Performance evaluation of Lustre on All-Flash Storage system at OIST

Eddy Taillefer – SCDA, OIST
Koji Tanaka – SCDA, OIST
Shuichi Ihara – Whamcloud

OIST

Whamcloud

OKINAWA INSTITUTE OF SCIENCE AND TECHNOLOGY GRADUATE UNIVERSITY
沖縄科学技術大学院大学

1. OIST - Okinawa Institute of Science and Technology

2. SCDA - Scientific Computing and Data Analysis Section

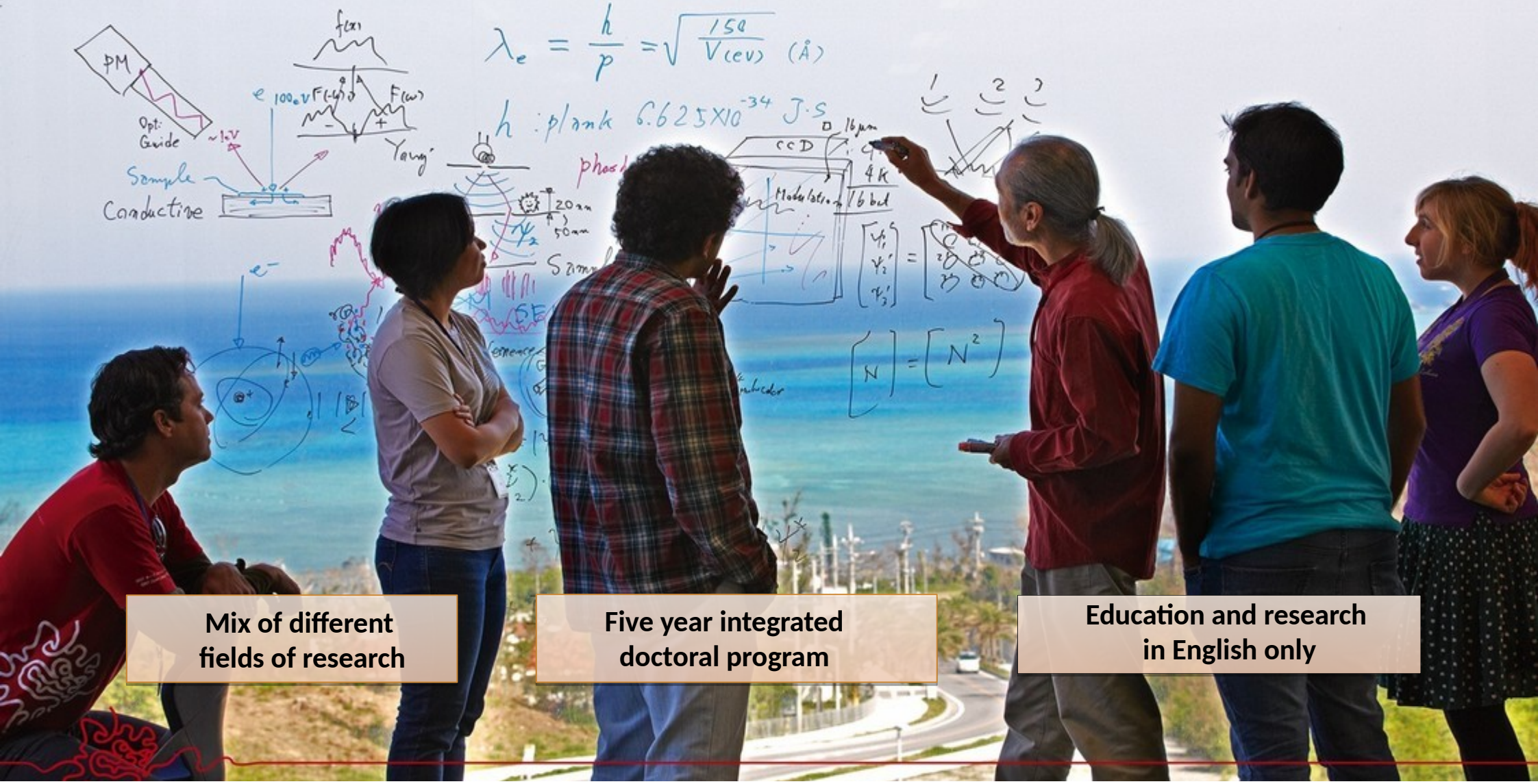
3. Benchmark Configuration and Results

4. Research Computing Example

OIST Campus

New Style Graduate University

Inaugural PhD class began Sep 3, 2012

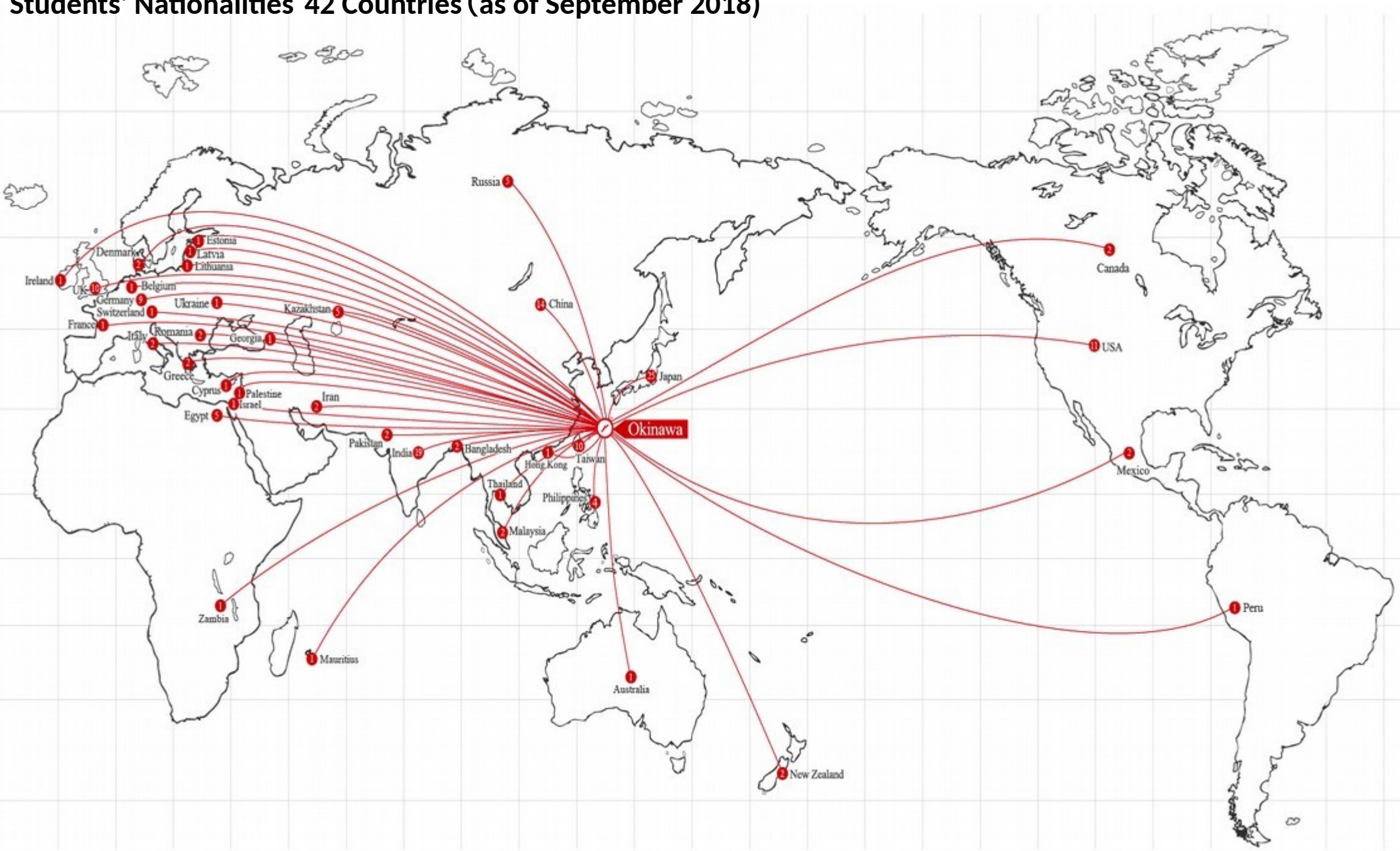


Mix of different fields of research

Five year integrated doctoral program

Education and research in English only

Students' Nationalities 42 Countries (as of September 2018)



