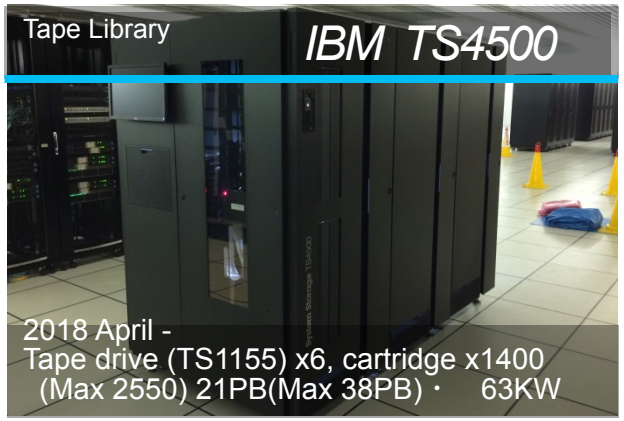


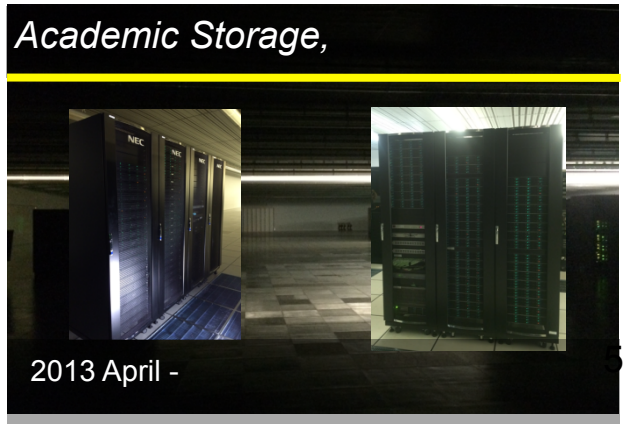
HPE APOLLO

NEC SC-ACE

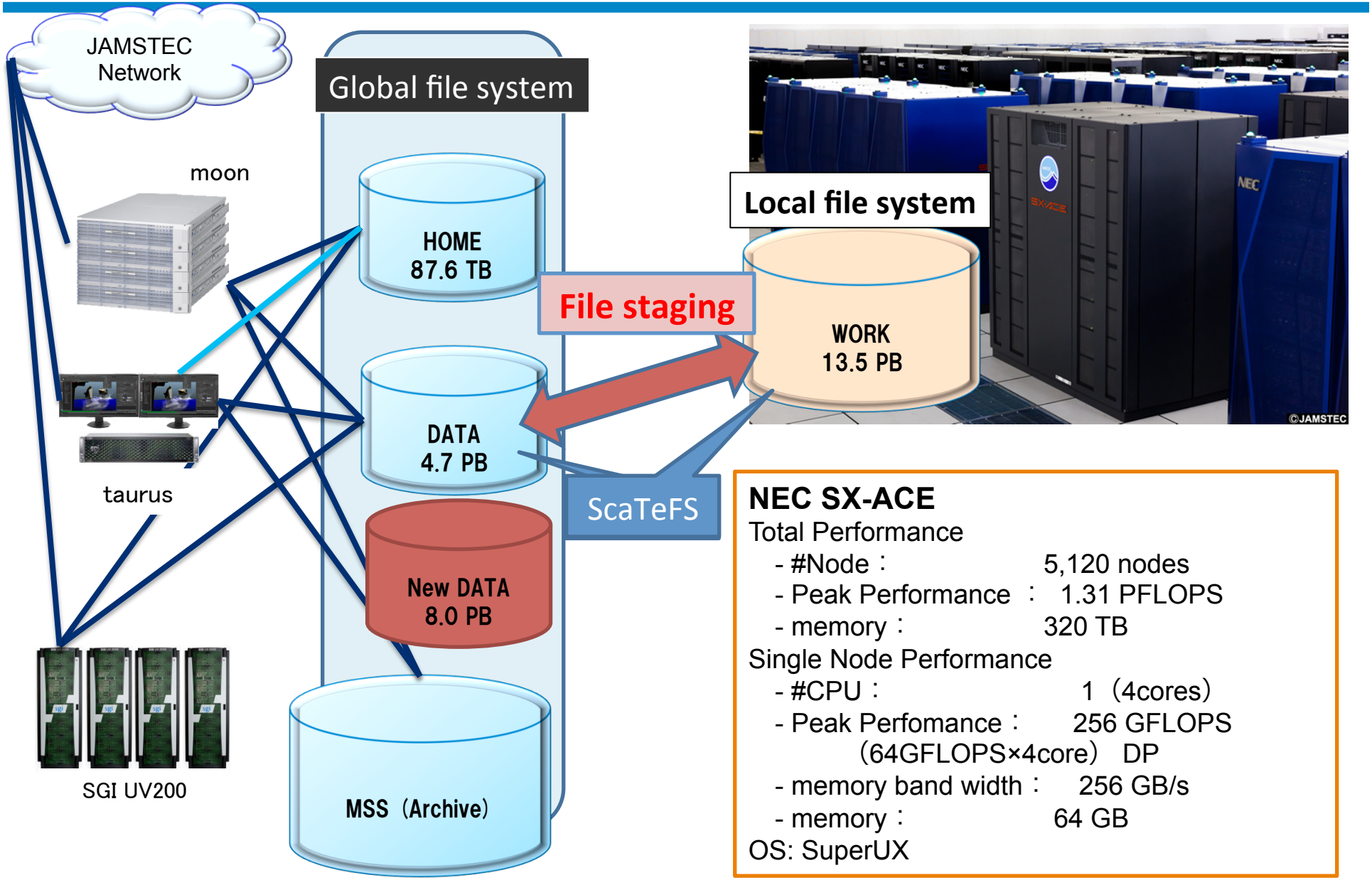
DATA







2: Earth Simulator (ES) system



NEC SX-ACE

Total Performance
- #Node： 5,120 nodes
- Peak Performance ： 1.31 PFLOPS
- memory： 320 TB

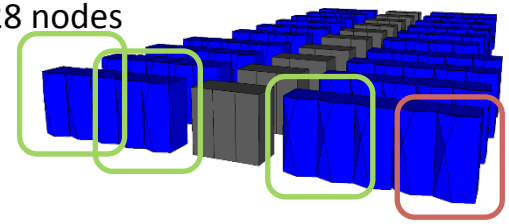
Single Node Performance
- #CPU： 1（4cores）
- Peak Perfomance： 256 GFLOPS
  （64GFLOPS×4core） DP
- memory band width： 256 GB/s
- memory： 64 GB

