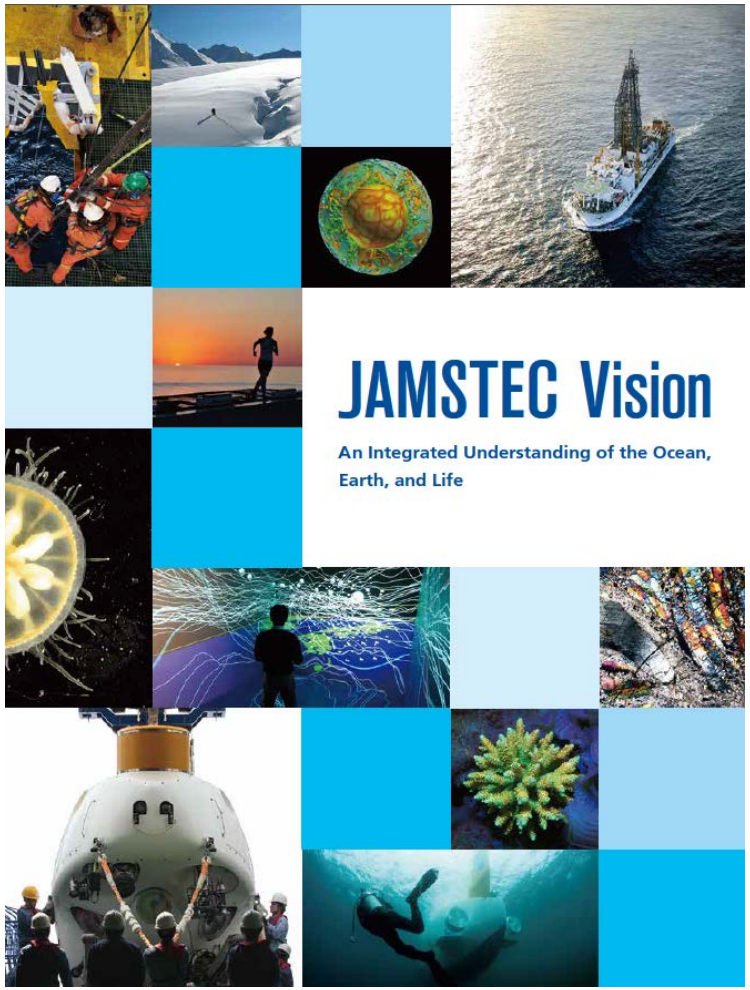
Improving data storage system structure and performance over various applications at JAMSTEC

CEIST, JAMSTEC
Tsuyoshi　NAKAGAWA
tnakagawa@jamstec.go.jp

1. about JAMSTEC
   - JAMSTEC Research Target
   - Facilities; Observation
2. System: Supercomputer and Storage
   - about Earth Simulator
   - File staging system
   - File system: ScaTeFS (NEC)
3. Evaluate I/O Performance
   - mdtest & IOR
   - Request Statistic Analysis
4. Summary

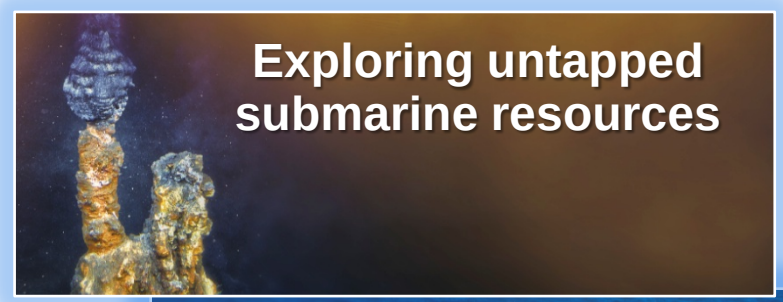
about our JAMSTEC

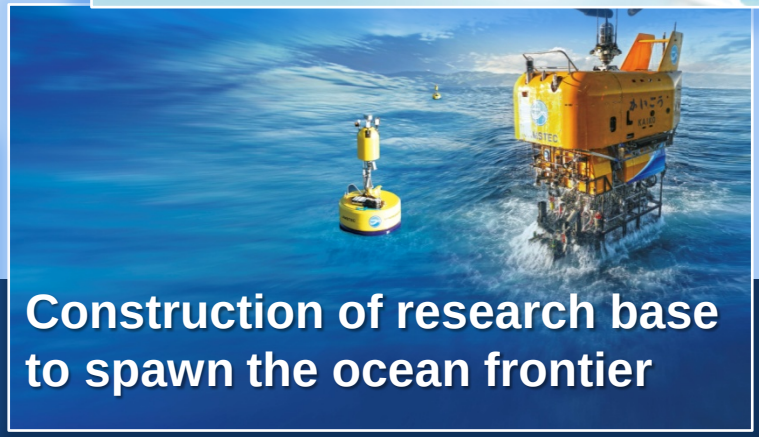
Japan Agency for Marine-Earth Science and Technology (JAMSTEC) has the main objective to contribute to the advancement of academic research in addition to the improvement of marine science and technology by proceeding the fundamental research and development on marine, and the cooperative activities on the academic research related to the Ocean for the benefit of the peace and human welfare.

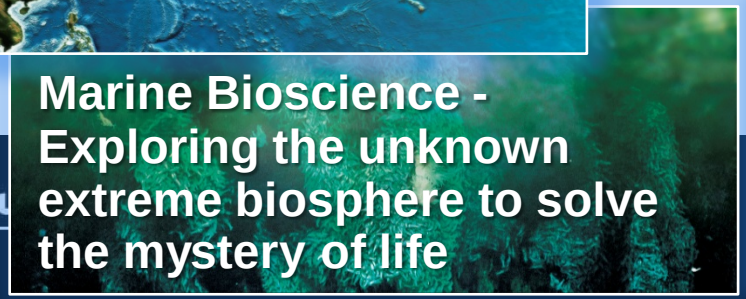


The main seven research and development issues during the third mid-term plan

During the third mid-term plan, we set and address the seven research and development issues with all our strength due to promote strategic and focused research and development based on the national and social needs.







JAMSTEC Site Location

Mutsu Institute for Oceanography

Tokyo office
東京都千代田区内幸町

Yokohama institute for Earth Sciences

Yokosuka HQ

Global Ocranographic Data Center（GODAC)

Kouti Institute for Core Sample Research

JAMSTEC Research Vessels

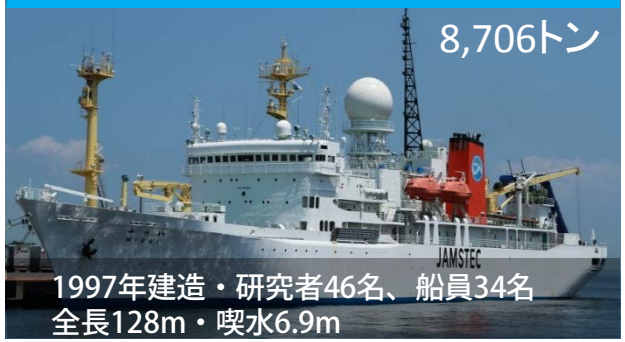
7 research ships to contribute with the issue of the earth environment, earthquake & tsunami and the resource survey.