OIST is

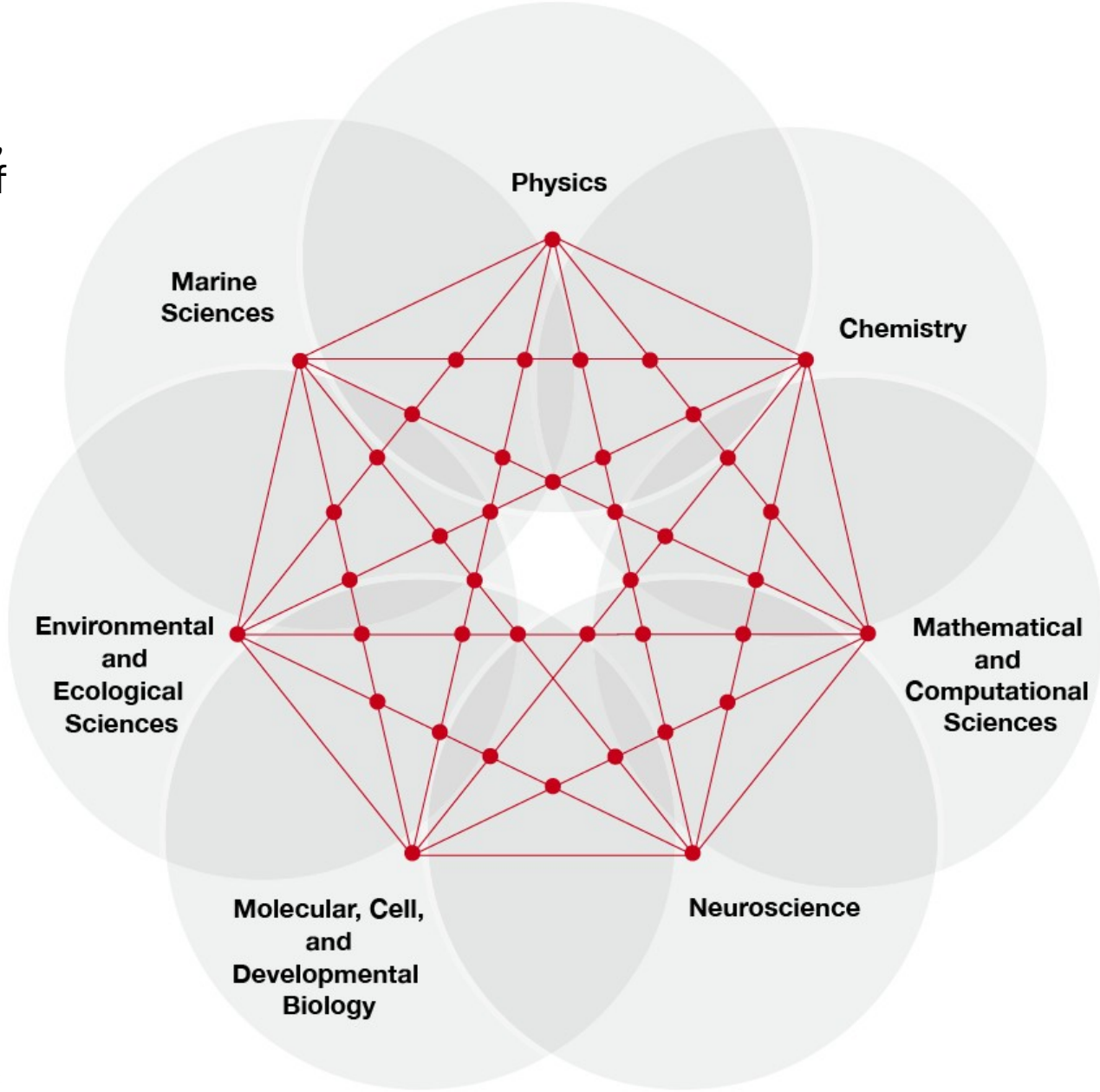
Over 1000 members

174 students, 65 professors, 446 researchers, 407 research support and administrative staff members

Seven Fields of Research

- Physics
- Chemistry
- Neuroscience
- Marine Science
- Environmental and Ecological Sciences
- Mathematical and Computational Sciences
- Molecular, Cellular, and Developmental Biology

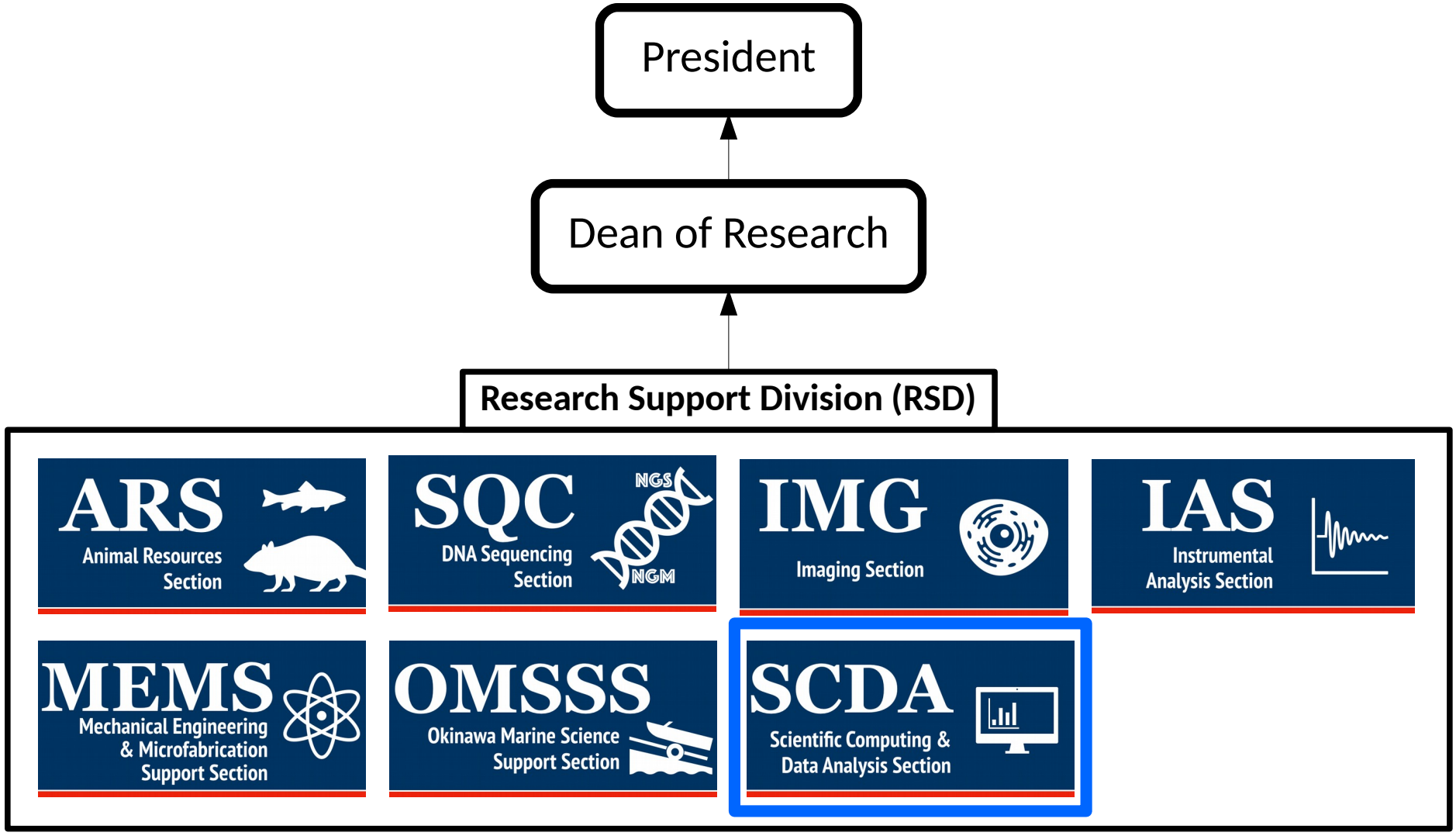
Cross-Disciplinary Approach



OIST's Key Comcepts

- World Leading Education and Research
- International
- Interdisciplinary
- Global Networking
- Collaboration with Industry