OS: SuperUX

ScaTeFS I/O servers of ES storage system

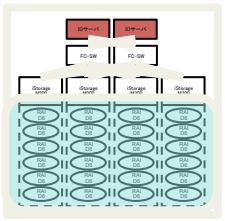
WORK region

Local file system
shared for 128 nodes



8GB/s

2 IO server pair (2 IOSV)= 1 partition; 345 TB

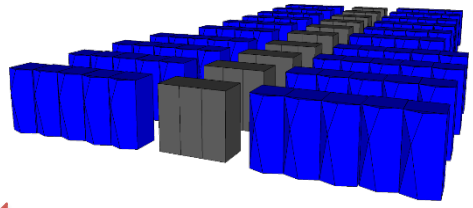


#iot = 24

×40 sets
= total 320 GB/s

DATA region

Global file system

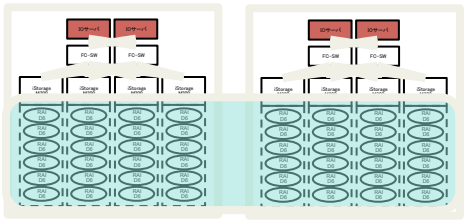


Login Server x4

16GB/s

2GB/s

4 IOSV = 1 partition; 690 TB

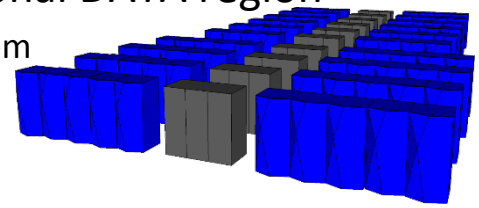


#iot = 48

×7 sets
= total 112 GB/s

New additional DATA region

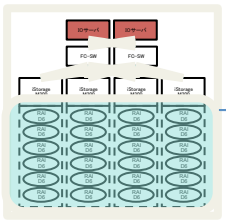
Global file system



Linux Server

16GB/s

2GB/s



#iot = 28

Total 8 PB ScaTeFS Storage
= 16IOSV = 8 IO unit

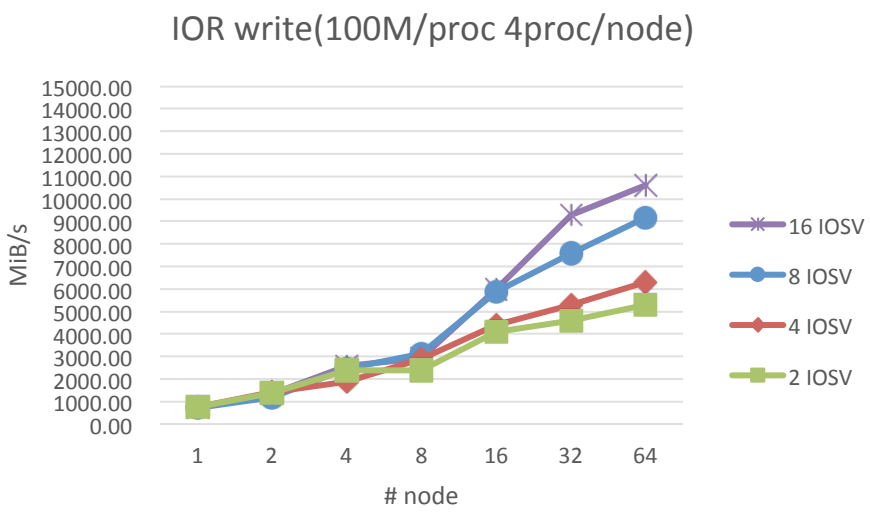
2 IOSV? = 1 partition; 1 PB x8

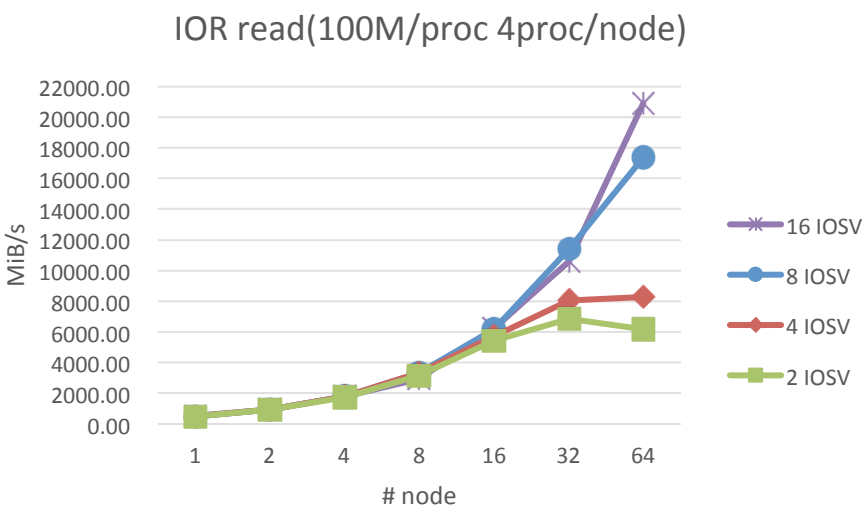
4 IOSV? = 1 partition; 2 PB x4

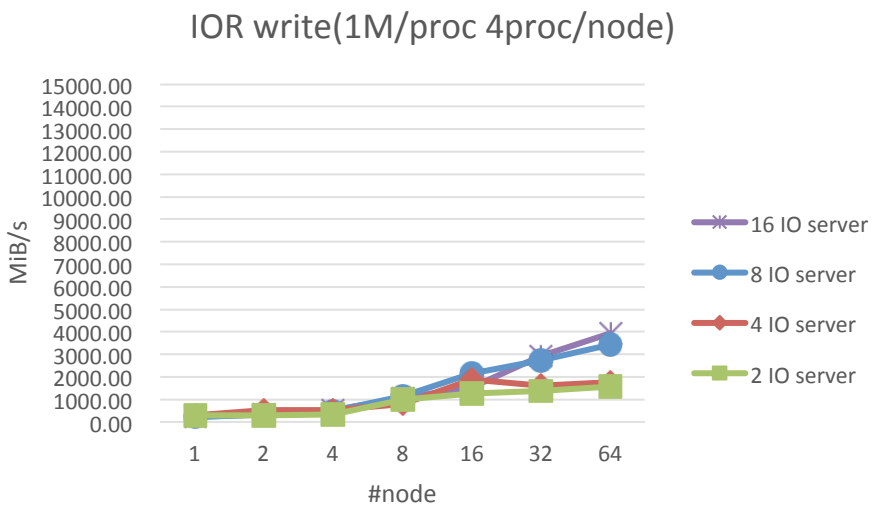
8 IOSV? = 1 partition; 4 PB x1

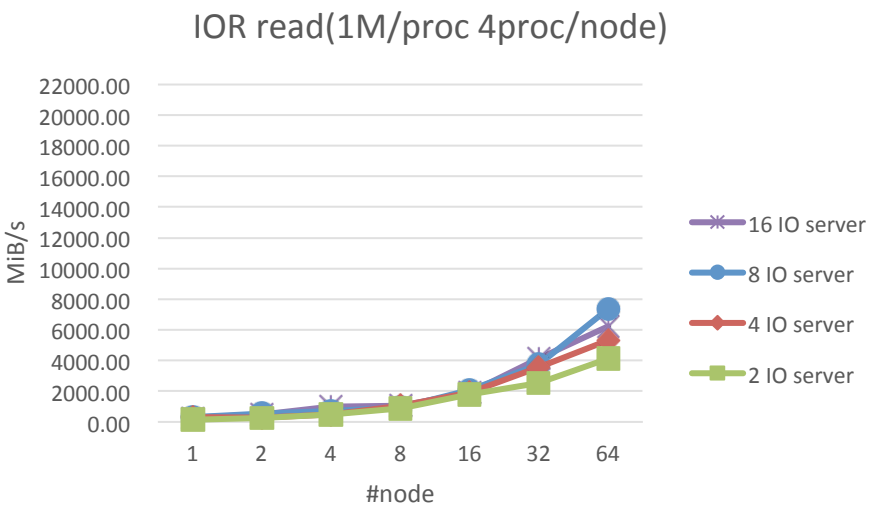
Benchmark them before the actual operation

IOR Benchmark New DATA (SX-ACE;Qfabric(10GbE))





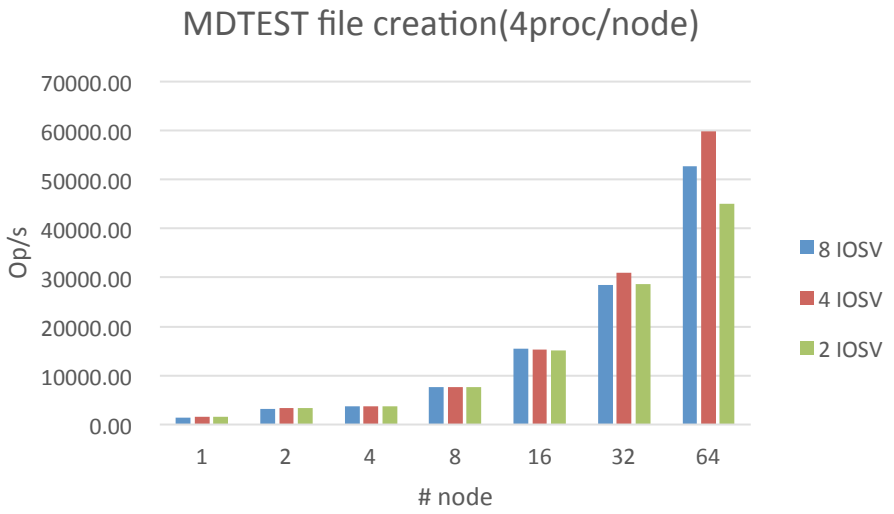


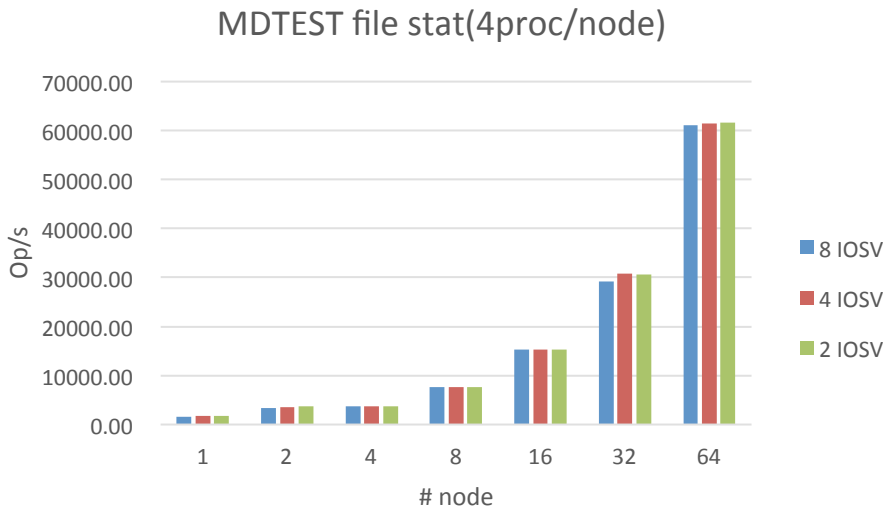


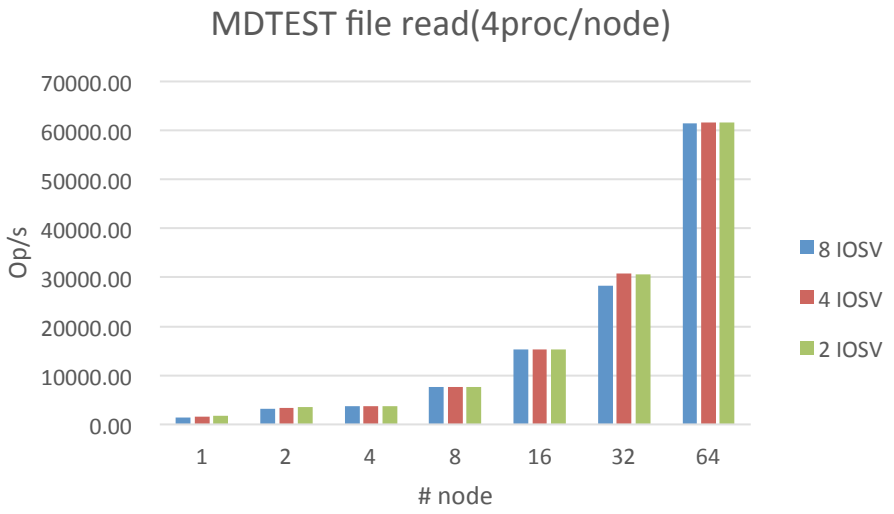
I/O BW performance is well scaled over 100MB