深海調査研究船 かいれい

4,517トン

1997年建造・研究者22名、船員38名
全長106m・喫水4.7m

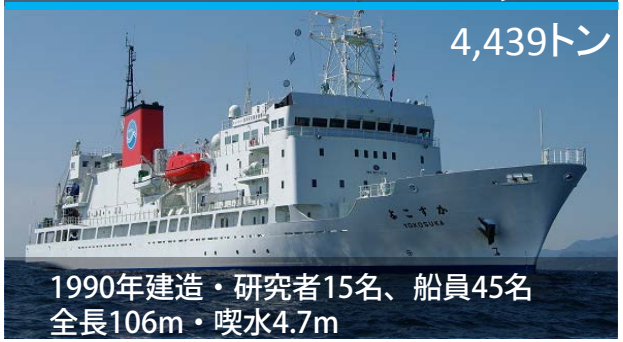
海洋地球研究船 みらい



8,706トン

1997年建造・研究者46名、船員34名
全長128m・喫水6.9m

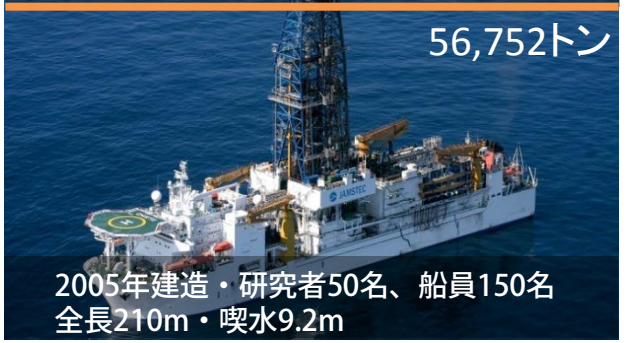
深海潜水調査船支援母船 よこすか



4,439トン

1990年建造・研究者15名、船員45名
全長106m・喫水4.7m

地球深部探査船 ちきゅう



56,752トン

2005年建造・研究者50名、船員150名
全長210m・喫水9.2m

学術研究船 白鳳丸

3,991トン

1989年建造・研究者35名、船員54名
全長100m・喫水6.0m

東北海洋生態系調査研究船 新青丸

1,629トン

2013年建造・研究者15名、船員26名
全長66m・喫水5.0m

海底広域研究船 かいめい

5,747トン

2016年建造・研究者38名、船員27名
全長100m・喫水6.0m

Research Vehicles & Observation System

Various type of data including picture and movie also come from them…

有人潜水調査船 しんかい6500

1989年建造・最大潜航深度 6,500m
搭乗可能人員3名　全長9.7m・幅2.7m
速力 0〜2.7ノット

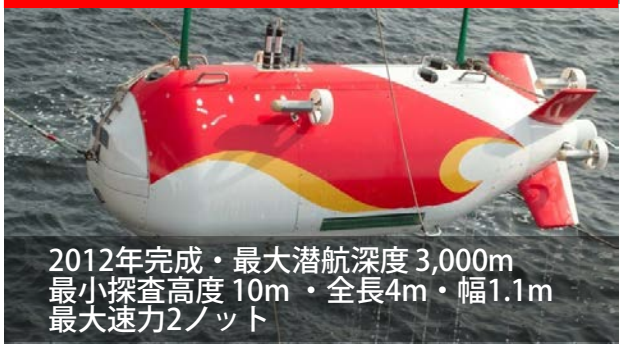
自律型無人探査機 うらしま

2000年完成・最大潜航深度 3,500m
航続距離 300km以上・全長10m・幅1.3m
速力 3ノット

自律型無人探査機 ゆめいるか

2012年完成・最大潜航深度 3,000m
最小探査高度30m ・ 全長5m・幅1.2m
最大速力3ノット

自律型無人探査機 じんべい



2012年完成・最大潜航深度 3,000m
最小探査高度 10m ・全長4m・幅1.1m
最大速力2ノット

自律型無人探査機 おとひめ

2012年完成・最大潜航深度3,000m
着底して観測可能 全長2.5m・幅1.4m
最大速力1.5ノット

無人探査機 ハイパードルフィン

1999年完成・最大潜航深度 3,000m
全長3.0m・幅2.0m・3.8トン

無人探査機 かいこう7000 II

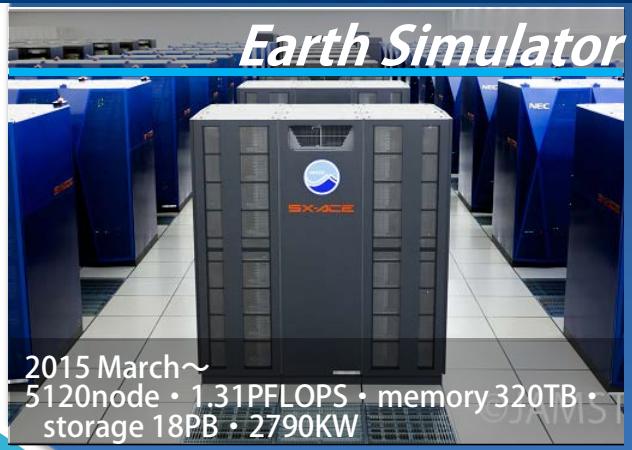
2006年運用開始・最大深度 7,000m
全長5.2m・幅2.6m・9.7トン

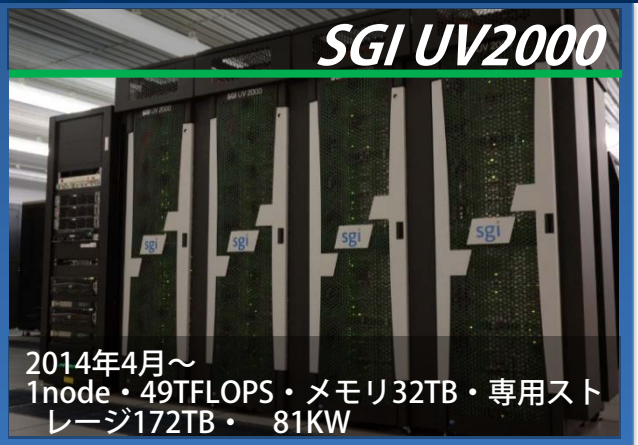
無人探査機 かいこうMk-IV

2013年完成・最大深度 7,000m・全長3m
幅2m・6トン　海底での重作業が可能

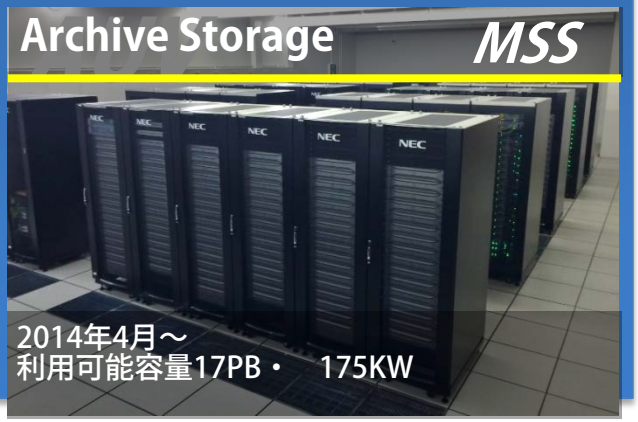
JAMSTEC Super Computer and Storage

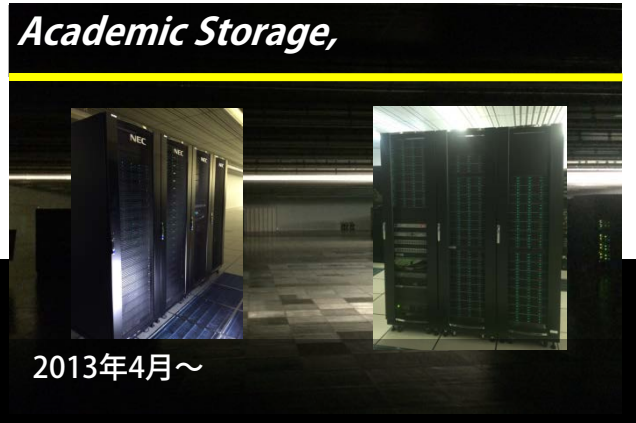
Supercomputers & Storages for the simulation data





NEC SC-ACE





- SYSTEM :

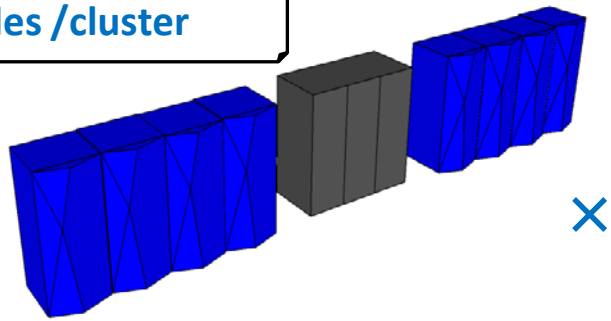
  Our Supercomputer and Storage
  (File staging, File system)

Earth Simulator Hardware (Specification)

		Earth Simulator
CPU	Clock	1.0GHz
	Peak Performance of Each Core, Each CPU	64GF(1core) 256GF(1CPU)
	Memory Bandwidth	64-256GB/s
	ADB size	1MB （1core） x4
Node	#CPU	1(4core)
	Peak Performance of Each Node	256GF
	Memory/Node	64GB
	Inter-node Transfer Speed	4GB/s x2
System	Total Number of Processor Nodes	5120
	Peak Performance of System	1.31PF
	Main Memory	320TB
	Interconnection Network	Fat-tree Network

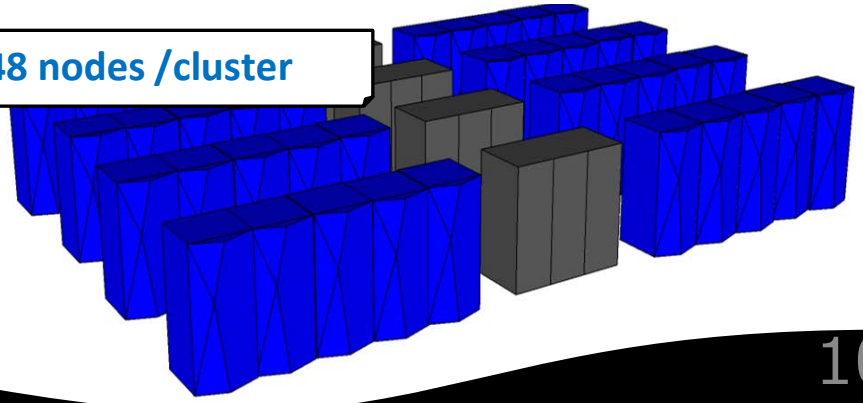
1 rack(64 nodes)