OIST Research Support



1. OIST - Okinawa Institute of Science and Technology

2. SCDA - Scientific Computing and Data Analysis Section

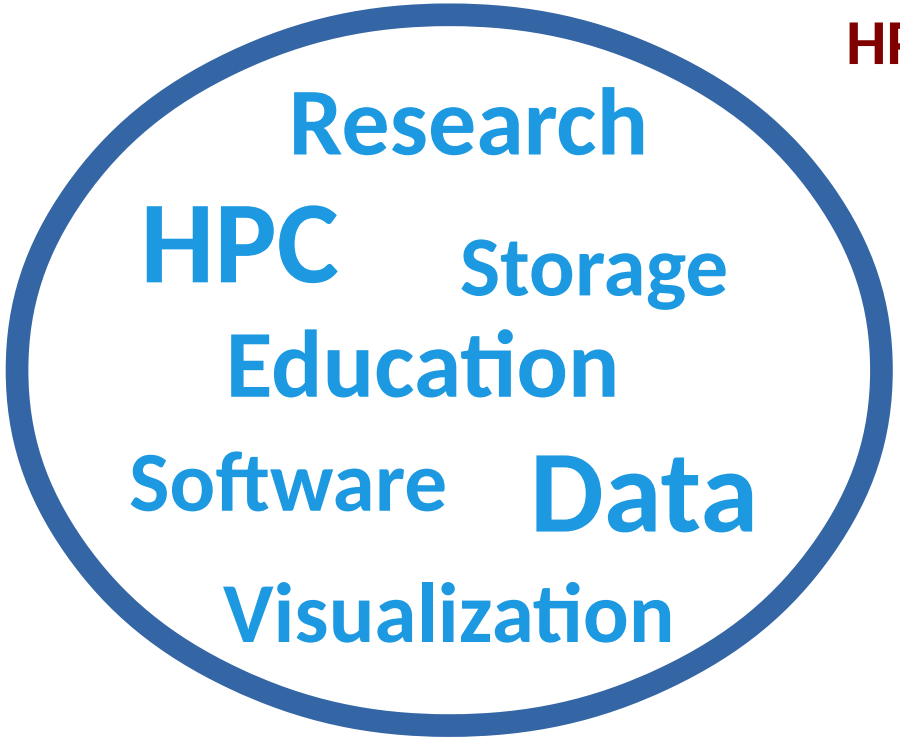
3. Benchmark Configuration and Results

4. Research Computing Example

SCDA - Scientific Computing and Data Analysis Section

Research Computing Support

- Research support
- Scientific software
- Open office hours
- Training and seminars
- HPC documentation
- *etc.*



HPC and Computing Resources

- Resource management
- Resource monitoring
- User support
- Research project assistance
- How-tos and wikis
- *etc.*

Scientific Services

- Data acquisition and analysis
- Scientific data visualization
- Web-based systems
- Training and seminars
- High performance databases
- *etc.*

Research

Acquisition instruments

Sequencers

Generated research data

OIST Cloud special account

Storage

/work

3000TB (scratch space)
50TB ~ available / unit

/bucket

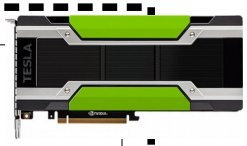
6000TB (research data)
50TB ~ available / unit

7500TB (tape archiving/backup)
Automatic incremental backup
Data archiving and restoring

/home

User home area
50GB / user

Saion (Highly parallel cluster):
4096 Intel KNL threads
(16x 64-core Intel KNL with 128GiB)
96 Tesla P100/V100 GPU
(14x 36-core Intel Xeon with 512 GiB)
200TB scratch space, 10TB per unit



Sango **(Main cluster)**:
10704 computing cores (Haswell)over 446 nodes
12 K80 dual-GPU and 16 P100 GPU (IBM-power)
400 generic nodes with 128 GiB of memory
46 large memory nodes with 512GiB/768GiB
One 3072GiB big memory compute node

HPC

/work

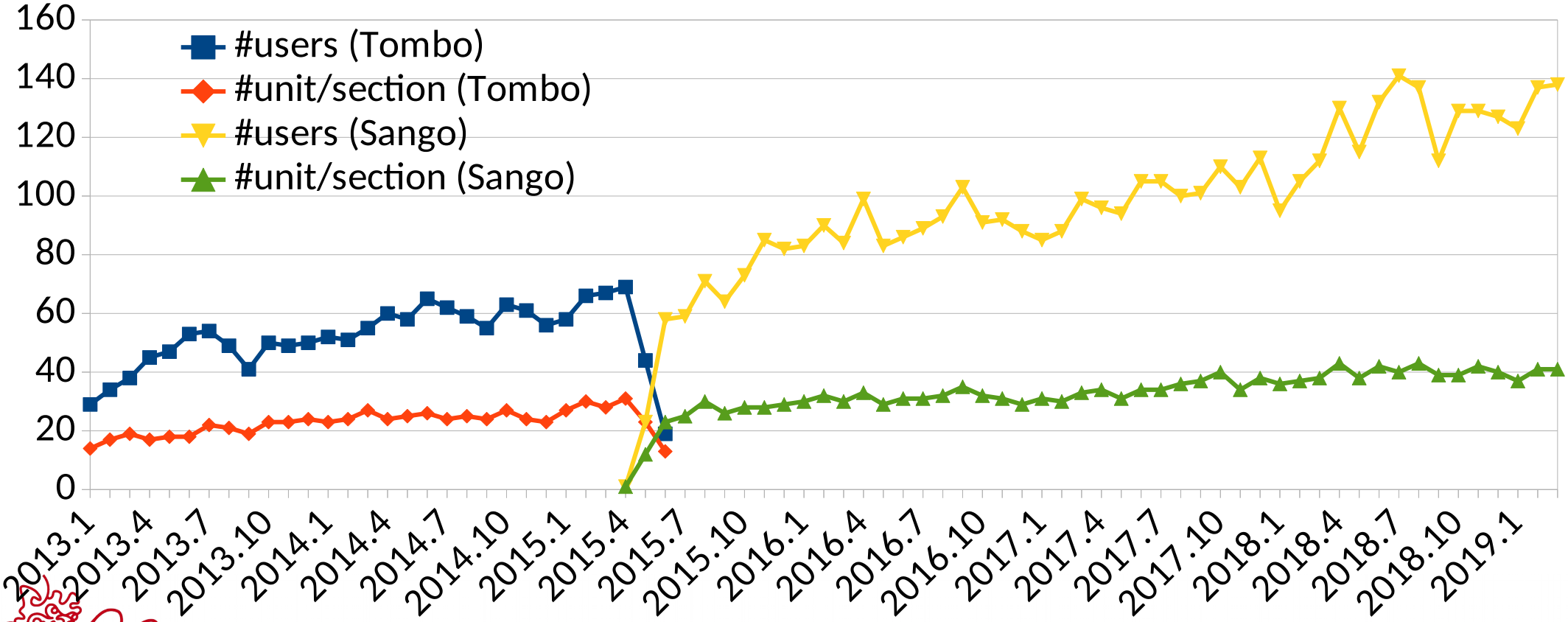
OIST LUG 19

https://groups.oist.jp/scs/equipment-gallery

History of HPC and Lustre at OIST(1)