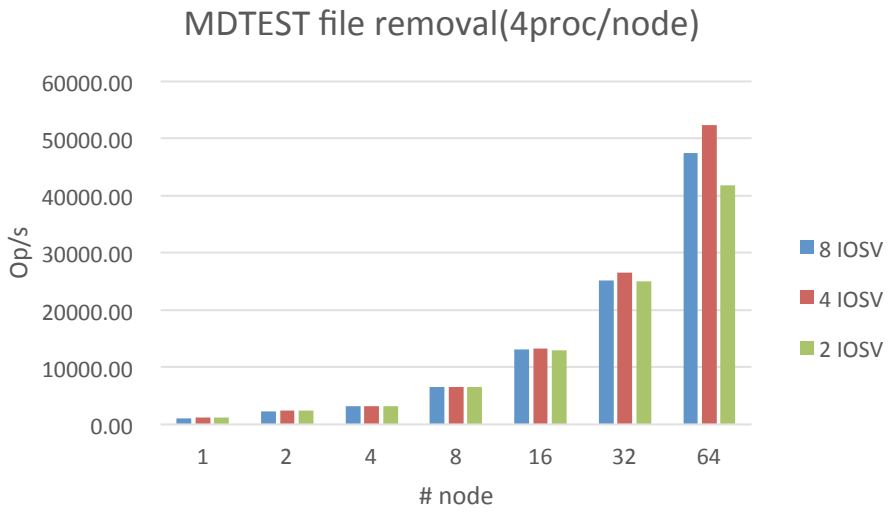
mpirun −nn ${N_NUM} −nnp ${P_NUM} ./IOR −F −i 3 −t 4M −b ${F_SIZE}

mdtest Benchmark (SX-ACE;Qfabric(10GbE))









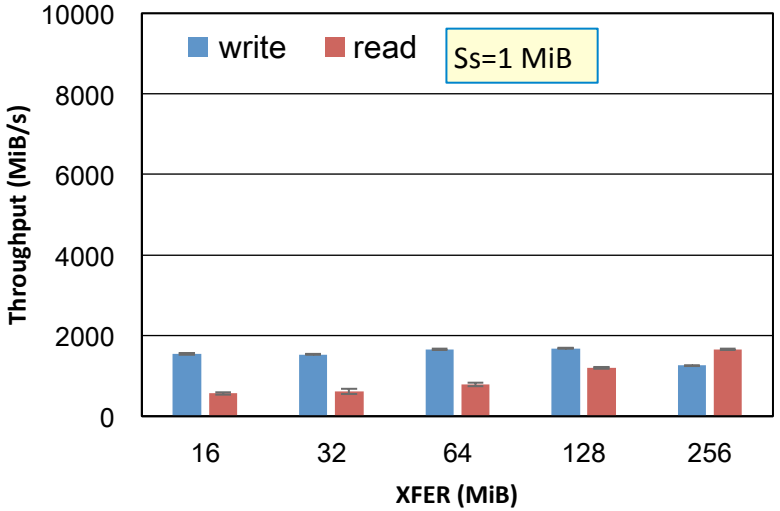
Metadata operation is independent on # IOSV

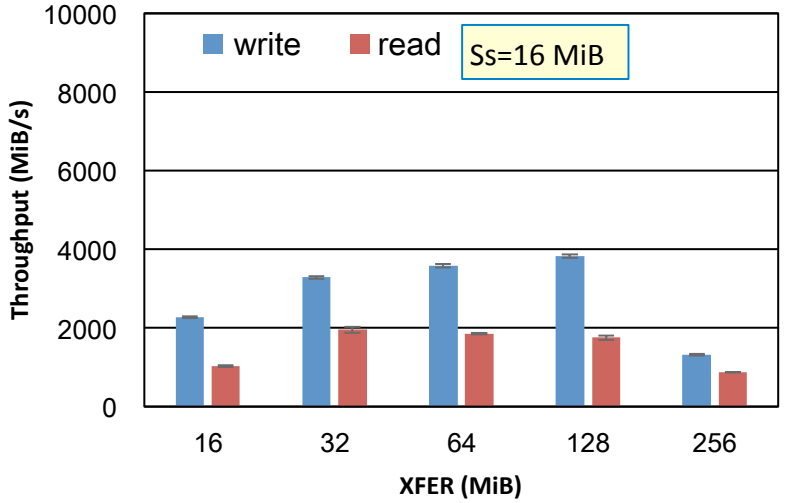
mpirun -nn ${N_NUM} -nnp ${P_NUM} ./mdtest -n 5000 -i 3 -V1 -p 10 -u -d ./

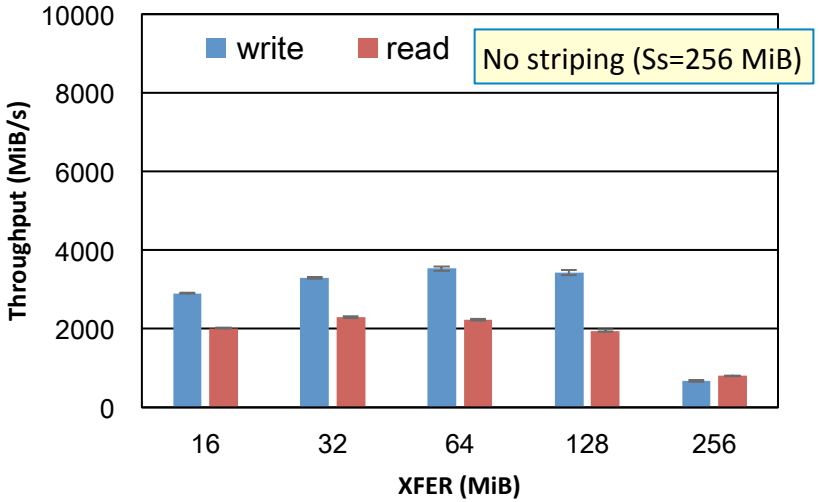
IOR benchmark・MPI-IO：BLK=256MiB（defaults setting）

- np=384@96nodes, Cs=-1 #384（Average, Max, Min in 5 times reputation.）
  - ✓ ROMIO; No Two Phase I/O using collect buffer
  - ✓ Each process I/O by XFER independently ; No data exchange communication occurs

```
Write:  ./IOR -i 5 -a MPIIO -c -k -m  –U ${HINTS_FILE}  –H -w -t ${XFER}m -b 256m -s 1 -o ../IOR-data/IOR-test -d 0.1
Read:   ./IOR -i 5 -a MPIIO -c -k -m –U ${HINTS_FILE}  –H -r -t ${XFER}m -b 256m -s 1 -o ../IOR-data/IOR-test -d 0.1
```





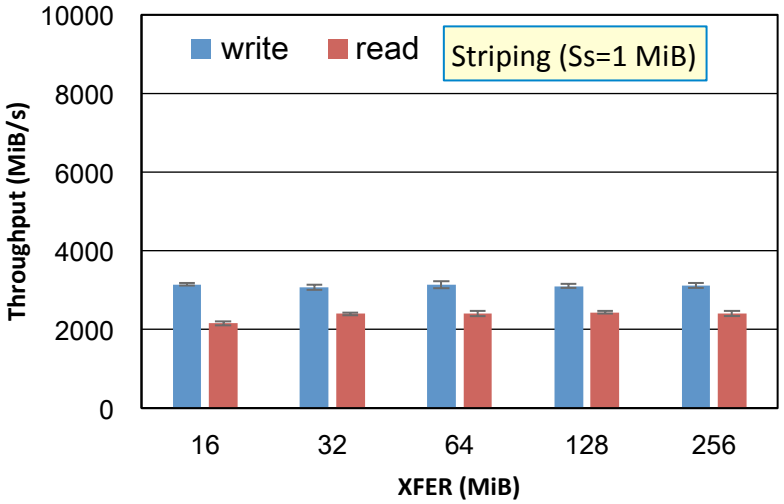


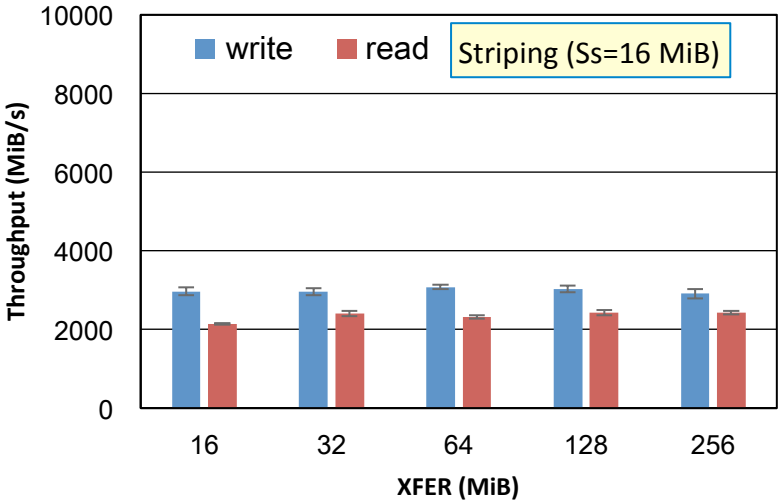
- Along with an increase in stripe size, performance improvement
- Performance improves as XFER increases, but performance drops at 256 MiB