512 nodes /cluster

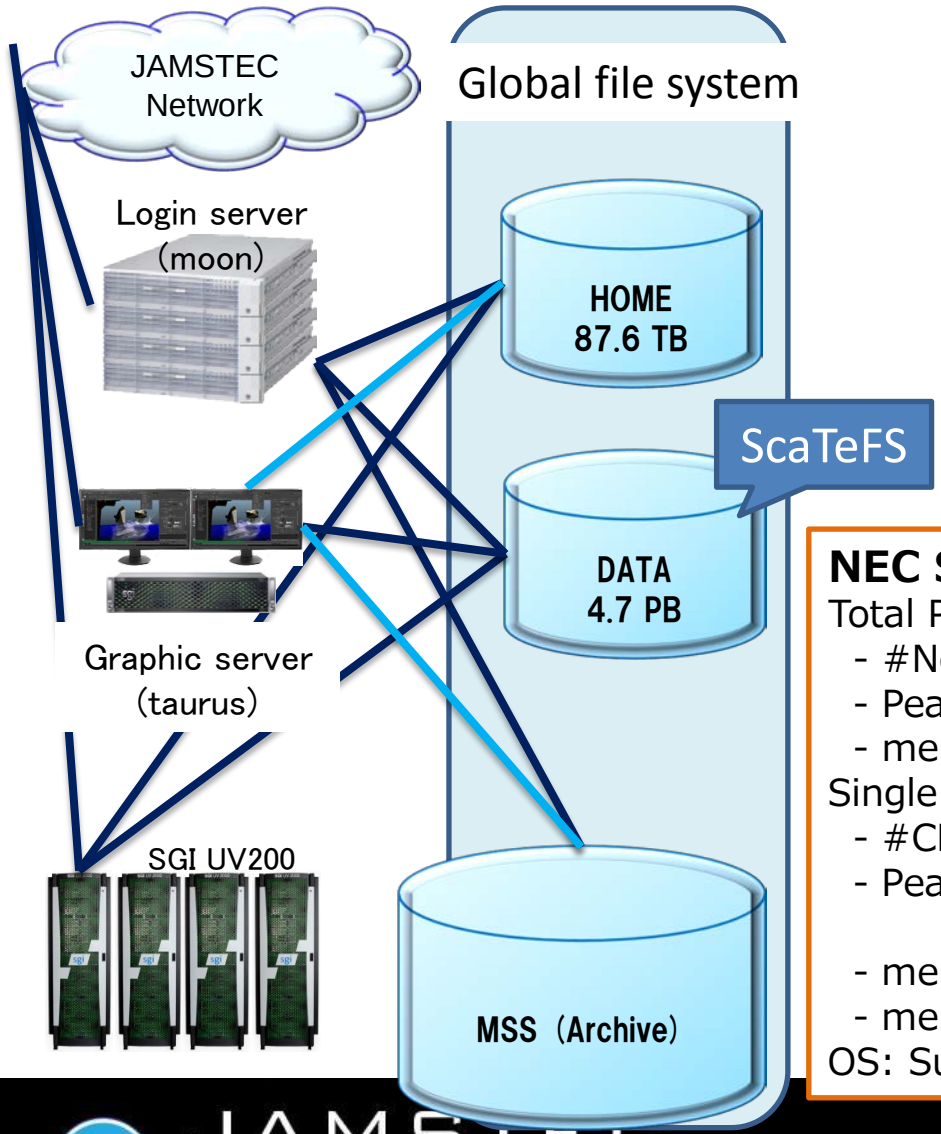


×6

2048 nodes /cluster



Abstract of Earth Simulator (ES) system

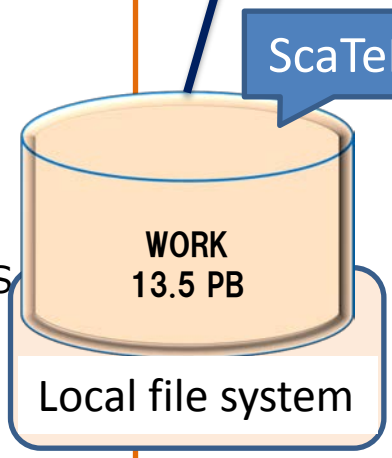


NEC SX-ACE

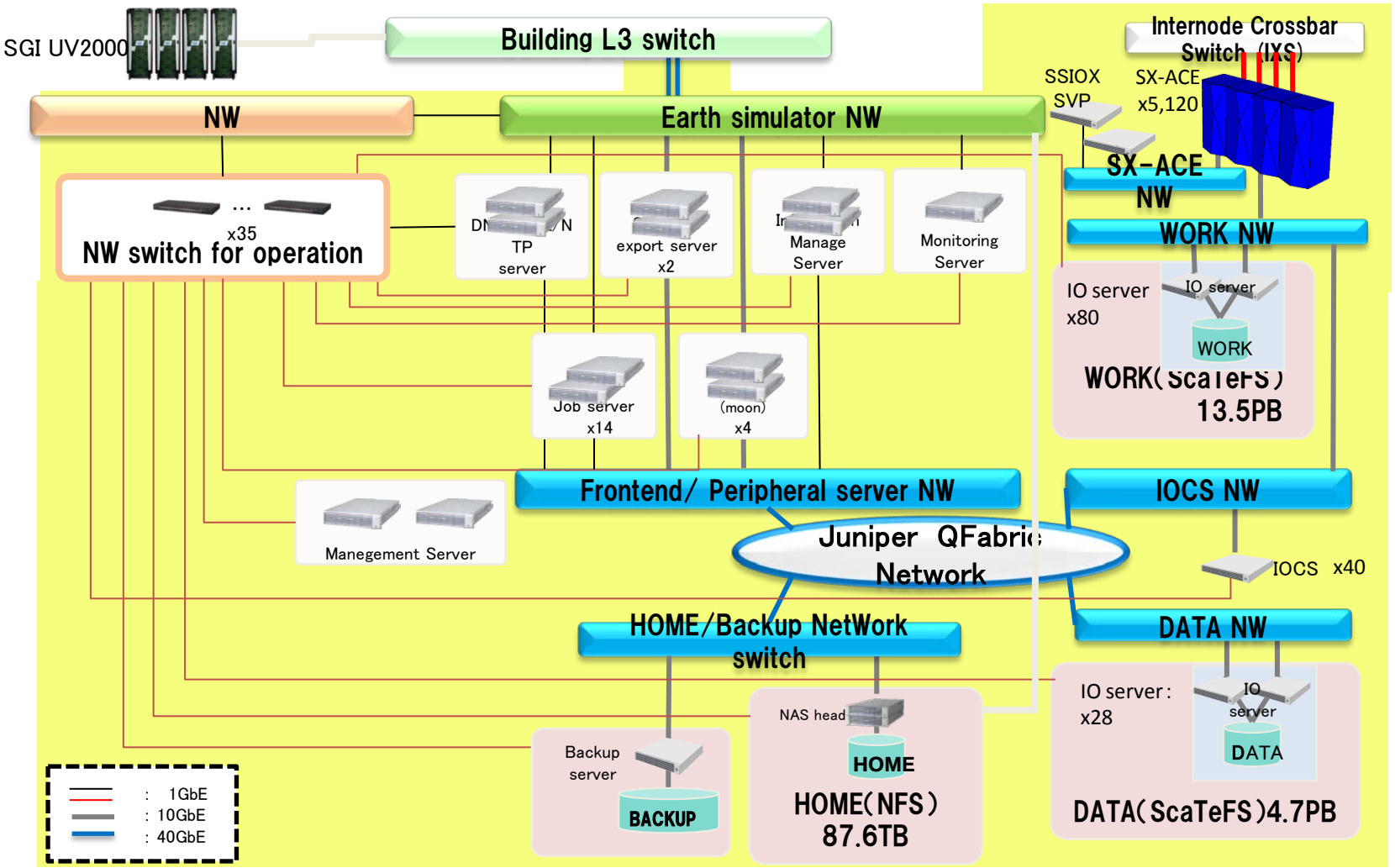
Total Performance
- #Node : 5,120 nodes
- Peak Performance : 1.31 PFLOPS
- memory : 320 TB