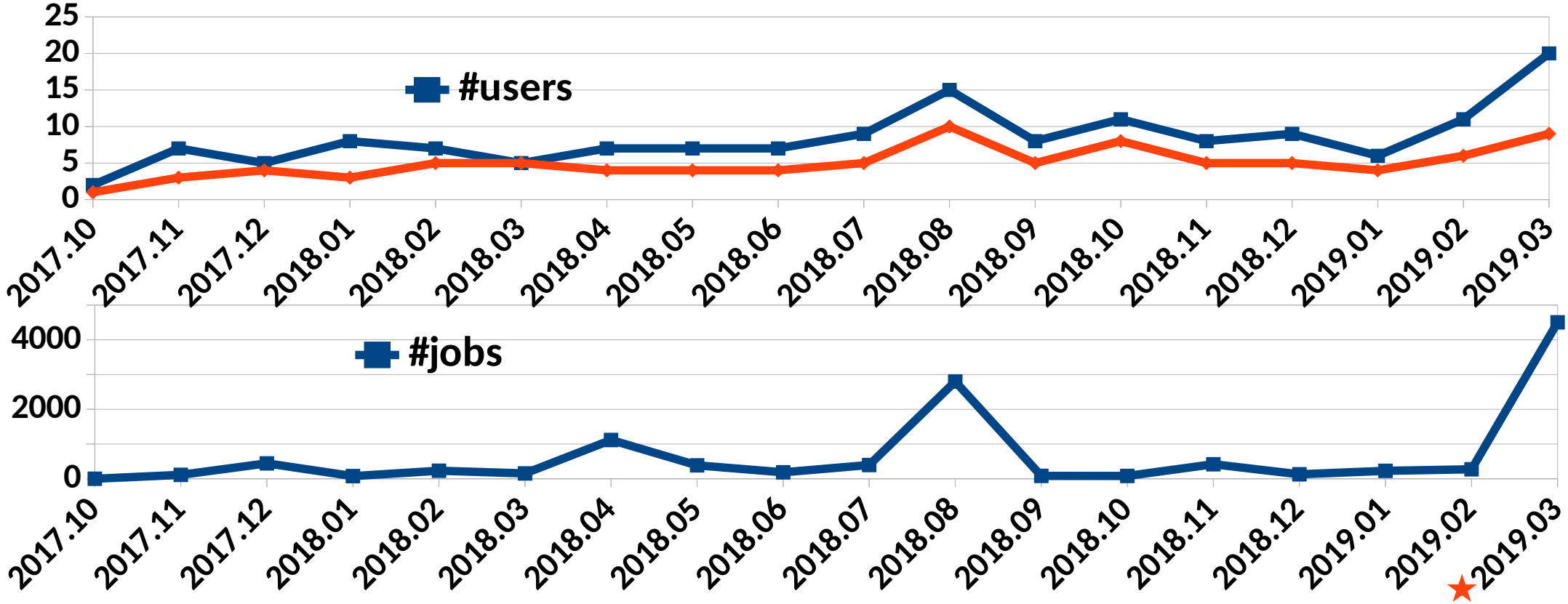
Tombo (2011~2015)
- 384 compute nodes
- QDR Infiniband / 10G Ethernet
- Lustre-2.1
- 400TB (HDDs)

Sango (2015~)
- 466 compute nodes
- FDR Infiniband
- Lustre 2.5, upgraded to 2.7 and going to upgrade lustre-2.10
- 3PB (HDDs)



History of HPC and Lustre at OIST(2)

Saion (2017~)
- 48 compute nodes
- EDR Infiniband
- Lustre 2.10, upgrade to 2.12 in the future
- 230TB (All-Flash)



★ Lustre with All-Flash installed

Overview of Saion - 蔡温

High Core-count Nodes

1440 cores

4x

2x 20-core Intel Xeon
512 GiB Memory / 960 GB local SSD

4x

2x 32-core AMD Napples
512 GiB Memory / 960 GB local SSD

16x

64-core Intel Xeon Phi (KNL)
192 GiB Memory
500 GB local hard drive

GPGPU Nodes

96 GPUs

14x

2x 18-core Intel Xeon
4x nVidia P100 GPU
512 GiB Memory
960GB local SSD drive

8x

2x 18-core Intel Xeon
4x nVidia V100 GPU
512 GiB Memory
1.6TB local NVMe drive

2x

2x 20-core IBM Power9
4x nVidia V100 GPU
512 GiB Memory
960GB local SSD drive

Campus 10Gb/s

Infiniband EDR 100Gb/s

230TB All-Flash storage

OIST LUG 19

1. OIST - Okinawa Institute of Science and Technology

2. SCDA - Scientific Computing and Data Analysis Section

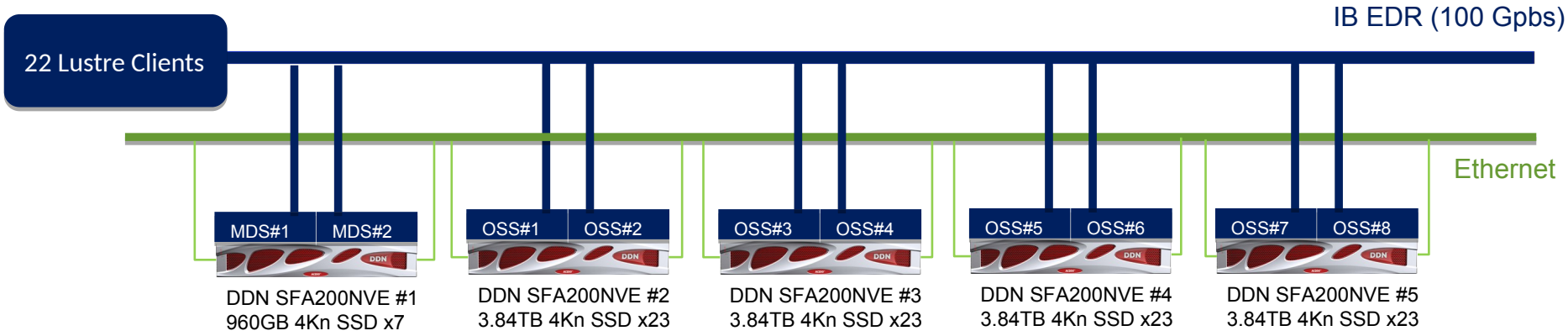
3. Benchmark Configuration and Results

4. Research Computing Example

Benchmark setup and method

Setup

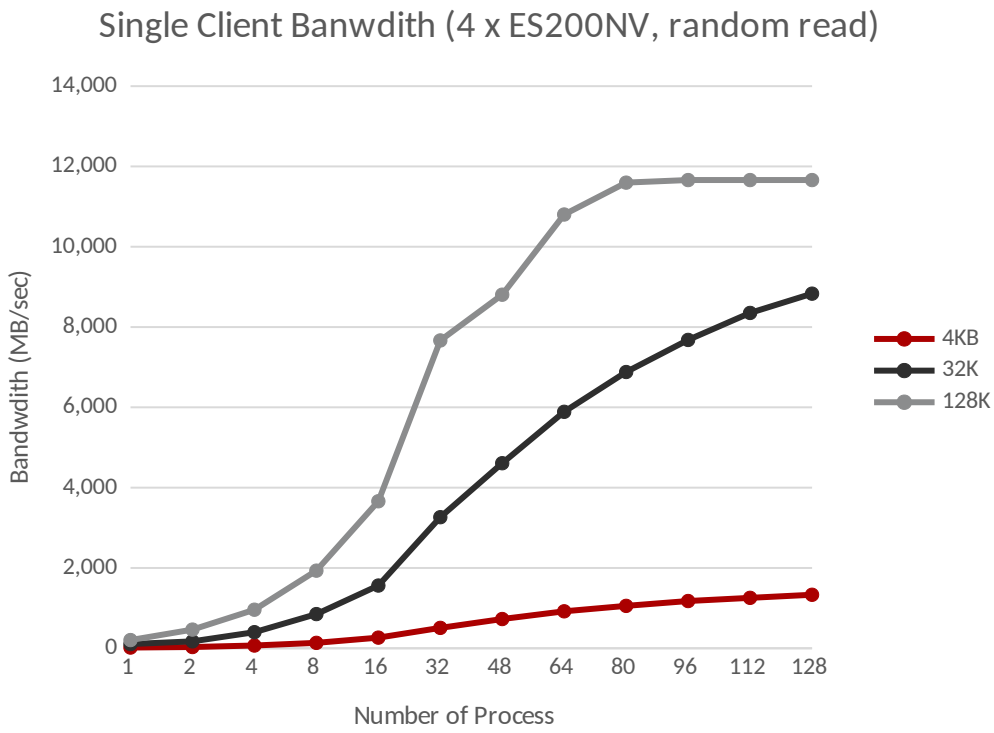
- Lustre 2.10.5-ddn7
- Capacity: 229.26 TiB
- Number of inodes: ~1.5 billion