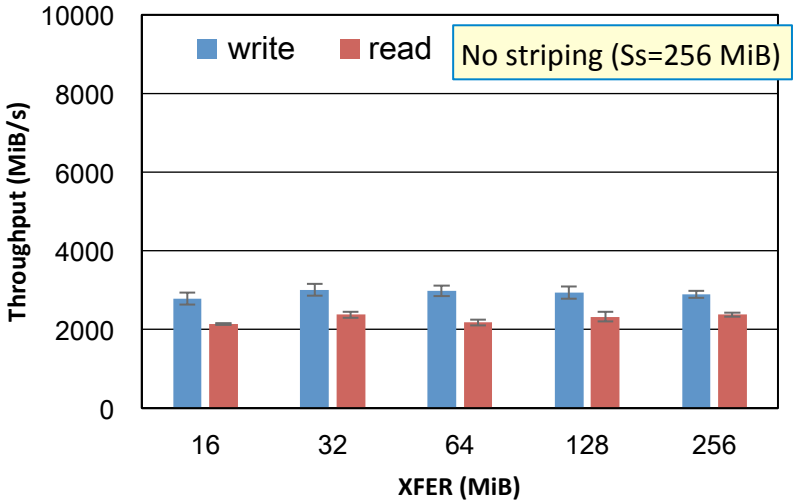
IOR benchmark•POSIX-I/O: BLK=256MiB

- np=384@96nodes （Average, Max, Min in 5 times reputation.）

Write: ./IOR -i 5 -a POSIX -k -m –H -w -t ${XFER}m -b 256m -s 1 -o ../IOR-data/IOR-test -d 0.1
Read: ./IOR -i 5 -a POSIX -k –m –H -r -t ${XFER}m -b 256m -s 1 -o ../IOR-data/IOR-test -d 0.1

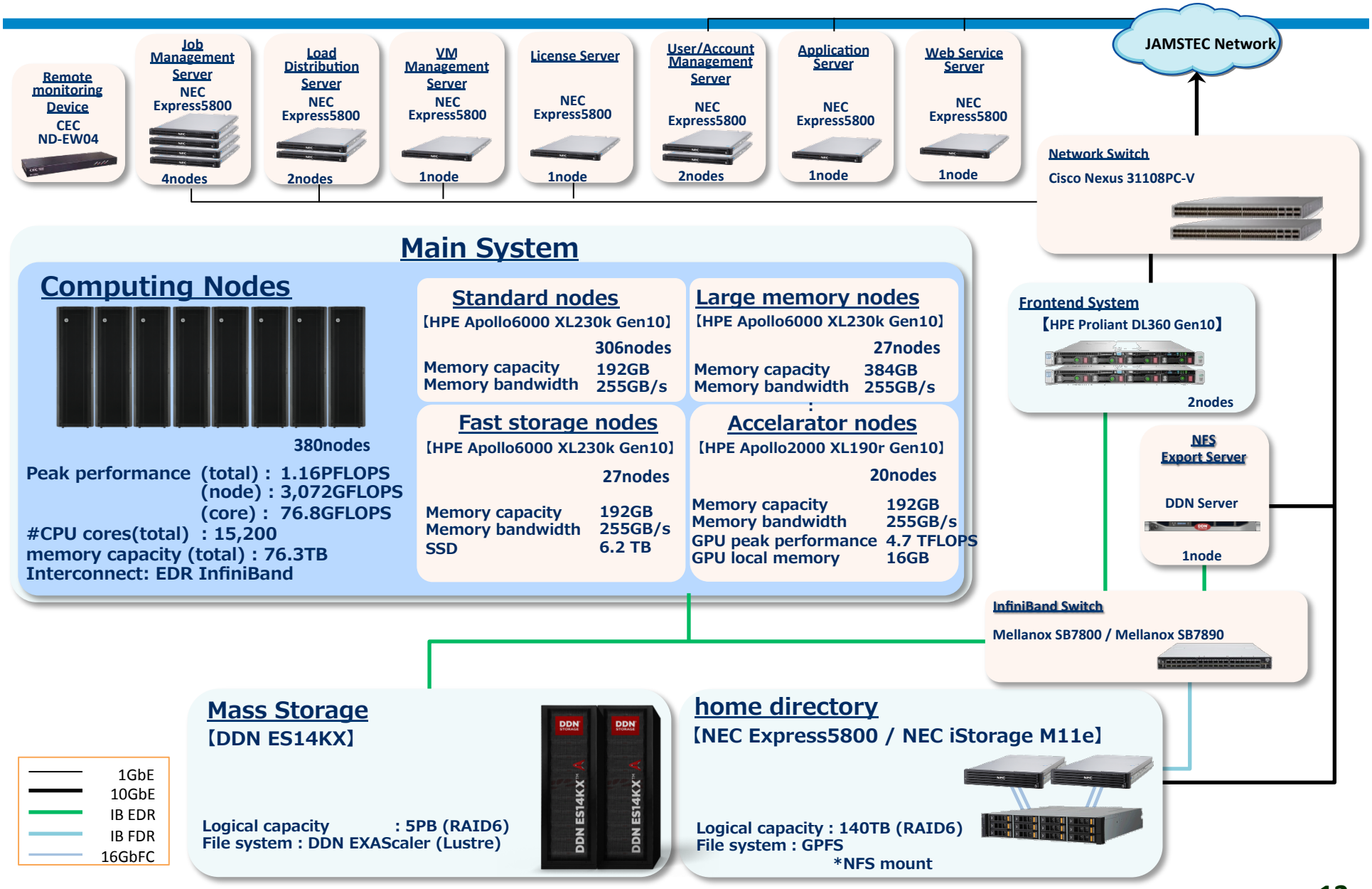






- Changes in stripe size do not change ScaTeFS characteristics
- The performance of POSIX-I/O is generally higher than MPI-I/O Because MPI-IO communicates by I/O pattern analysis at the beginning and make the communication cost increase