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

OS: SuperUX

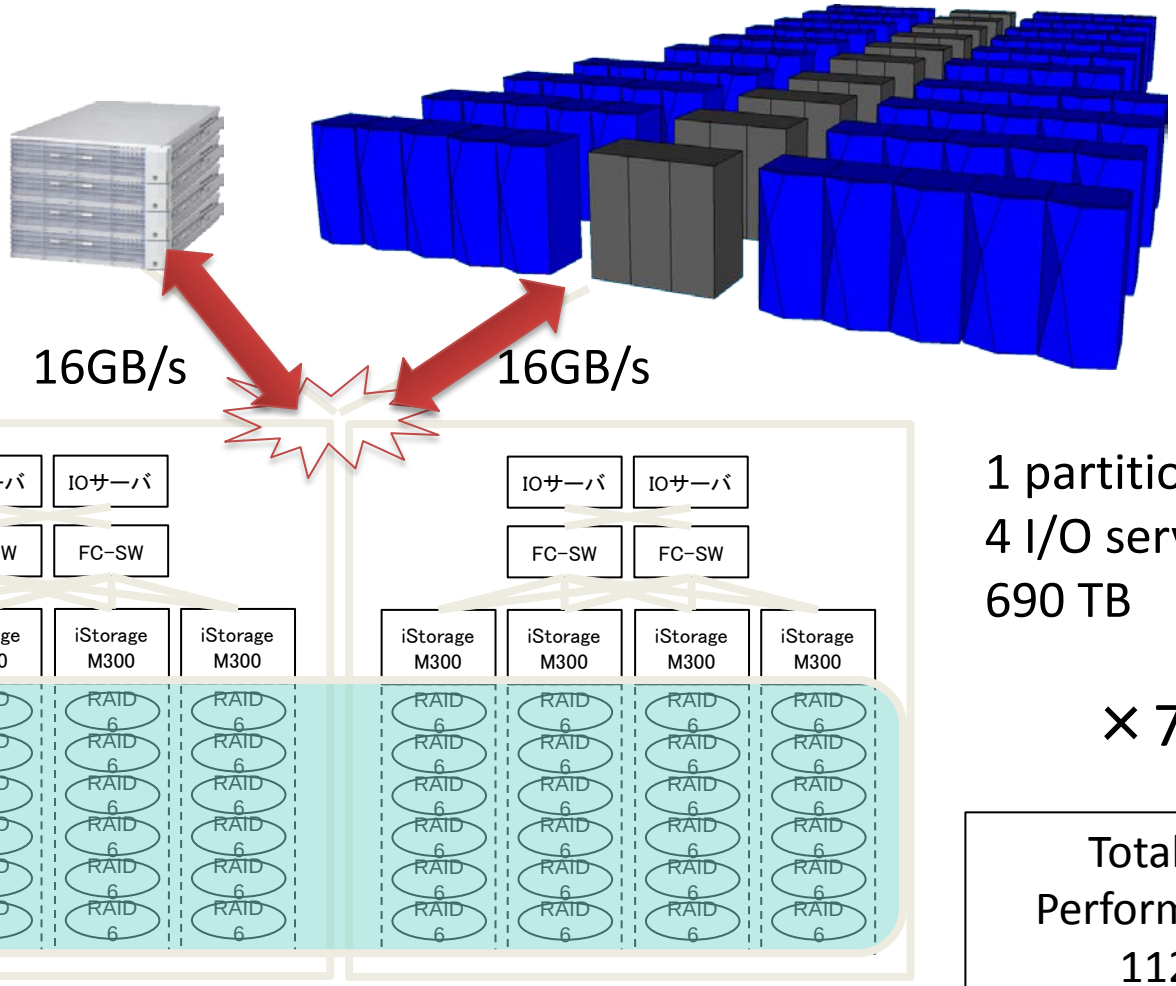


Abstract of ES Network



DATA Region

Login Server
(moon)