Method

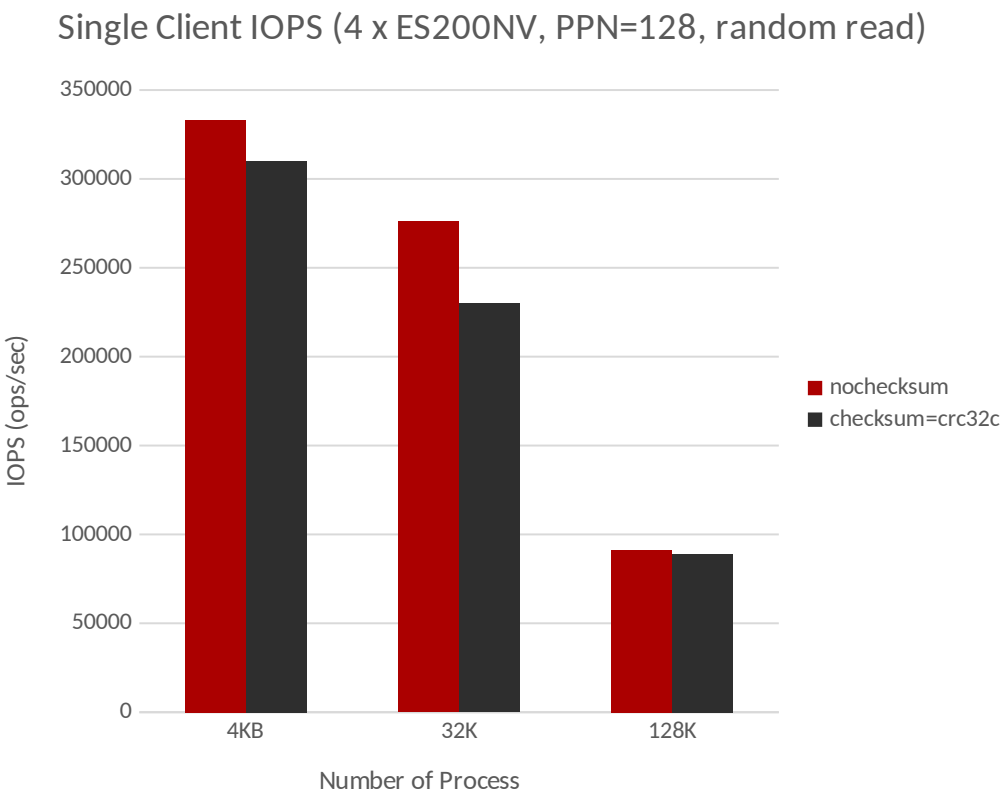
- Use fio with server/client mode
- Create test files(4GB file x processes) first, then random read (4KB, 32KB, and 128KB)
- Minimize I/O cache effects (“sync” and O_DIRECT mode)
- 10 sec run time

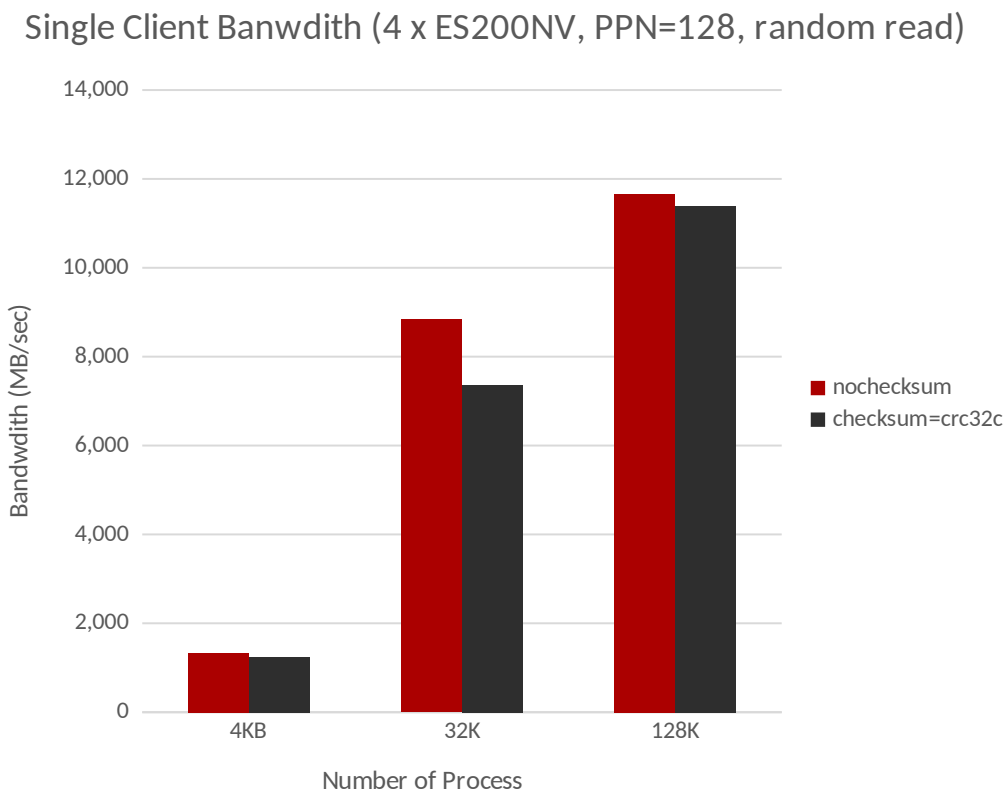
Single Client IOPS and Bandwidth(1) (1 x client and 4 x ES200NV)



- Achieved 333K IOPS at 4K Random Read and 11.5GB/sec at 128KB Random Read
- 128K Random Read is limited by EDR Bandwidth at Client

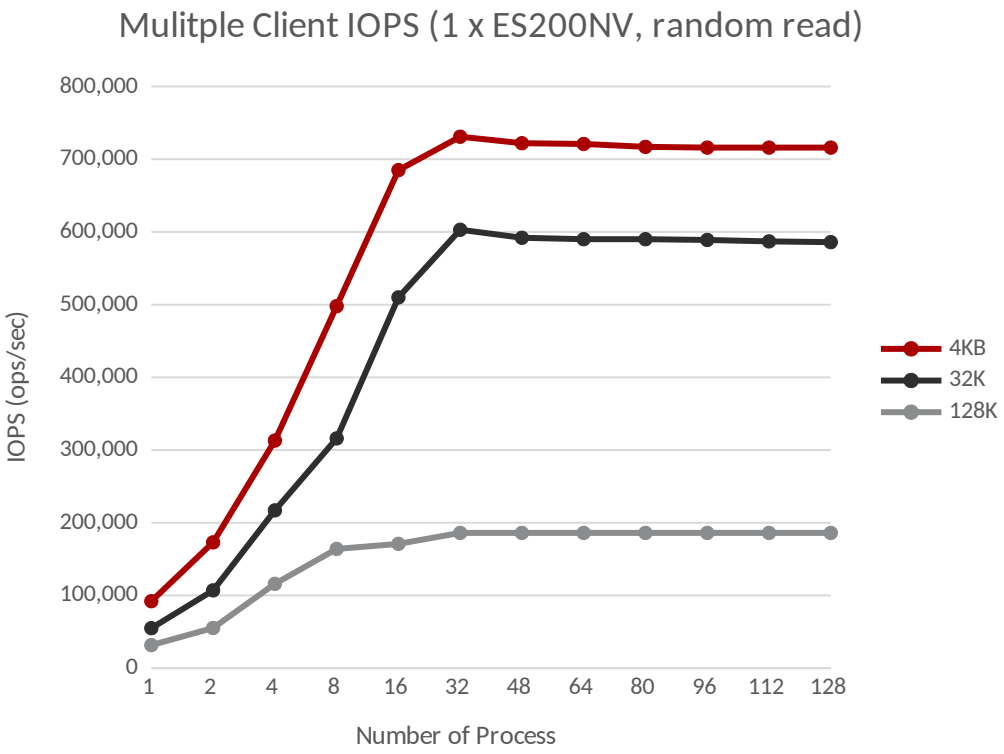
Single Client IOPS and Bandwidth(2)
Performance impacts with/without Lustre checksum