3 : Data Analyzing system – system configuration



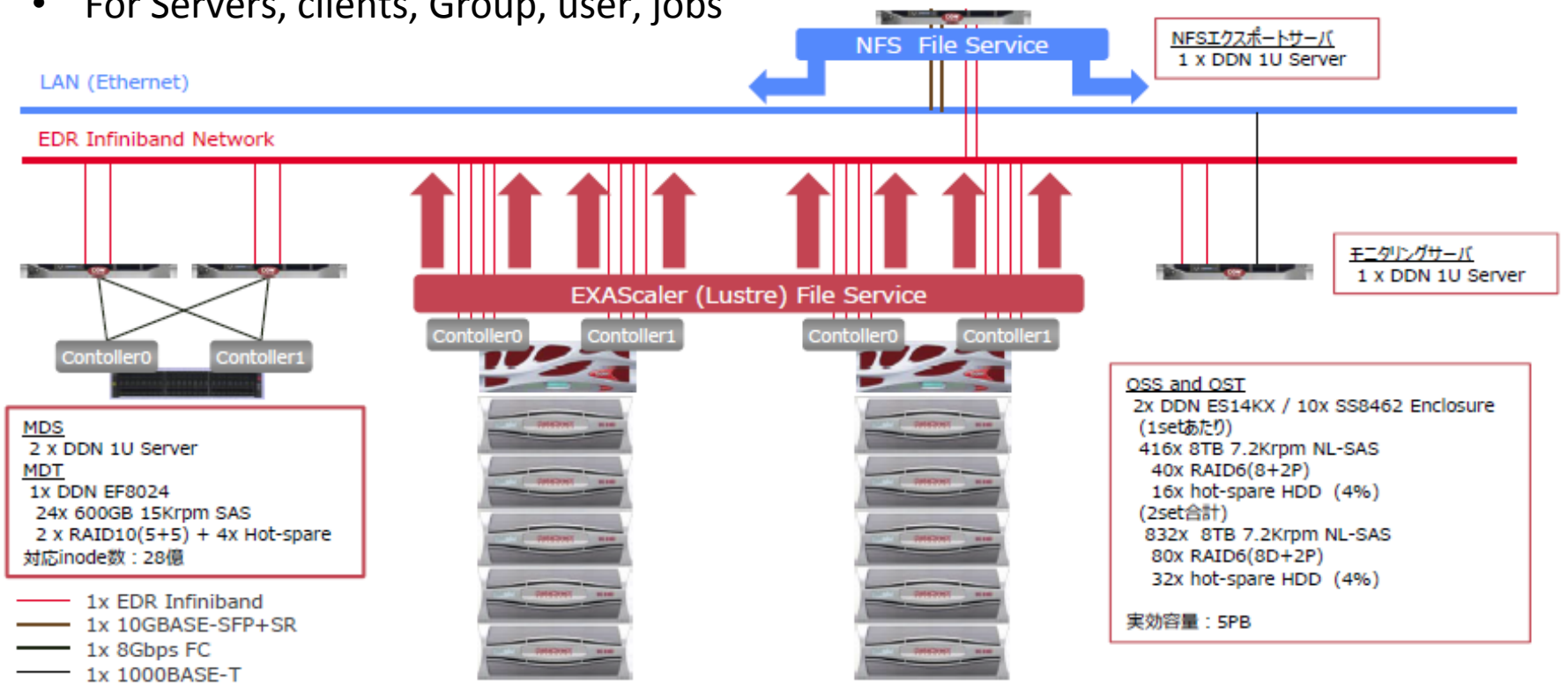
DA Storage system

Work storage

- DDN ES14000X
- File system : Lustre (EXAScaler) ; 5 PB

DDN Lustre Montor

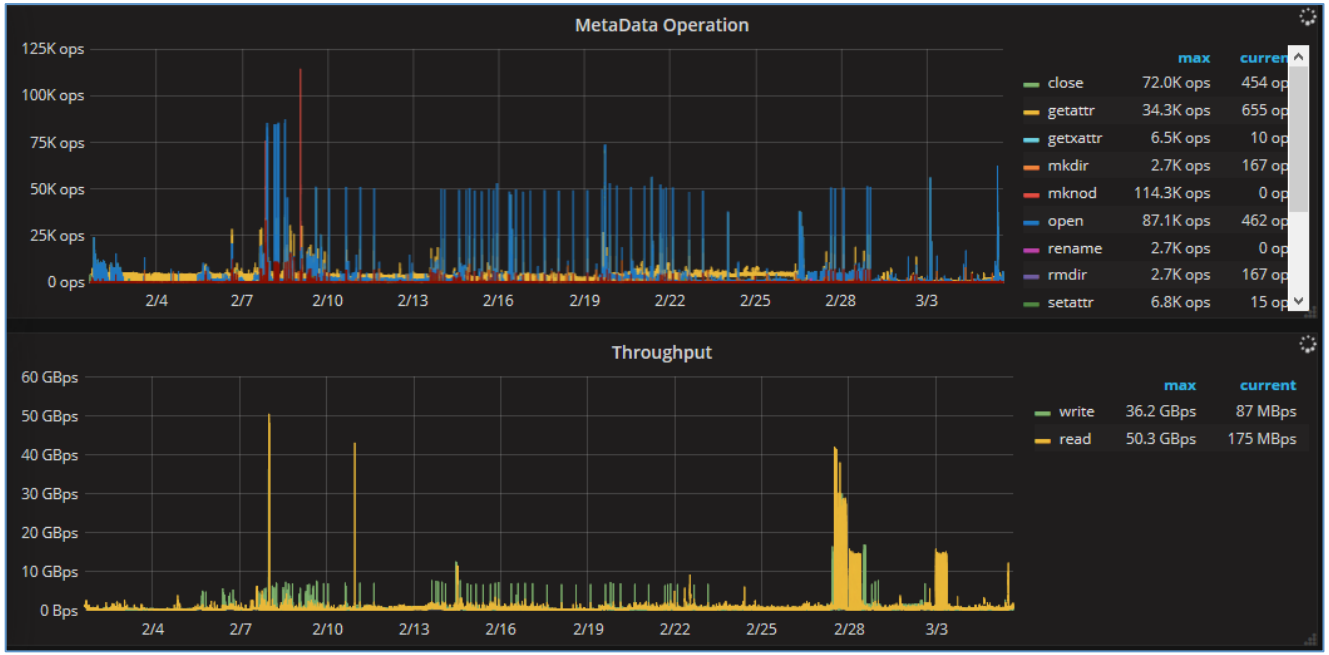
- Consecutive I/O monitoring and logging
- Throughput, IO size, File System Usage, Load Average (MDS, OSS)
- Metadata operation
- For Servers, clients, Group, user, jobs

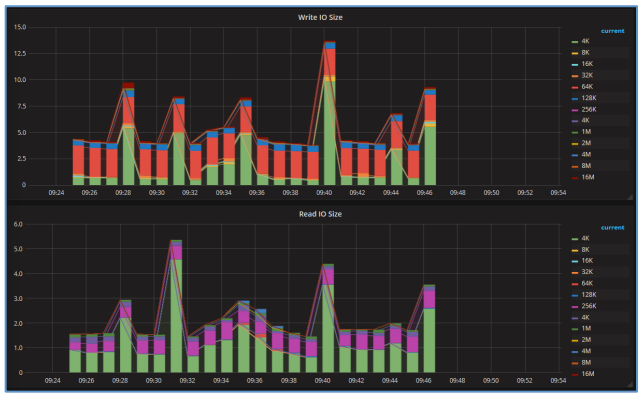


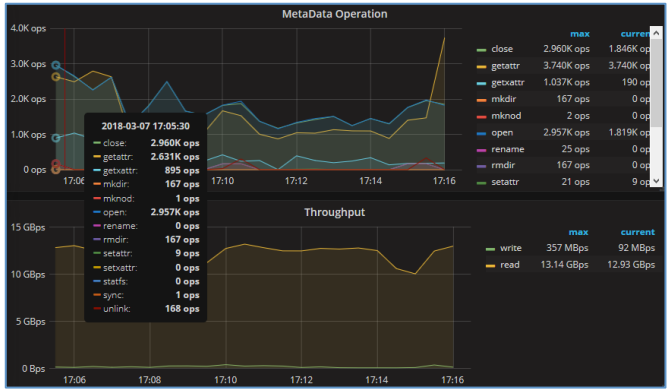
I/O monitoring on DA by DDN Lustre Monitor

- Analyze I/O characteristic of each applications toward optimization

- Statistic data of I/O performance and Trouble shooting for system management.