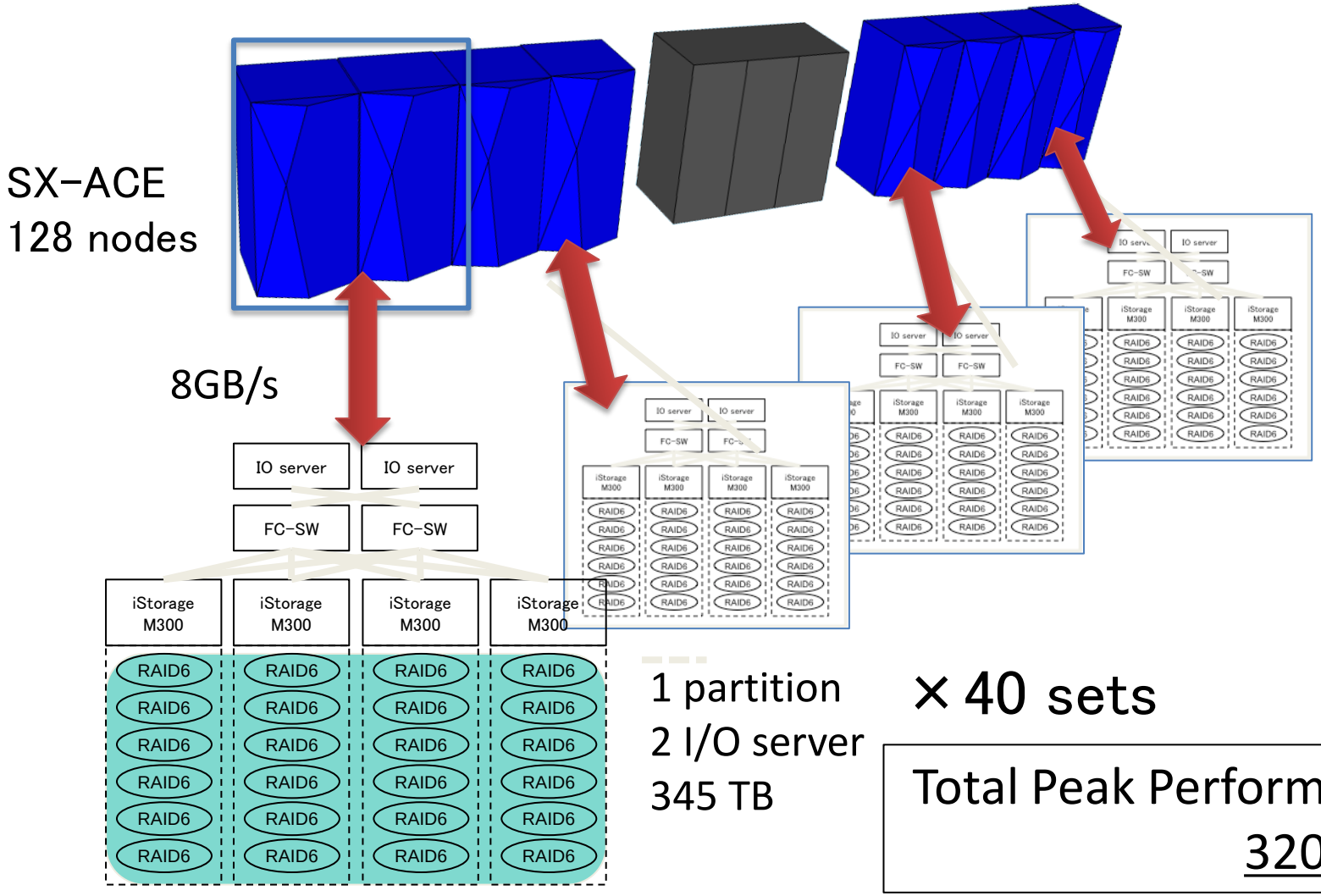


1 partition
4 I/O server
690 TB

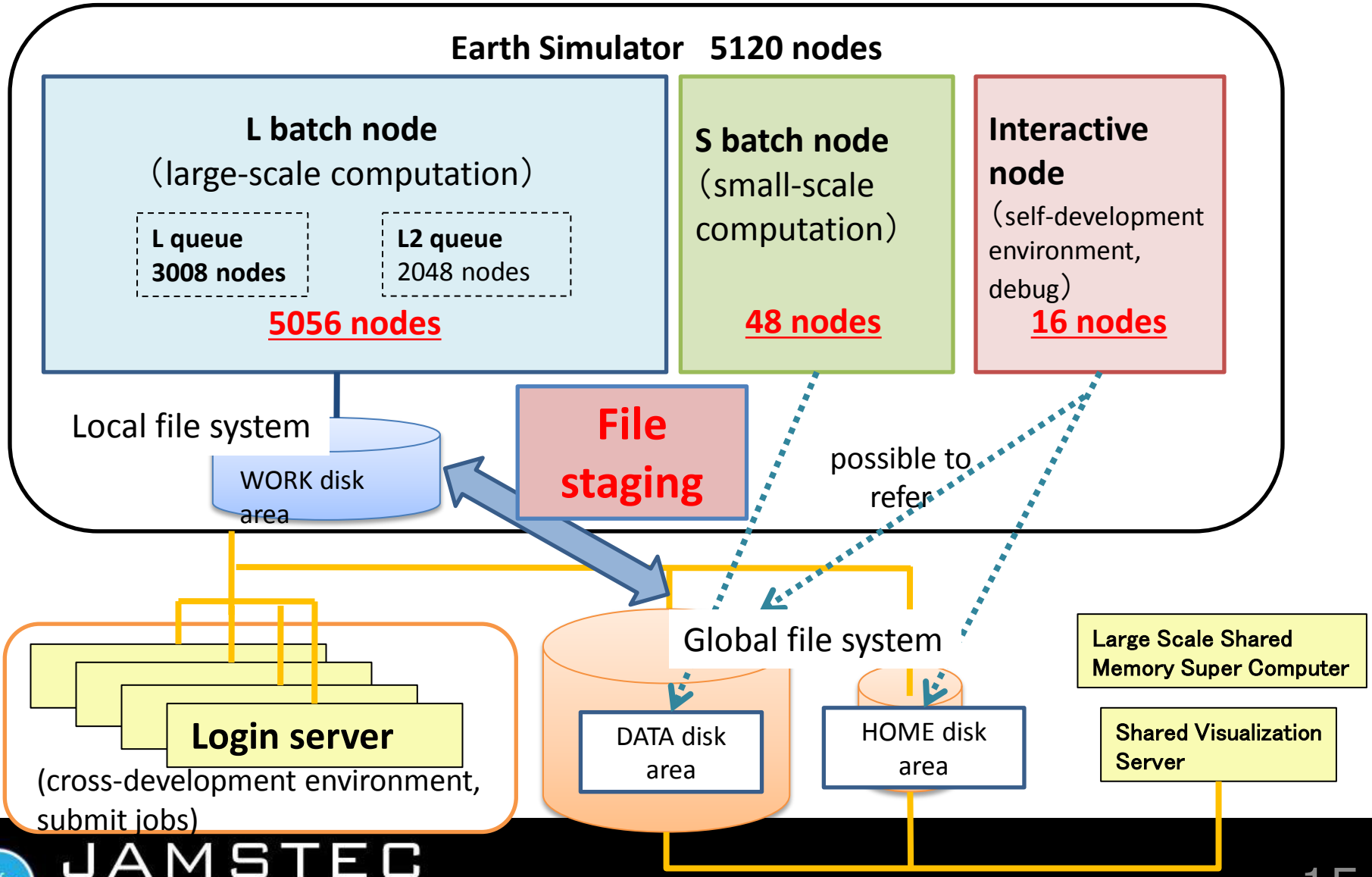
× 7 sets

Total Peak Performance
112GB/s

WORK Region


SX-ACE
128 nodes
8GB/s
IO server
IO server
FC-SW
FC-SW
iStorage M300
iStorage M300
iStorage M300
iStorage M300
RAID6
1 partition
2 I/O server
345 TB
× 40 sets
Total Peak Performance
320GB/s

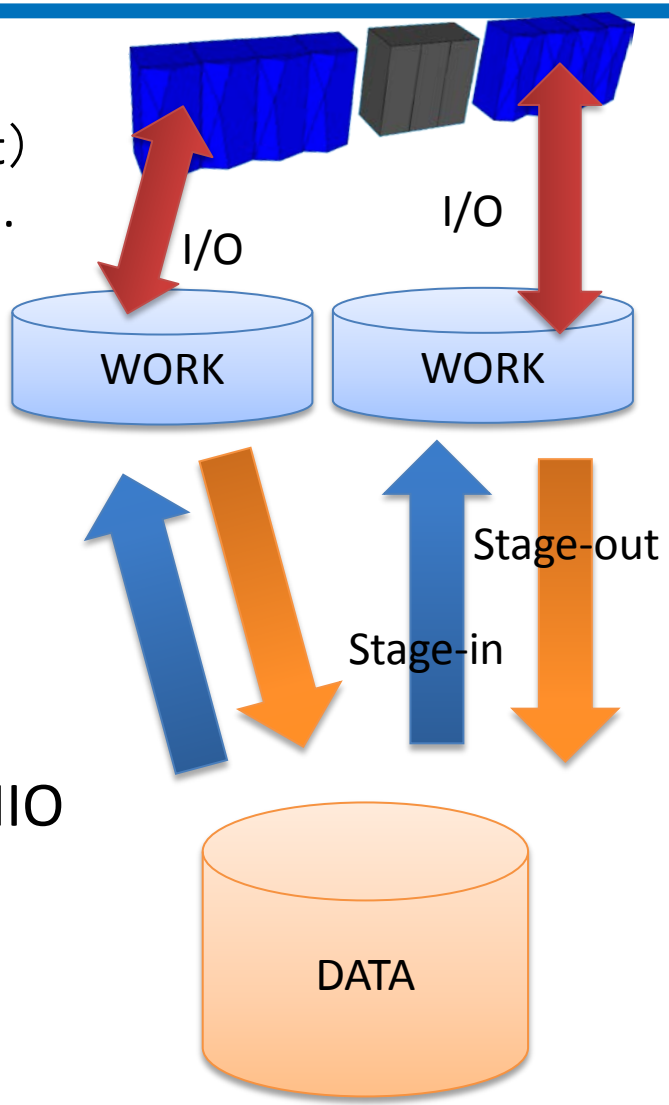
File Staging system



File Staging system

Copy file automatically between global file system and local file system before/after jobs（stage-in & stage-out）by jobs scheduler considering the rest of the disk space.

- Merit
  - Prevent small I/O conflicts among applications
- Demerit
  - Cost & Size
    DATA 4.7PB ＜ WORK 13.5 PB!
  - Not shear the files between nodes; MPIIO is not available

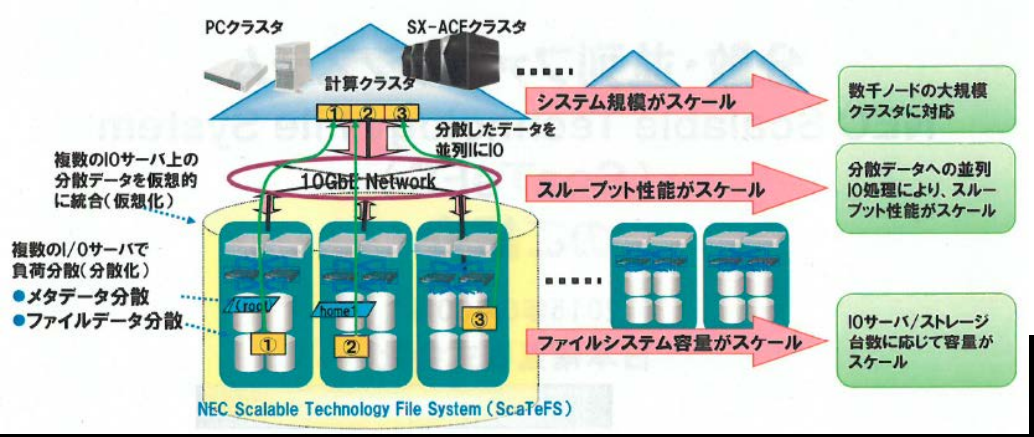


About File System: ScaTeFS (by NEC)

ScaTeFS: NEC Scalable Technology File System

NEC says the features of ScaTeFS are;

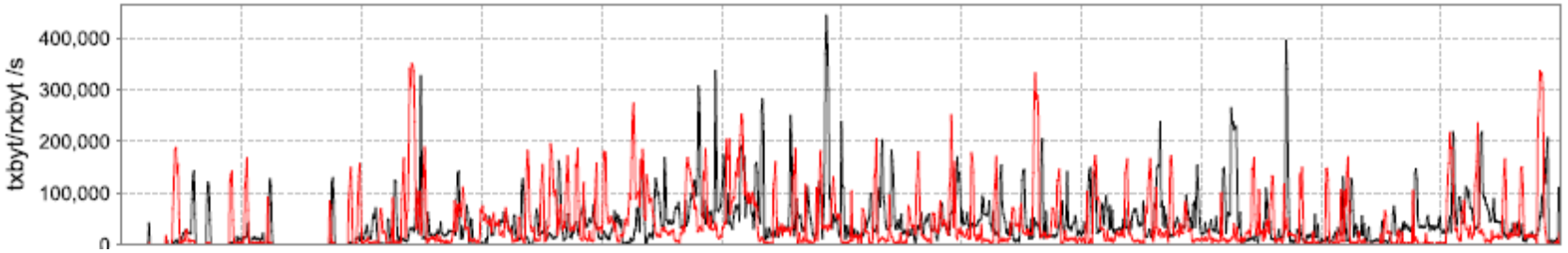
1. Scalable I/O throughput
2. Distribute data and metadata
3. High reliability
4. InfiniBand ready (ES use 10GbE)



- Evaluate I/O performance:
  - BM: mdtest, IOR

sar(read&write): Work vs Data

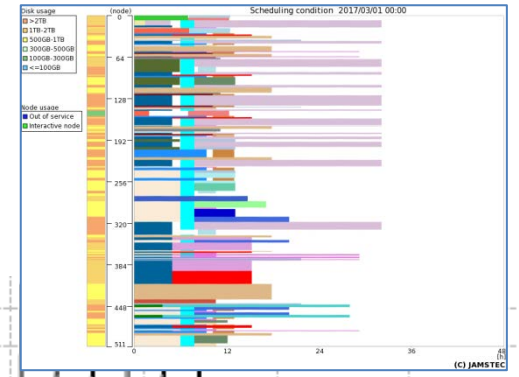
Data (Global File System)

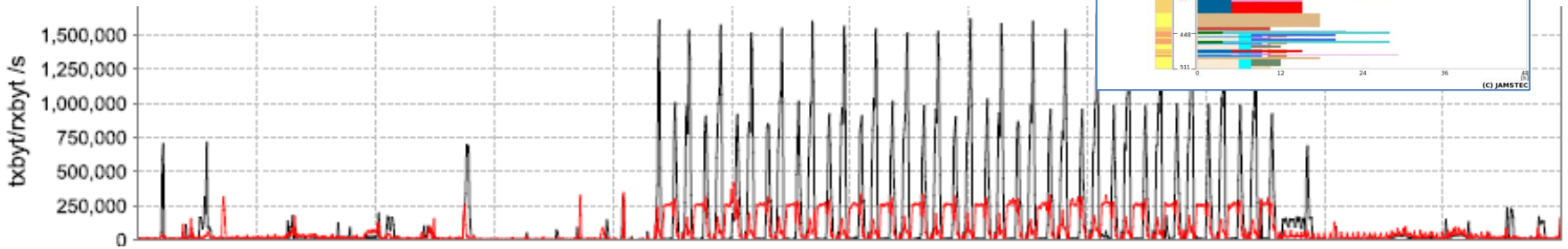


Work (Local File System)

- I/O work load: data < work
- I/O Patterns → Use for a realistic I/O BM?