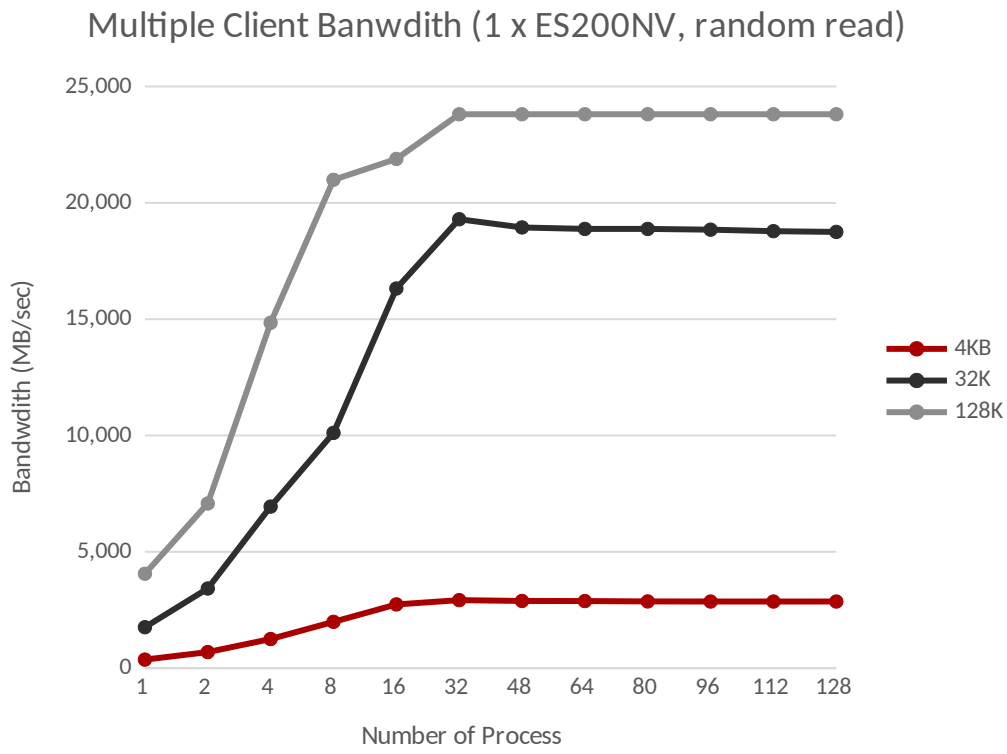




- ~6% checksum overhead at 4KB Random Read and almost no impacts at 128KB Random Read.
- ~17% Performance impact at 32KB Random Read. Will Investigate why.

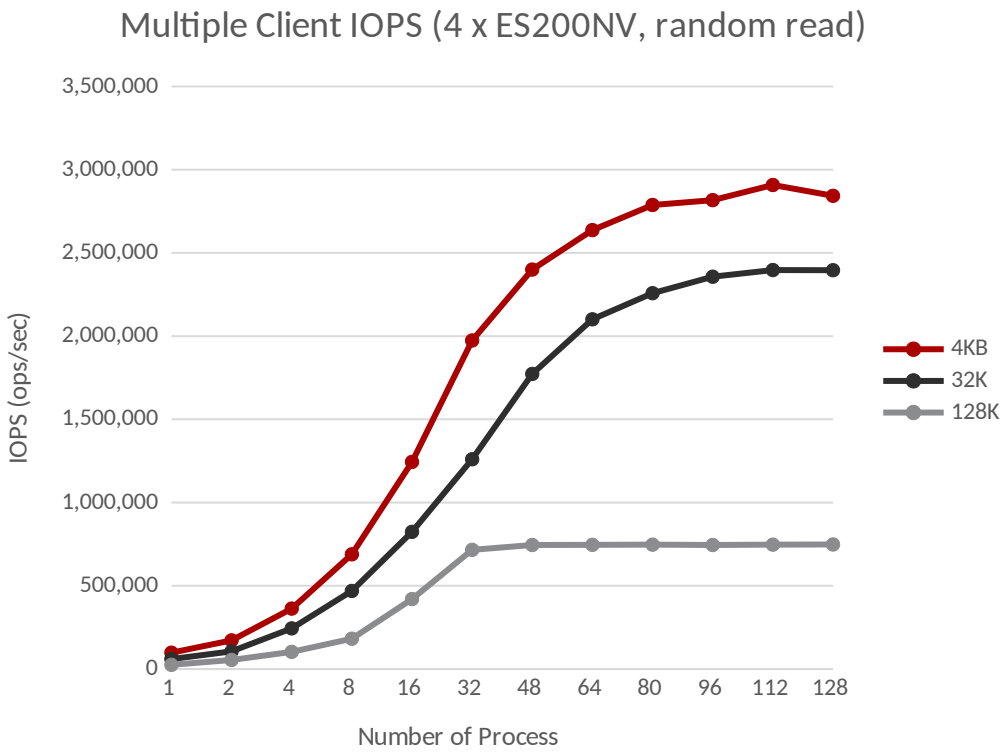
Multiple Client IOPS and Bandwidth(1) (22 x client and 1 x ES200NV)

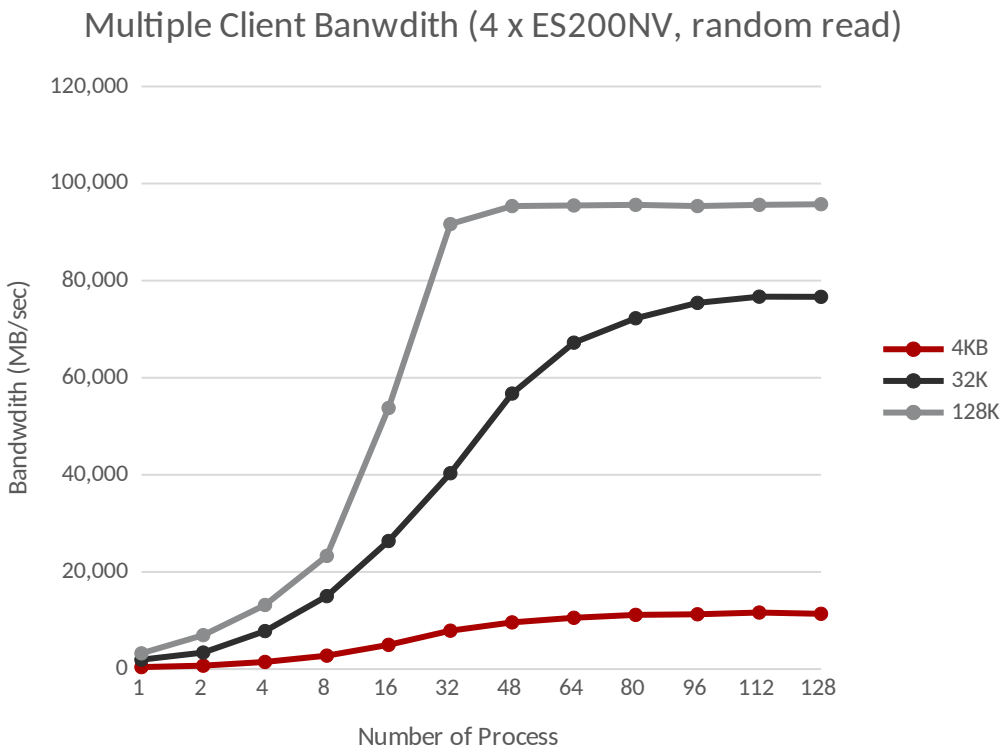




- Single ES200NV achieved ~710K IOPS at 4K Random Read and ~24GB/sec at 128K Random Read
- 128K Random Read is limited by EDR Bandwidth at OSSs

Multiple Client IOPS and Bandwidth(1) (22 x client and 4 x ES200NV)





- Achieved 2.84M IOPS at 4K Random Read and 95.7GB/sec at 128KB Random Read
- Delivered balanced IOPS(2.39M IOPS) and Bandwidth(76.9GB/sec) for 32KB Random Read

Summary

Lustre on All-Flash storage system handles huge number of small Random Read requests and unleashes the data analysis and AI/DL computing capability of Saion