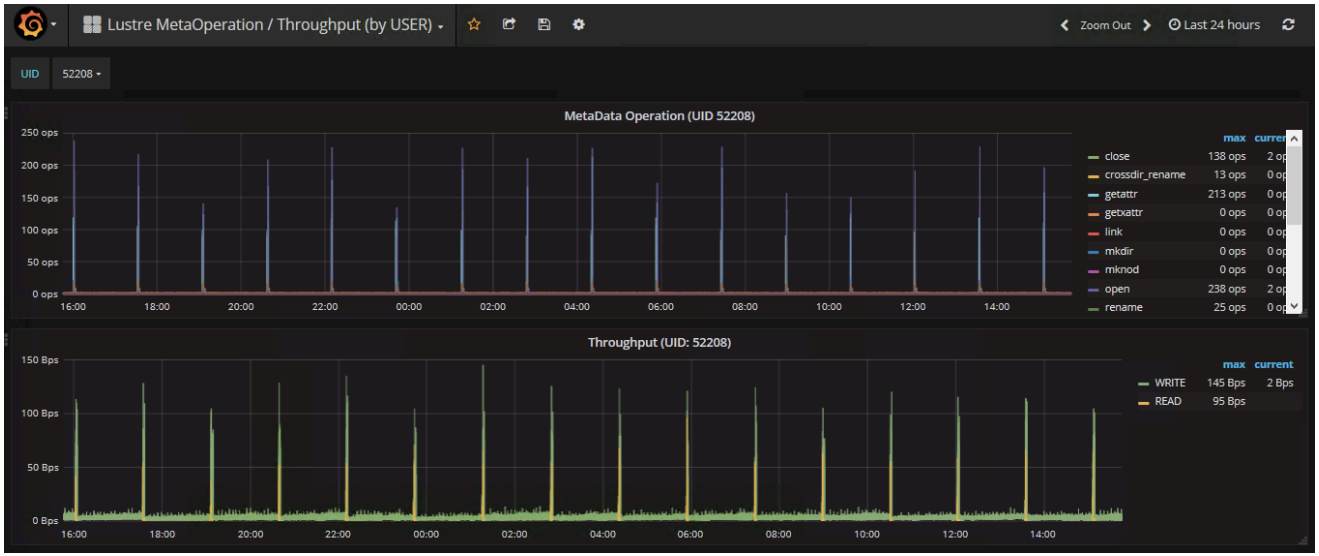


Time series DB: InfluxDB and Visualization by Grafana

I/O pattern

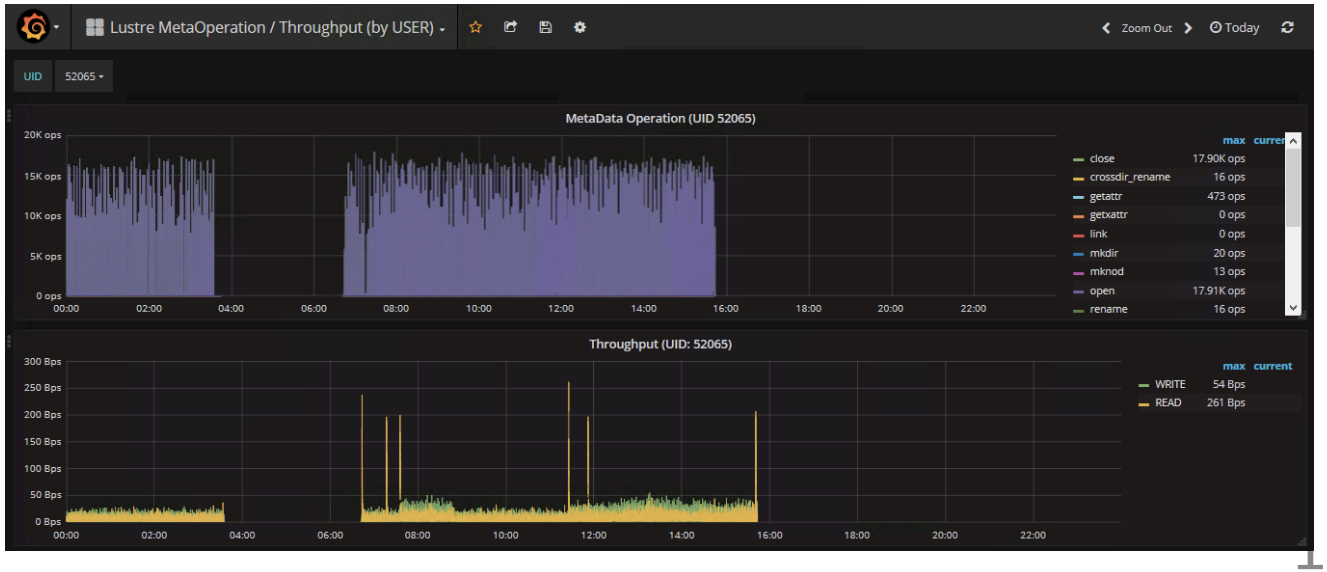
Simulation

Periodic
Sequential



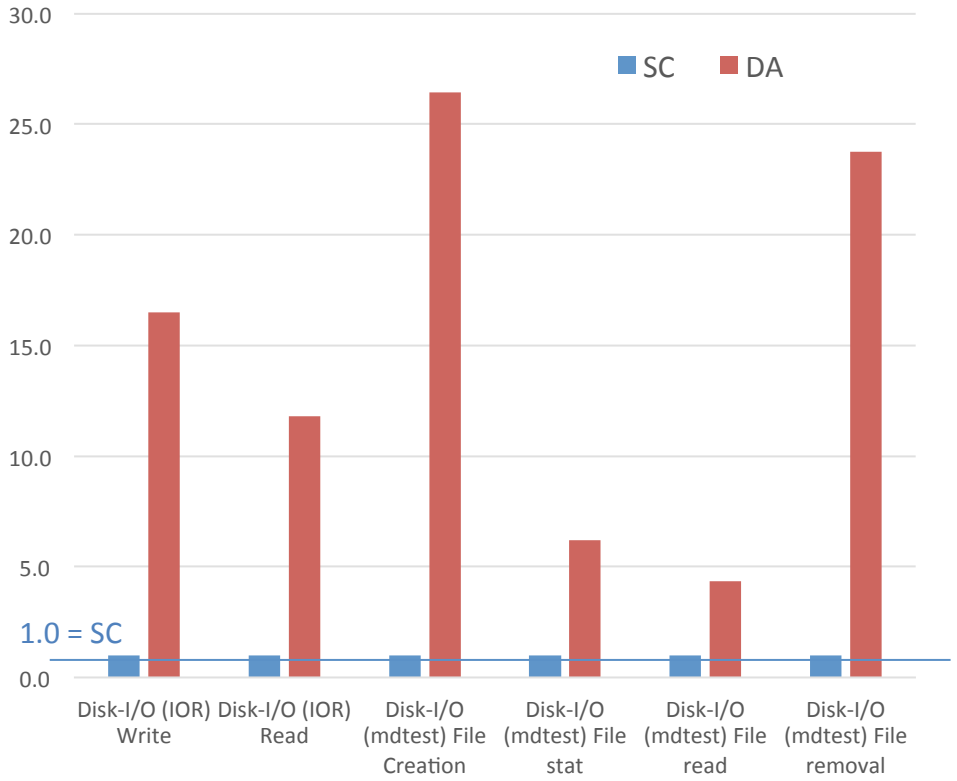
Data Analytics

No Periodic
Random



Performance of DA storage from 32 nodes

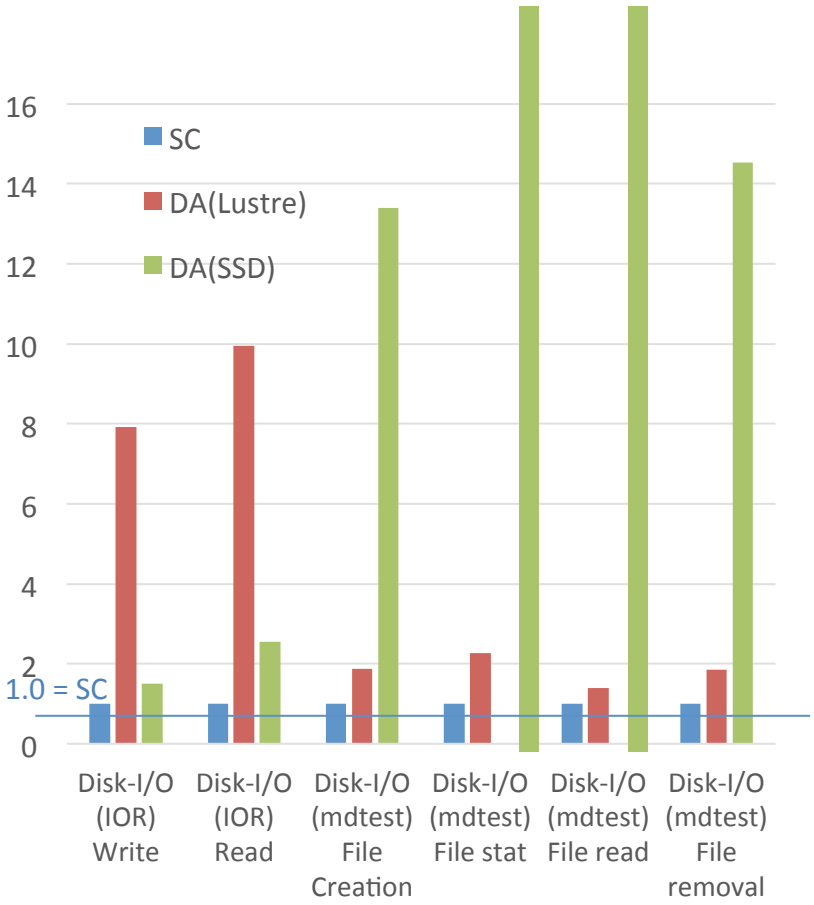
		SC	DA
		SGI IS16000	DDN ES14KX
		Lustre 1.6 base	Lustre 2.7 base
		512MPI/ 32nodes	128MPI/ 32nodes
Throughput	Write (GB/s)	4.0	67.3
	Read (GB/s)	5.6	70.7
Metadata Operation	File Creation	2195.0	68550.0
	File Stat	24500.0	163520.3
	File read	40172.0	179876.9
	File removal	4212.0	126751.4



x 9.0 greater I/O performance than the previous SC system

Performance of DA storage from 1 node

		SC	DA	
		SGI IS16000	DDN ES14KX	SSD
		Lustre 1.8 base	Lustre 2.7 base	XFS RAID0
	#process	16	40	40
Throughput (1nodes)	Write	1545.5	12233.2	2323.3
	Read	929.0	9252.0	2101.9
Metadata Operation (1nodes)	File Creation	3265.6	6134.8	43723.1
	File Stat	4433.2	10063.0	26727228.7
	File read	6067.7	8464.3	17616464.3
	File removal	2498.5	4603.0	36311.1



x 9.0 greater performance for I/O BW but x 2.0 for metadata than the previous SC system

4: Application I/O survey

Motivation:

To estimate the required I/O performance of JAMSTEC codes for the future. To prepare the benchmark reflecting the real I/O workload for the next supercomputer procurements as some kernels.

I/O profiler:

Darshan 3.1.5 (Light weight I/O profiler tool developed at Argonne National Laboratory)

%export LD_PRELOAD=/xxx/darshan/lib/

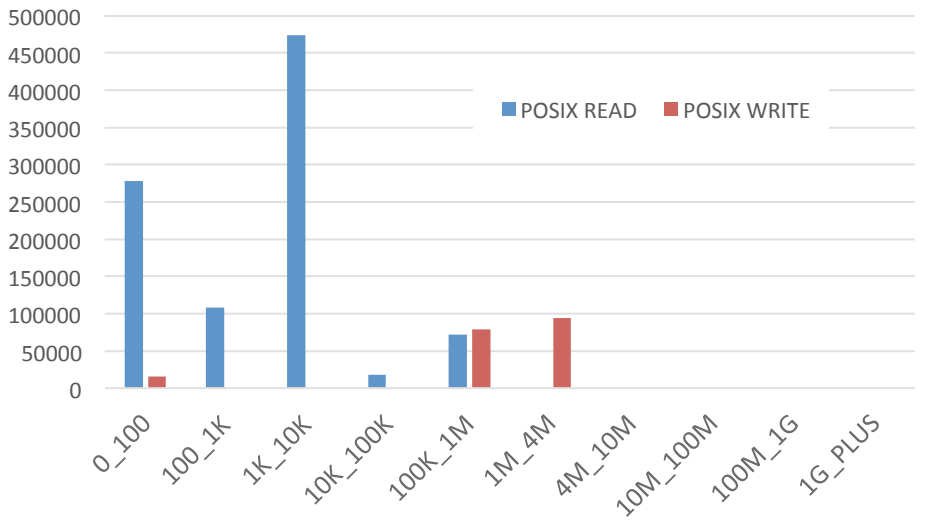
Darshan-riken 2.3.0 (extended version for recoding I/O time history and non MPI application, https://www-sys-aics.riken.jp/releasedsoftware/ksoftware/darshan/)

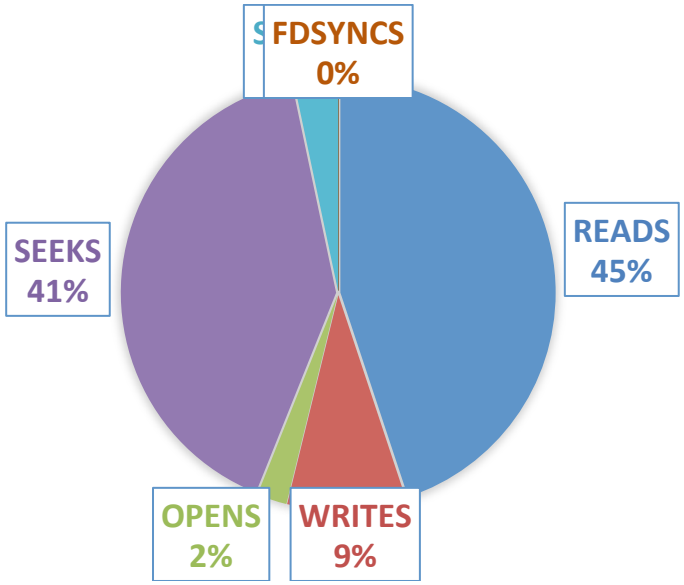
Execution hosts:

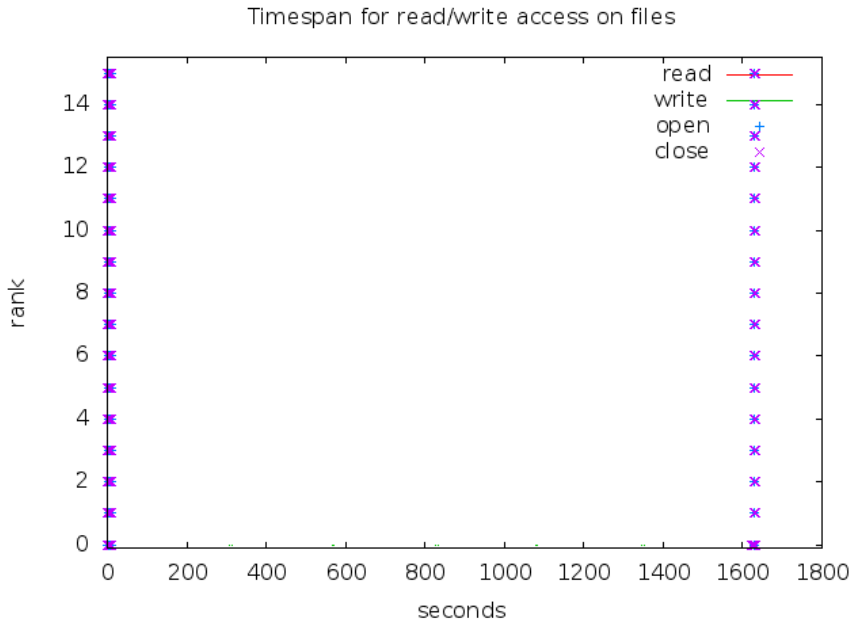
DA (because the installation for ES has some difficults) ; SC

CASE1 : *~Regional atmosphere model~*

1Hour Simulation, 1km grid, 1792x1792x32
1024MPI
Output interval(grad, restart): 1 Hour
Run time: 1632s
Total#files: 14,705
read_only: 10,385
write_only: 4,320
Total READ: 60.64 GB
Total WRITE: 261.87 GB
Total META: 2,120,040 op
Total READ_TIME: 1053.70 s
Total WRITE_TIME: 234.72 s
Total META_TIME: 1124.43 s
I/O time: 9.05 s (1% of Run time)
Estimated I/O rate: 38.022 GB/s



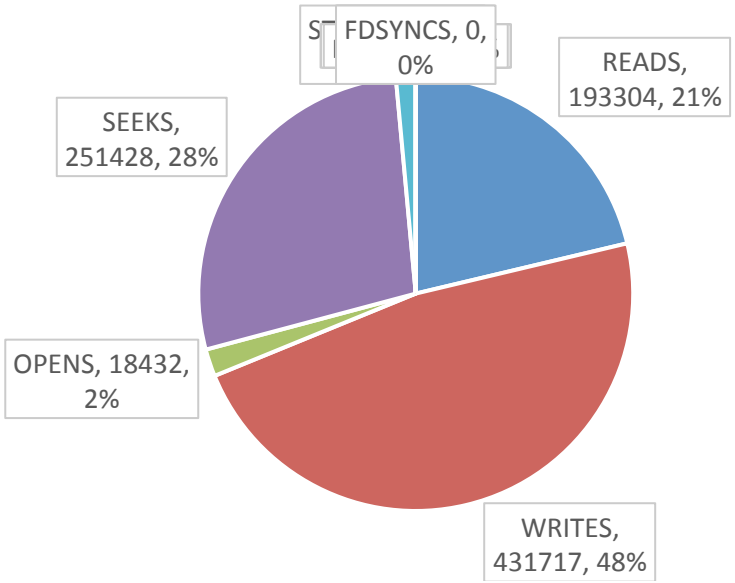


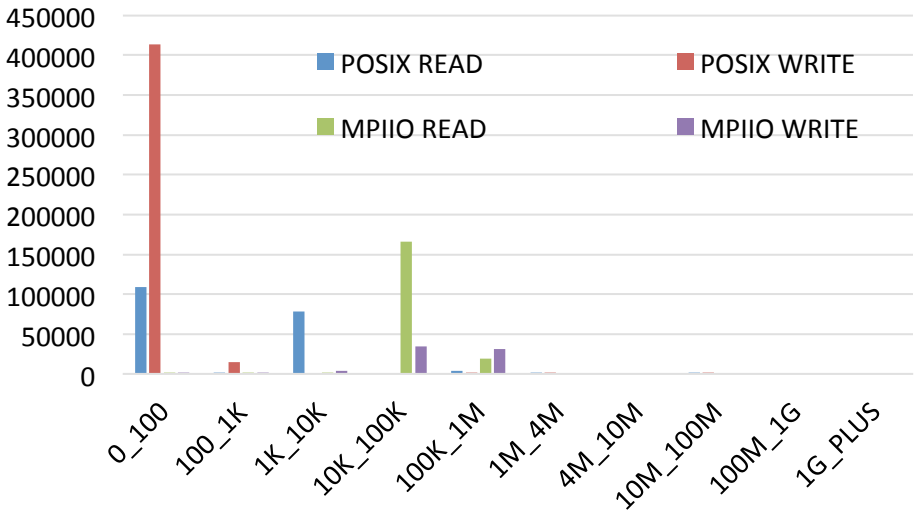


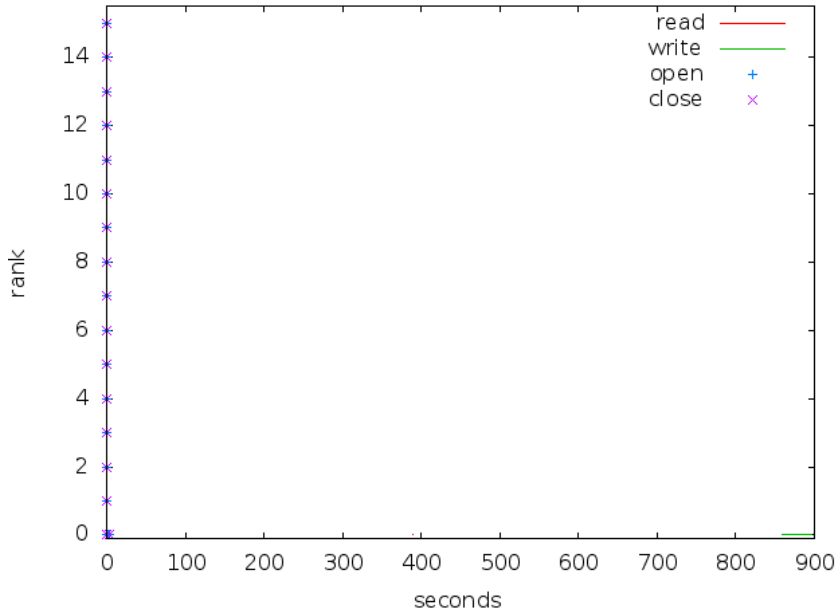
CASE2 : *~Global Ocean model~*

1Month Simulation, 25km grid
1024MPI

Output interval:	month
Run time: 898s	
Total#files: 1,041	
read_only: 12	
write_only: 1,029	
Total READ:	21,3 GB
Total WRITE:	28,7 GB
Total META: 908,193 op	
Total READ_TIME:	123.11 s
Total WRITE_TIME:	667.73 s
Total META_TIME:	183.99 s
I/O time: 42.6s (5%)	
Estimated I/O rate:	1.11 GB/s