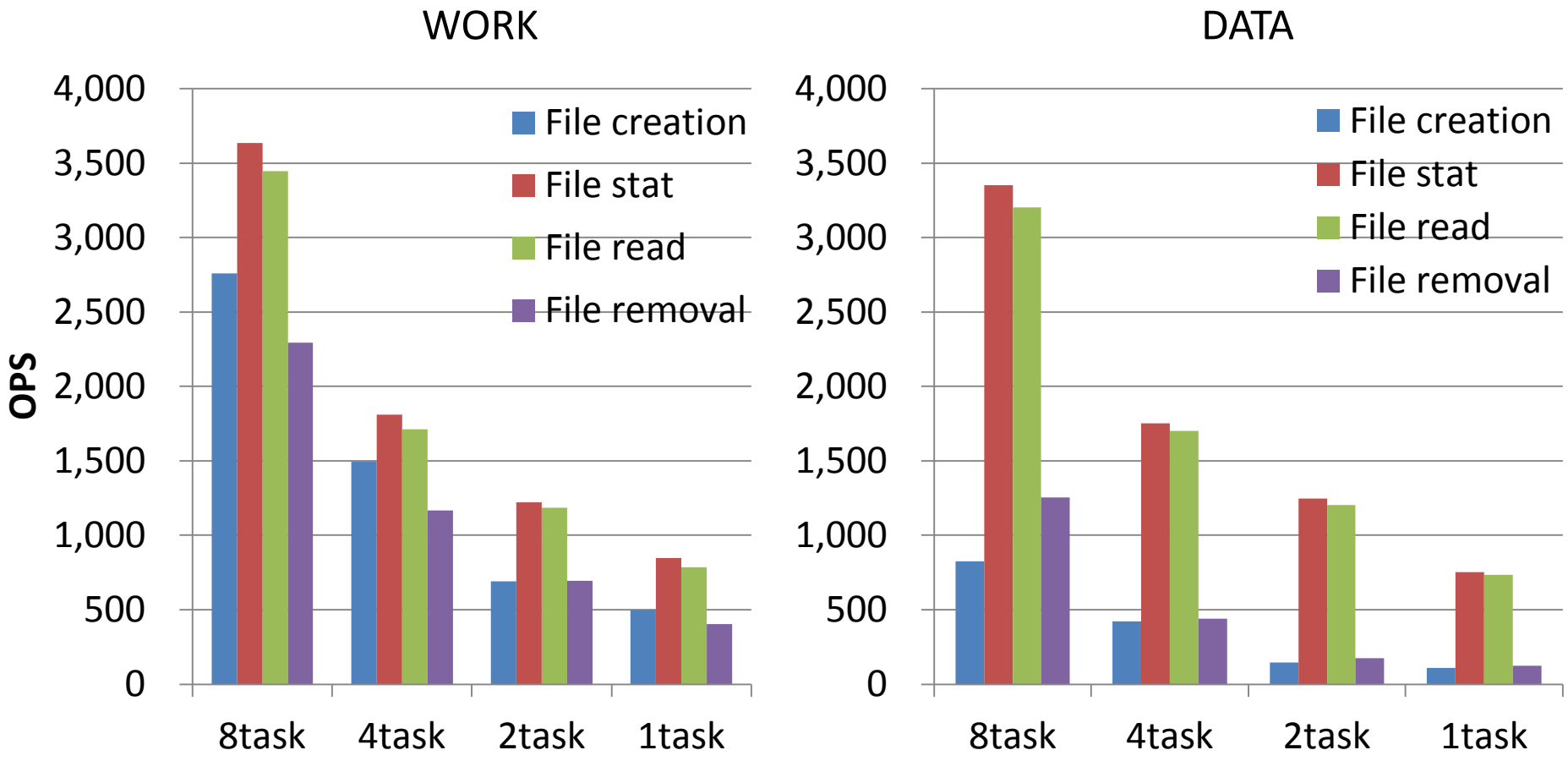
job scheduling





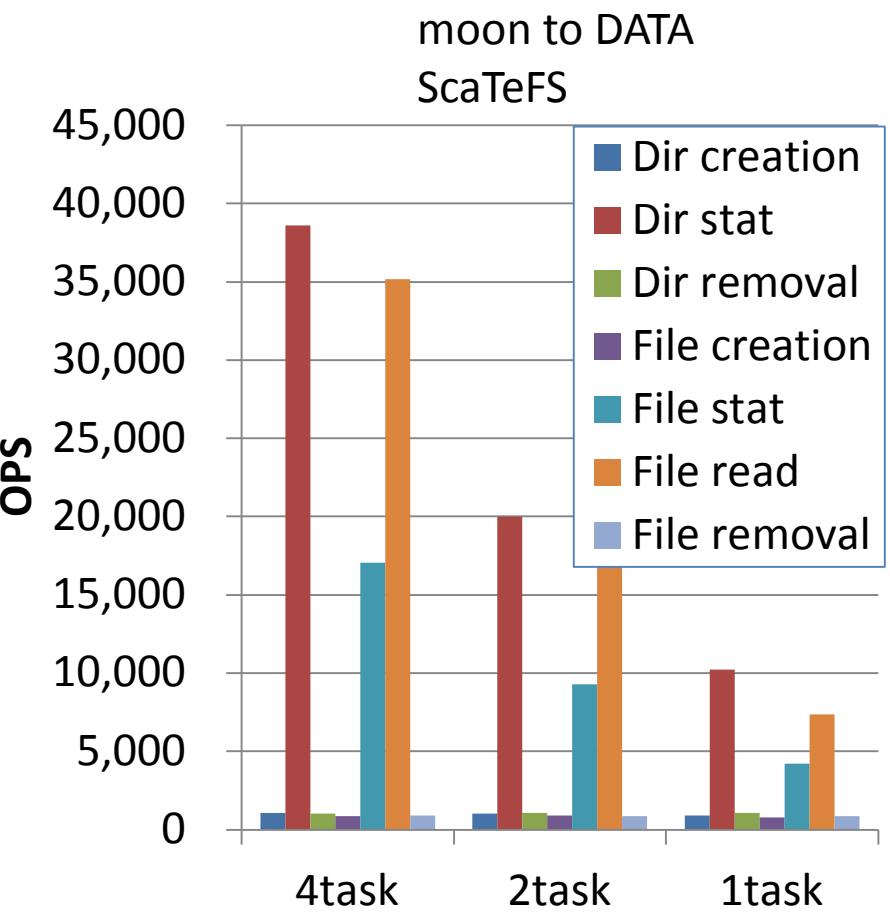
ES job scheduling condition in March 1

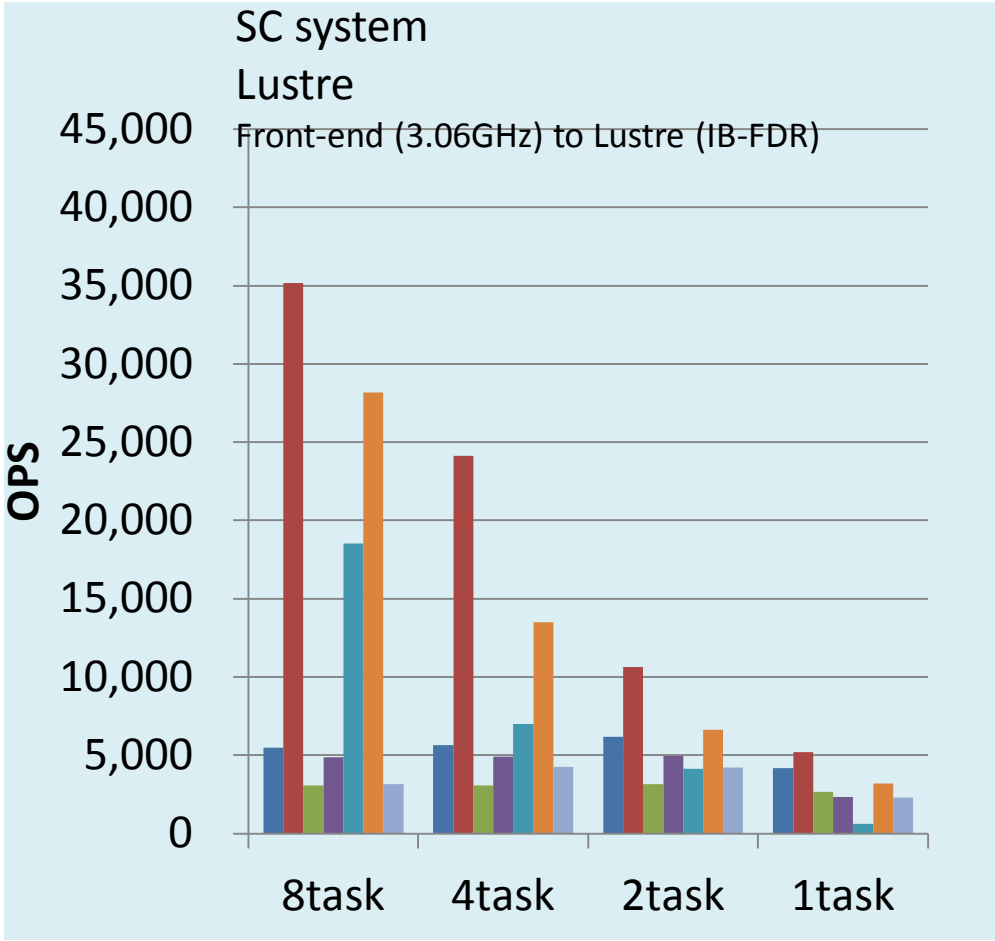
mdtest from SX-ACE



mpirun –nn 1 -np X ./mdtest –n 5000 -i 5 –p 10 –u
(-u: unique directory)

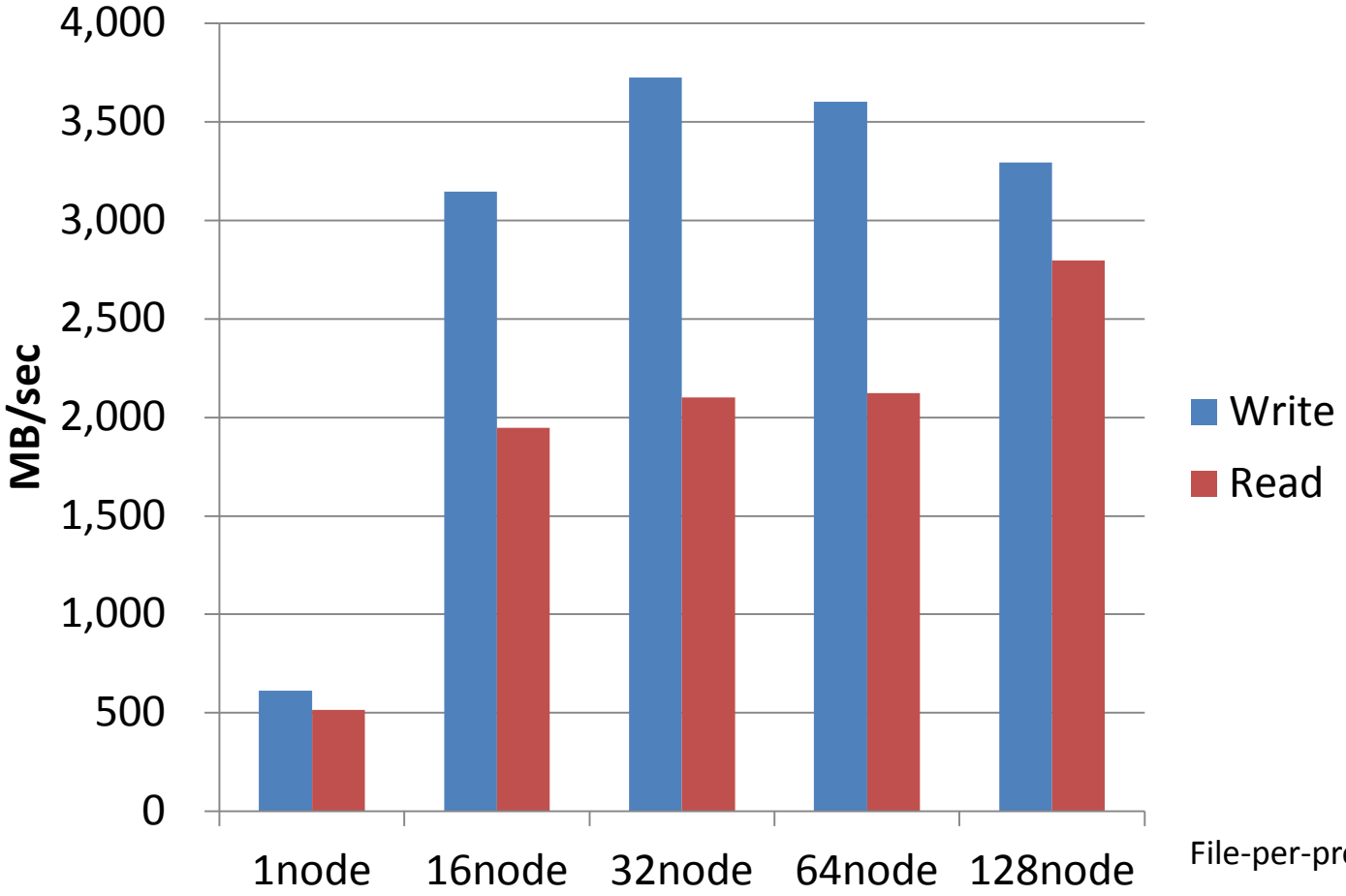