- **Single client (1 x client and 4 x ES200NV)**
  - 333K IOPS at 4K Random Read
  - 11.8GB/sec at 128KB Random Read
- **Multiple client (22 x client and 4 x ES200NV)**
  - 2.84M IOPS at 4K Random Read
  - 95.7GB/sec at 128KB Random Read
- **Lusture Checksum impacts**
  - Small enough at 4K(~6%) and 128K(~0%), except 32KB

1. OIST - Okinawa Institute of Science and Technology

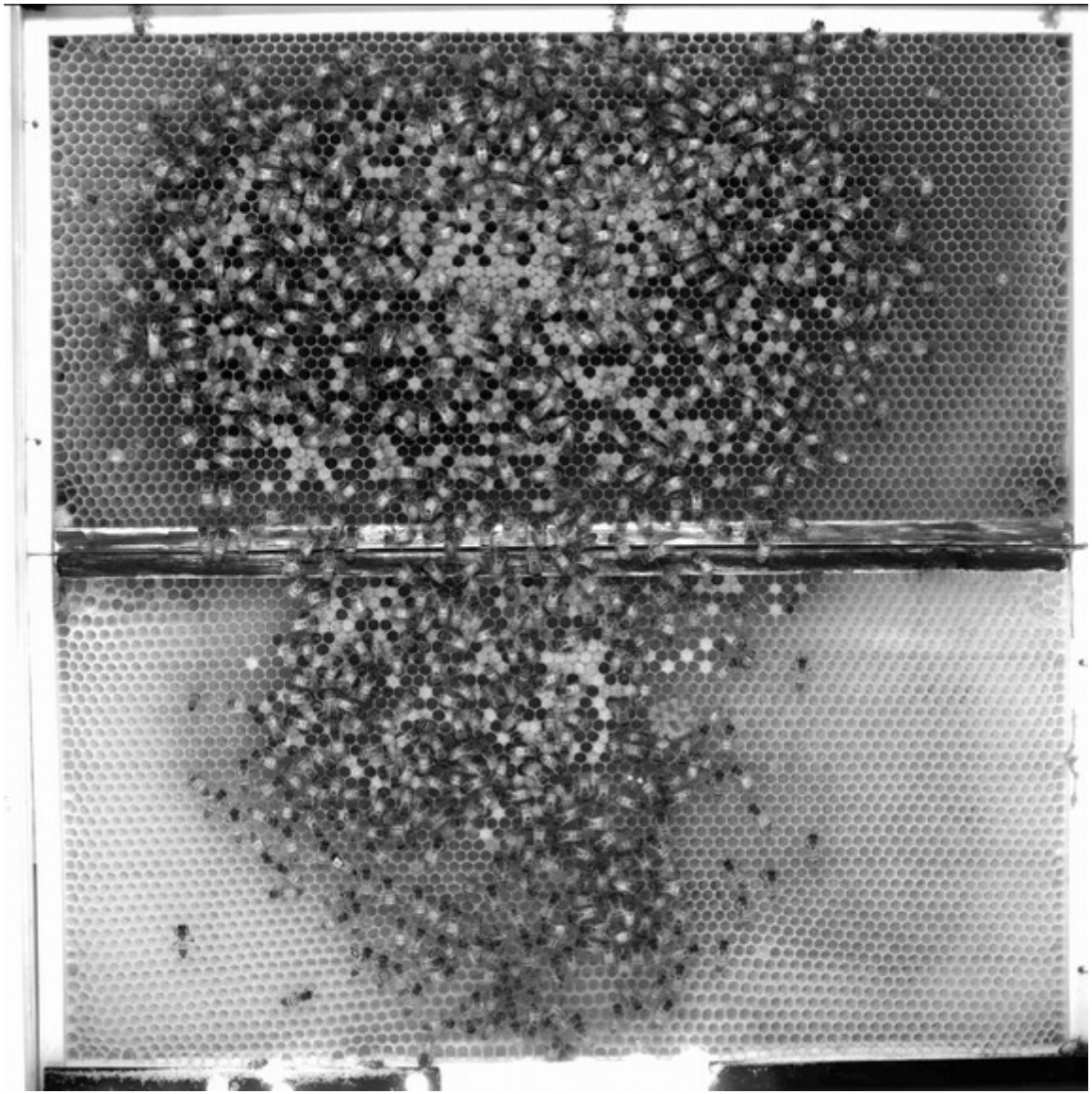
2. SCDA - Scientific Computing and Data Analysis Section

3. Benchmark Configuration and Results

4. Research Computing Example

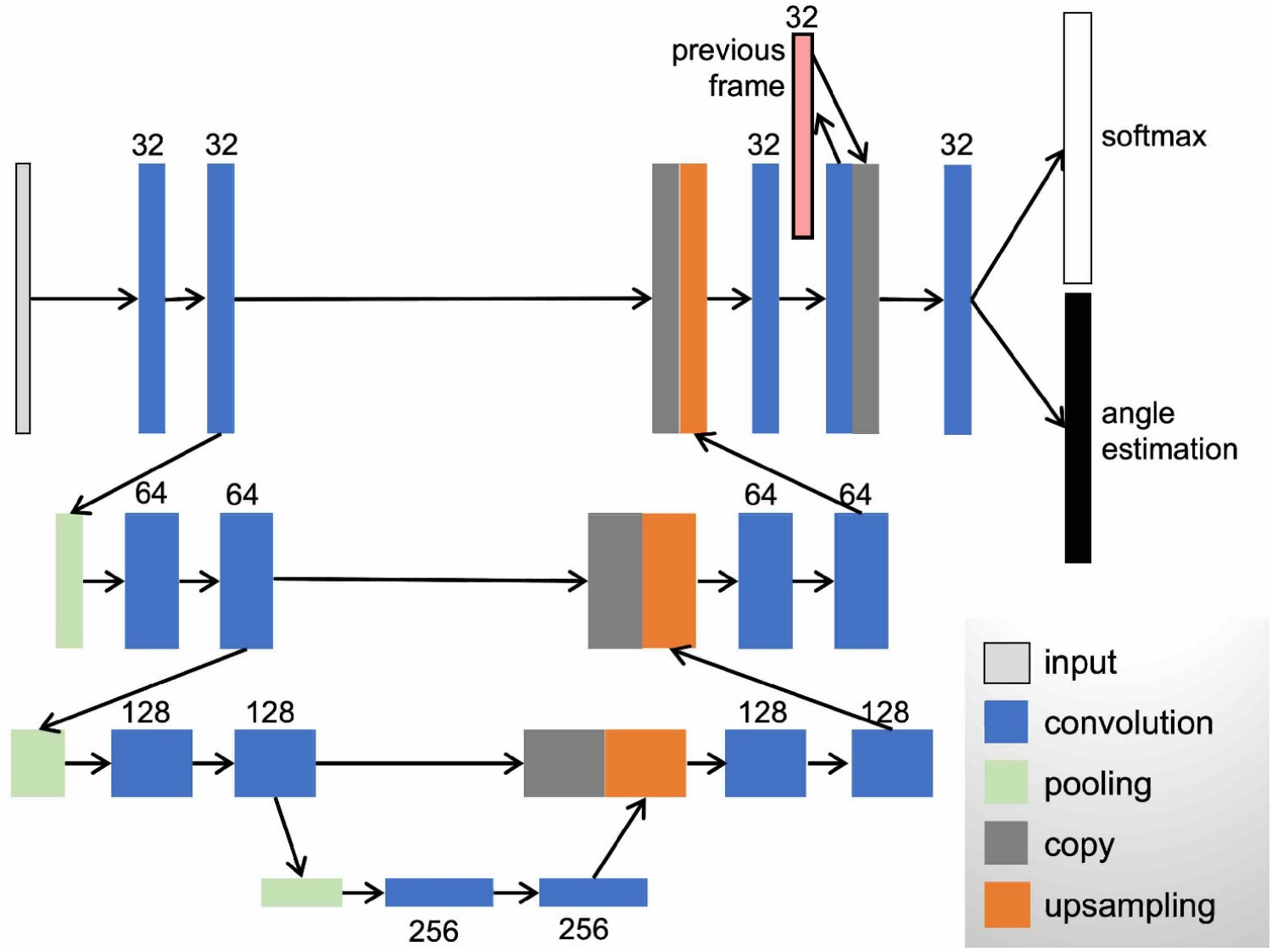
Detection and tracking of individual bees in a hive