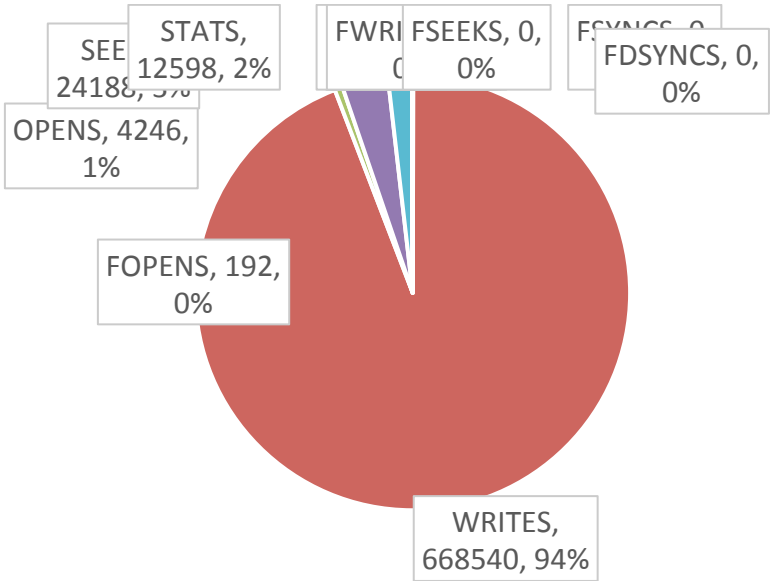


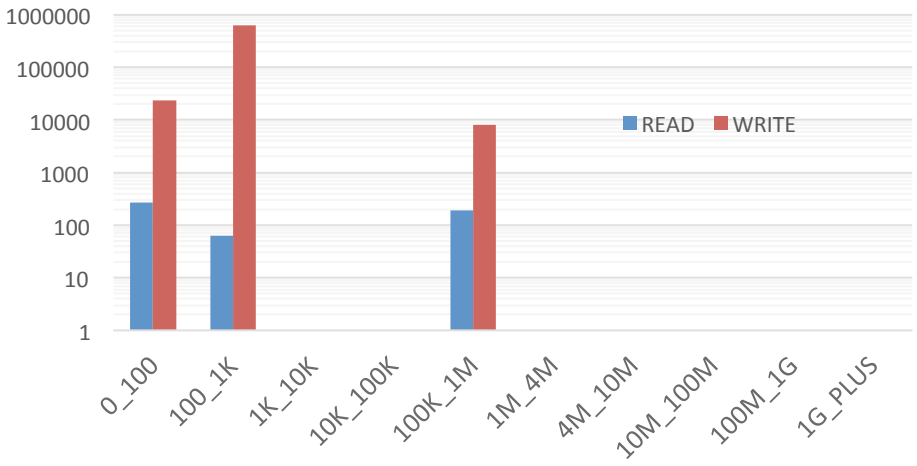
CASE3: *~Tsunami Simulation~*

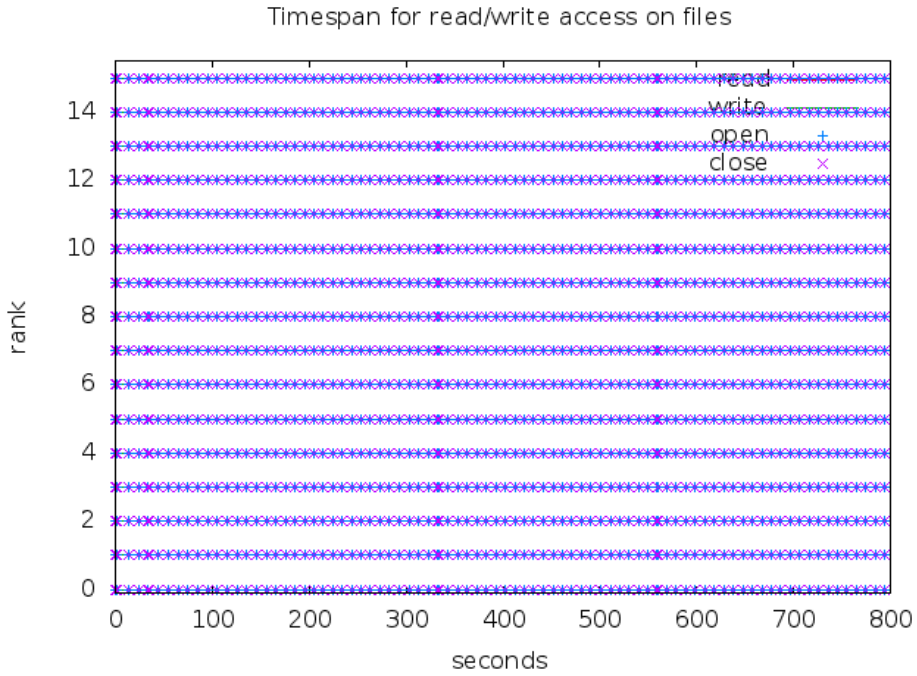
https://github.com/jagurs-admin/jagurs

80 hours simulation ; 30m grid
16MPI16SMP

Output interval:	60s
Run time:	795s
Total#files:	4,105
read_only:	99
write_only:	4,006
Total READ:	125.5 MB
Total WRITE:	5.2 GB
Total META:	710,292 op
Total READ_TIME:	4.53 s
Total WRITE_TIME:	23.51 s
Total META_TIME:	5.31 s
I/O time:	8.45s (1%)
Estimated I/O rate:	601.87 MB/s

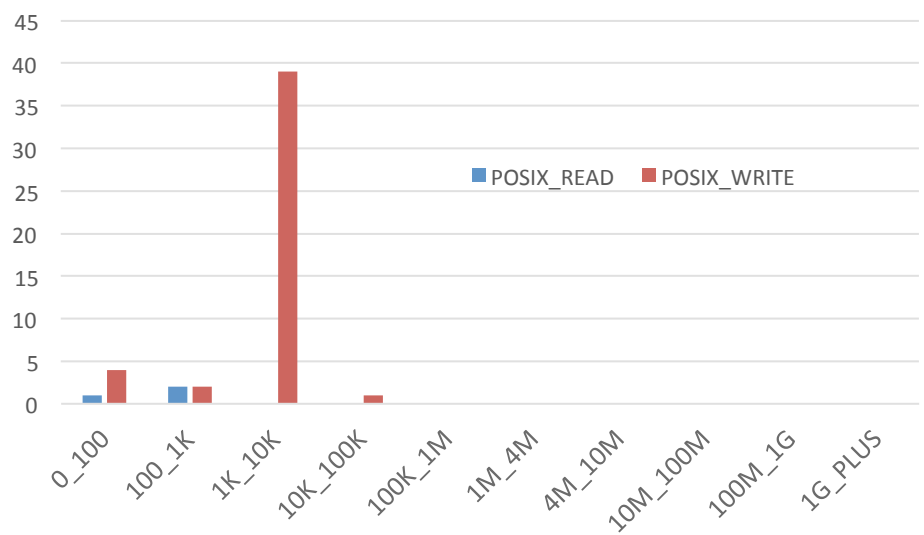


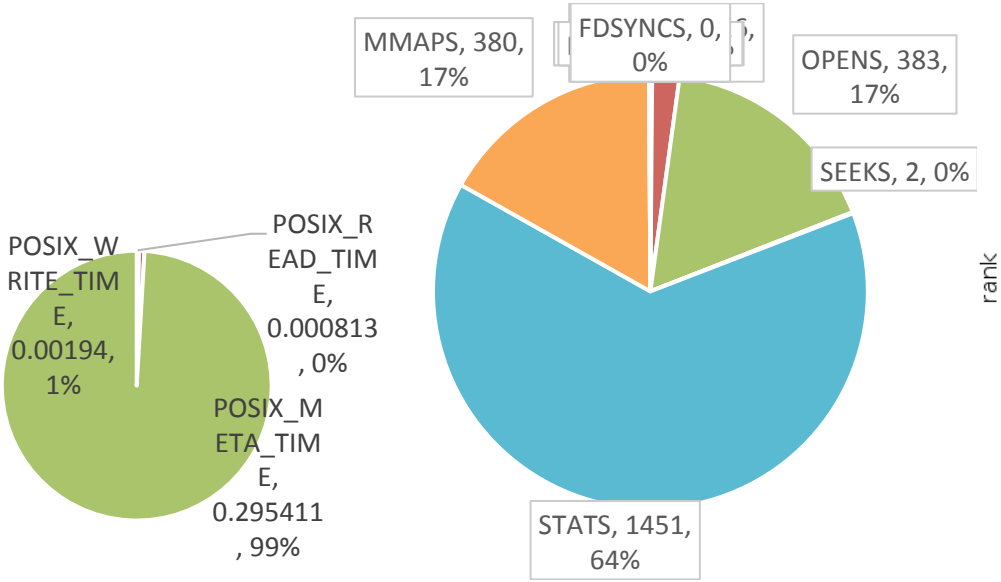


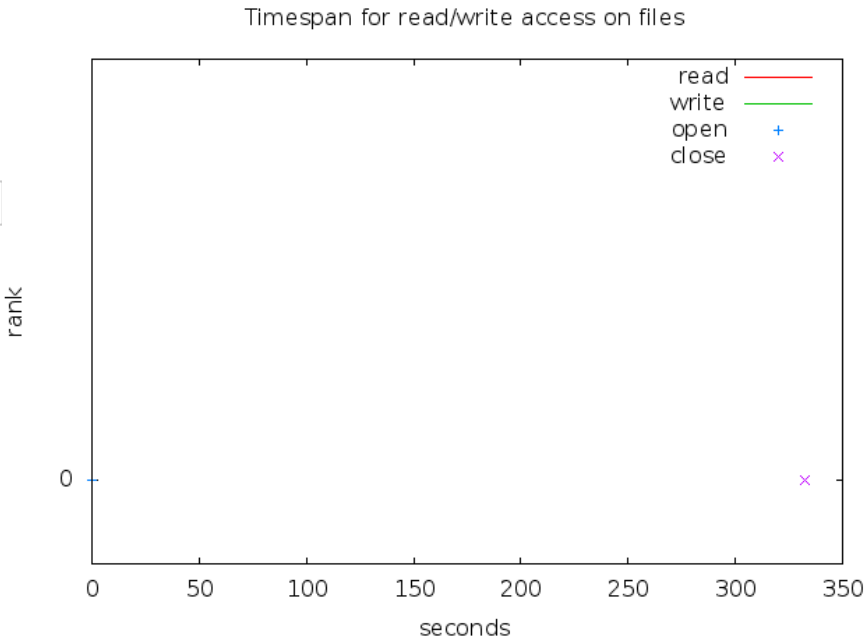


CASE4 : *~similarity between biological sequences~*

Ave. input query size: 131 ~ 450 depends on sample
DB size: 7,003,66 seq.; 2.3 GB
No MPI 4Threads
Output interval: per Input
Run time: 334 s
Total#files: 384
read_only: 383
write_only: 1
Total READ: 650 B
Total WRITTEN: 326.01 KB
Total META 2,267 op
Total READ_TIME: 0.000813 s
Total WRITE_TIME: 0.00194 s
Total META_TIME: 0.295 s
I/O time: 0.593 s (1%)
Estimated I/O rate: 1.04 MB/s







I/O profiling summary

	①	②	③	④
Research area	Atmosphere	Ocean	Solid Earth	Bio
Total I/O size	320 GB	50 GB	5 GB	300 KB
Running time	1632 s	898 s	795 s	334 s
% I/O time	1 %	5 %	1 %	0.1 %
I/O speed	38 GB/s	1.0 GB/s	0.6 GB/s	0.004 GB/s
# I/O File	~15,000	~1000	~4000	~400
# meta operation	2 Mops	0.9 Mops	0.7 Mops	0.002 Mops
% meta / IO time	46 %	21 %	14 %	99 %
# Case run simultaneously	~10?	~10?	~100?	~10000?

5: Summary

- I/O Benchmark
  - Benchmark the additional ScaTeFS storage with changing the number of I/O server pairs
  - Benchmark the work storage system of DA system
- I/O monitoring
  - Lustre monitor start logging I/O information
- I/O profiler environment
  - I/O profiling for some of applications with darshan
  - Continue to analyze for the proper benchmark setting on the next procurement
- I/O statistics DB
  - Further Plan : To merge them to Job summary information DB

Thank you for your attention.