mdtest from login server





mpirun -np X ./mdtest –n 5000 –i 5 –p 10

IOR from SXACE



File-per-process using POSIX

mpirun -np X -ppn 4 ./IOR -t 4M -i 3 -b 256m -F –s Y -vv -o Work_dir

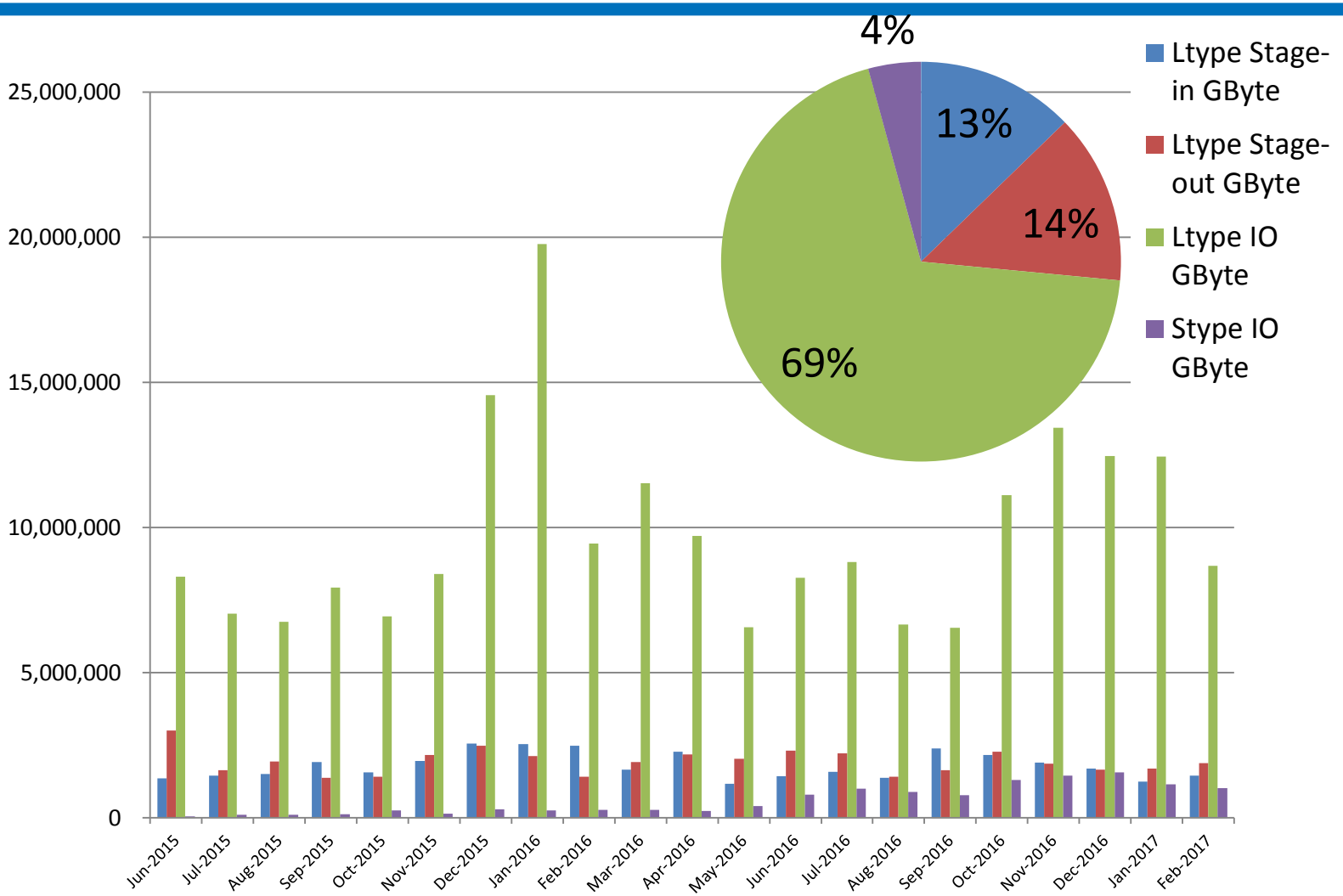
- I/O Statistic monitoring:

Application using in JAMSTEC

Most of our (in-house) simulation code written by FORTRAN.