"Towards dense object tracking in a 2D honeybee hive" Katarzyna Bozek, *et al*. In arXiv:1812.11797v1



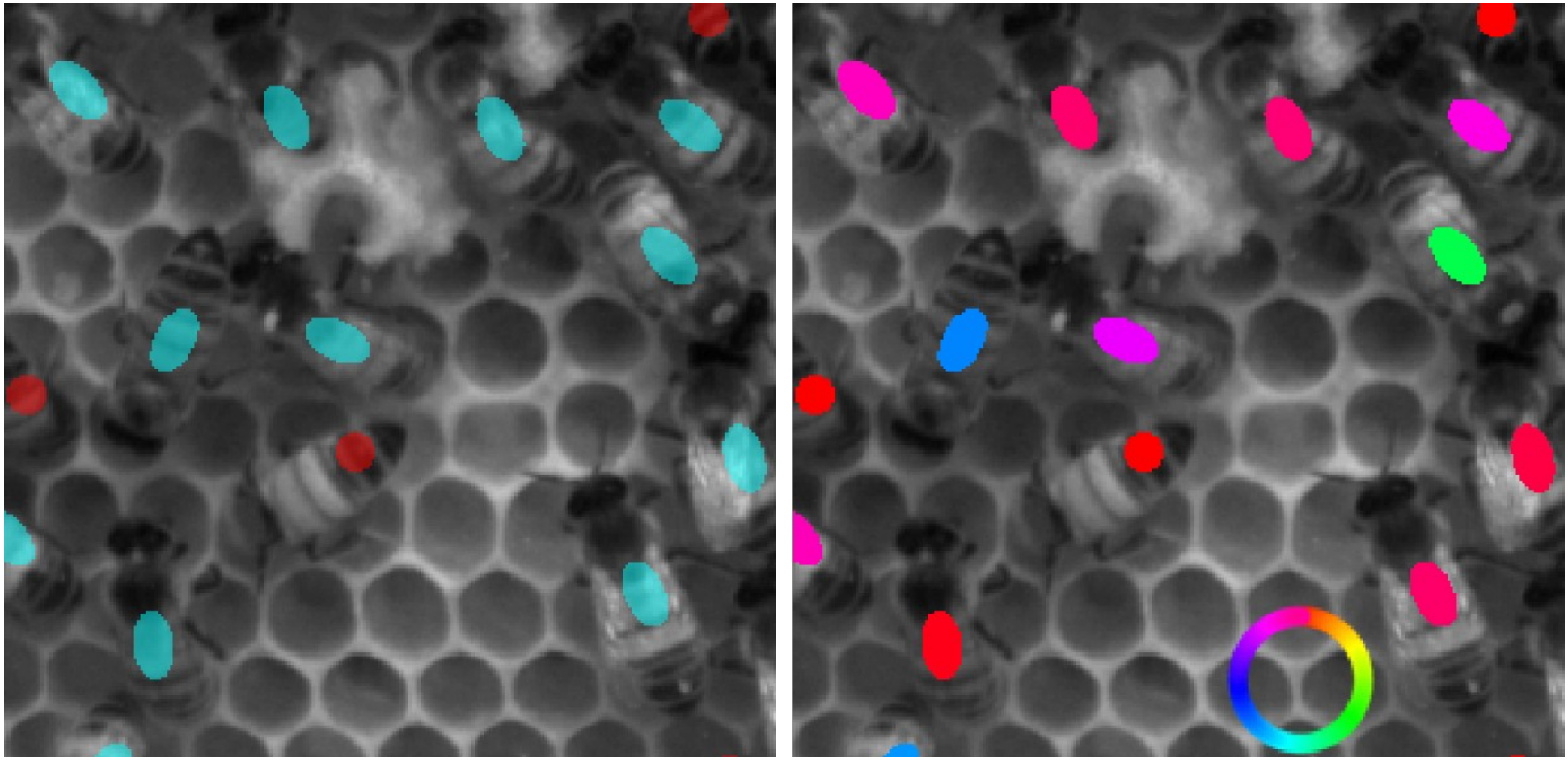
https://groups.oist.jp/bptu/honeybee-tracking-dataset

UNNet architecture for object detection



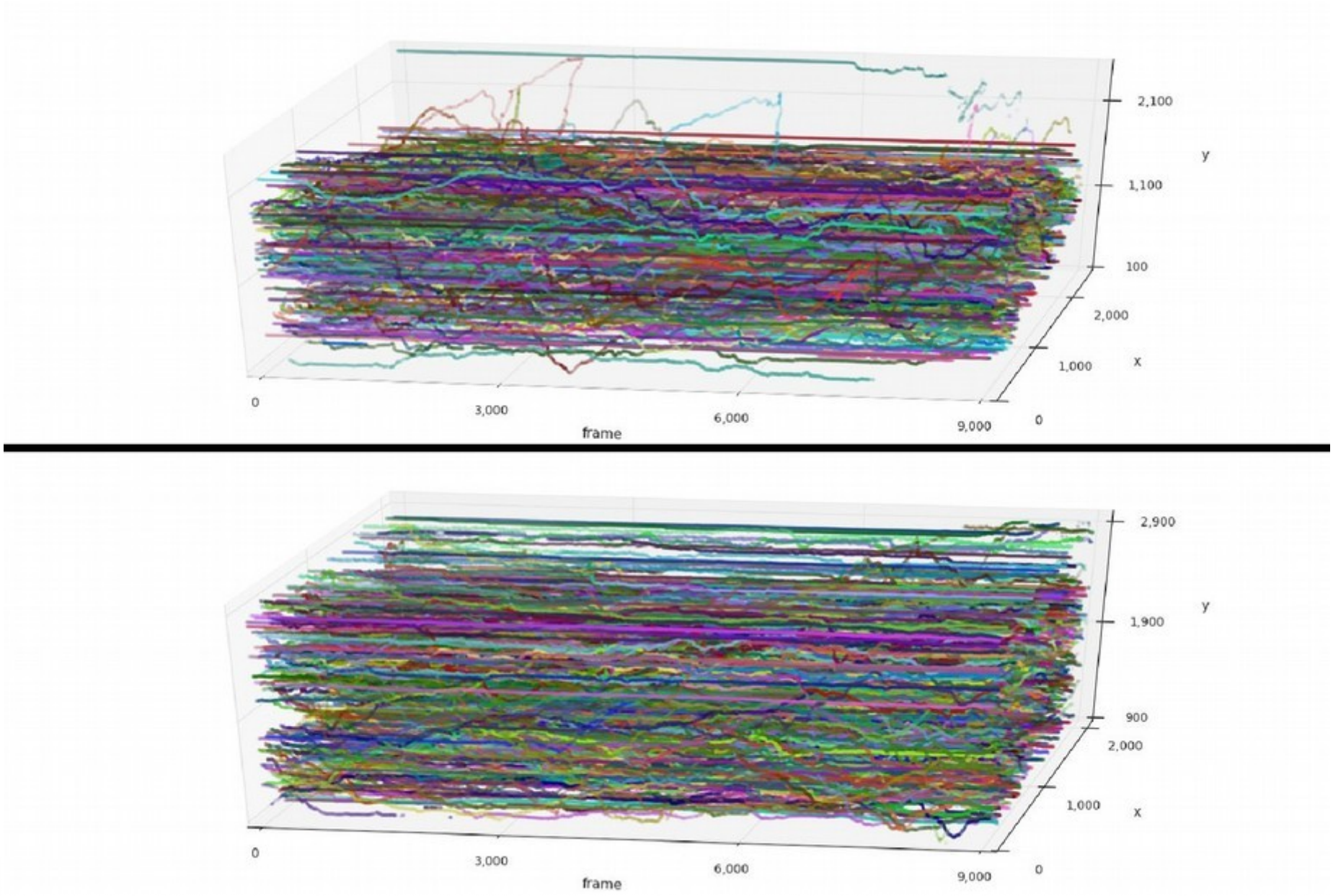
https://groups.oist.jp/bptu/honeybee-tracking-dataset

Example of labels and images used for detection



https://groups.oist.jp/bptu/honeybee-tracking-dataset

Reconstructed trajectories for several minutes of recording



https://groups.oist.jp/bptu/honeybee-tracking-dataset