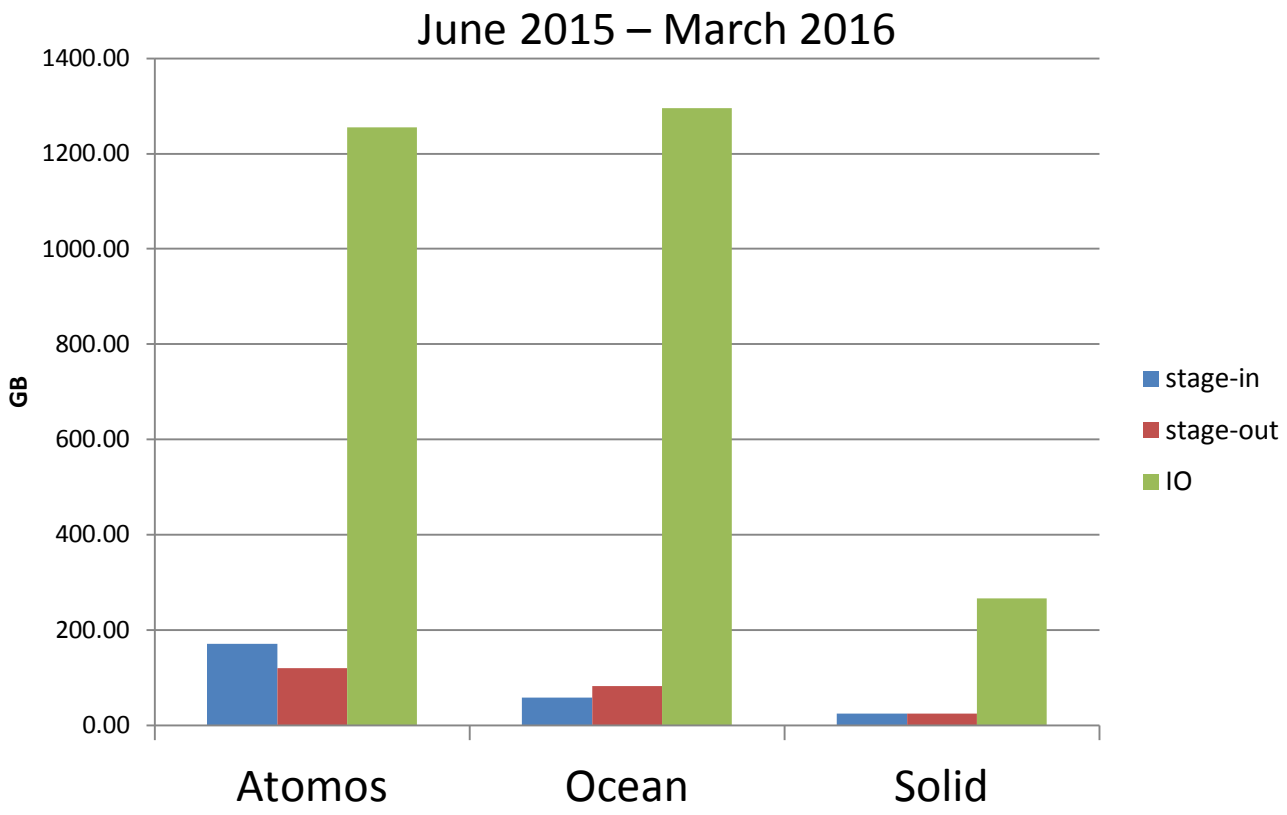
- Weather & Climate model
  - WRF, MSSG, NICAM, MIROC
  - ROMS, COCO, M4DV(data assimilation)
- Earthquake, Tsunami, Mantle convection
  - SPECFEM3D, JAGURS, ACuTEMan
- Bio-infomatics(Genome)
  - BLASTX, GhostX, Mothur
- Data processing such as Satellite data
  - R, python, perl, ruby
- Chemistry & Material
  - Phase
- R&D for research vessels
  - MATLAB, COMSOL Multiphysics

I/O Statistics Analysis

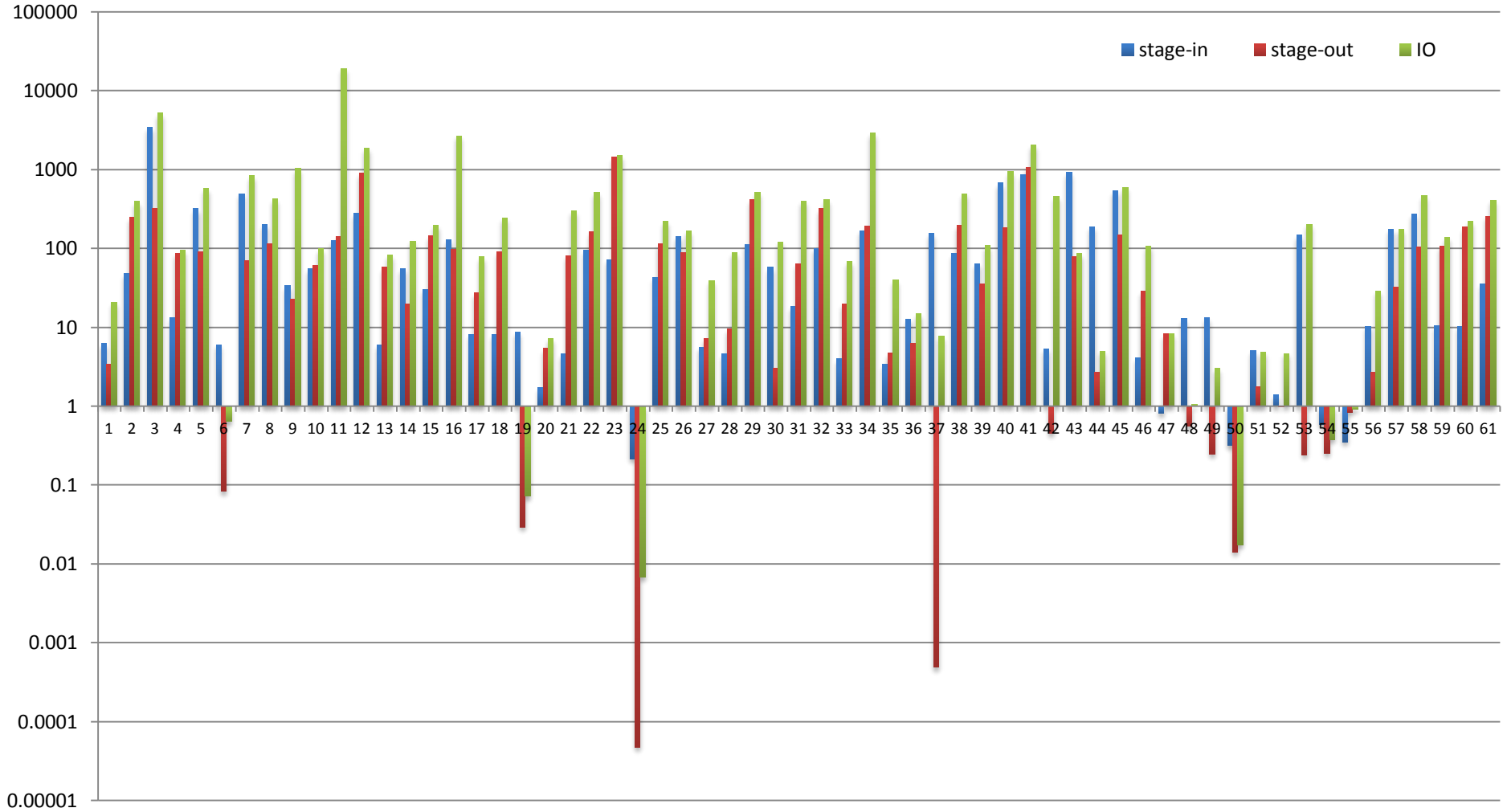


I/O Statistics Analysis: Aver. I/O per job

- 69 research projects
- total L batch requests: 251,955



Statistics I/O Analysis: Ave. I/O over each projects



Summary

- To extend the research target and combine them, I/O performance is critical
- File staging system is good for jobs. But trade-off on cost, user disk space and availability. We need to evaluate DATA vs WORK
- Next generation of file staging system would be realized by using burstbuffer or NVMe?
- Our next steps:

  Application I/O profiling for the realistic I/O BM

- Thank you for your attention!