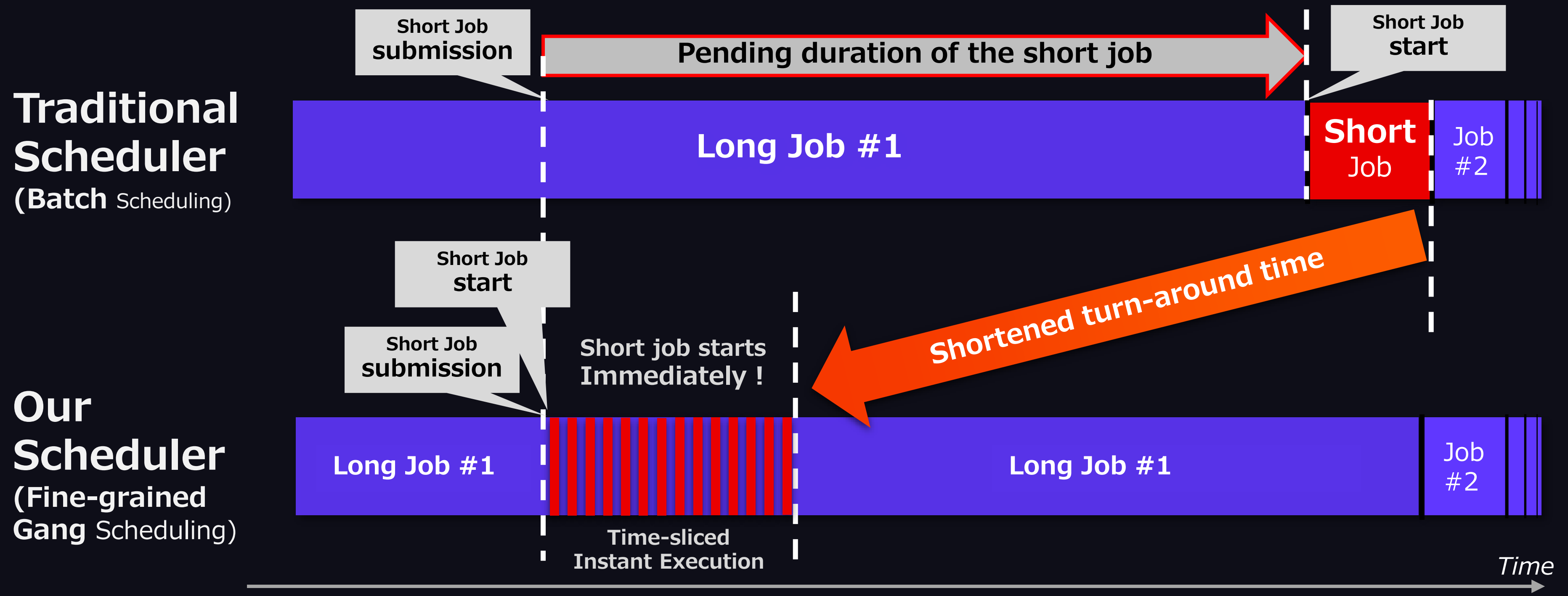
Interactive HPC

Fujitsu Limited

Bringing Interactivity on HPC

FUJITSU



Traditional Scheduler
(**Batch** Scheduling)

Short Job submission

Pending duration of the short job

Short Job start

Long Job #1

Short Job

Job #2

Shortened turn-around time

Our Scheduler
(**Fine-grained Gang** Scheduling)

Short Job start

Short Job submission

Short job starts Immediately !

Long Job #1

Long Job #1

Job #2

Time-sliced Instant Execution

Time

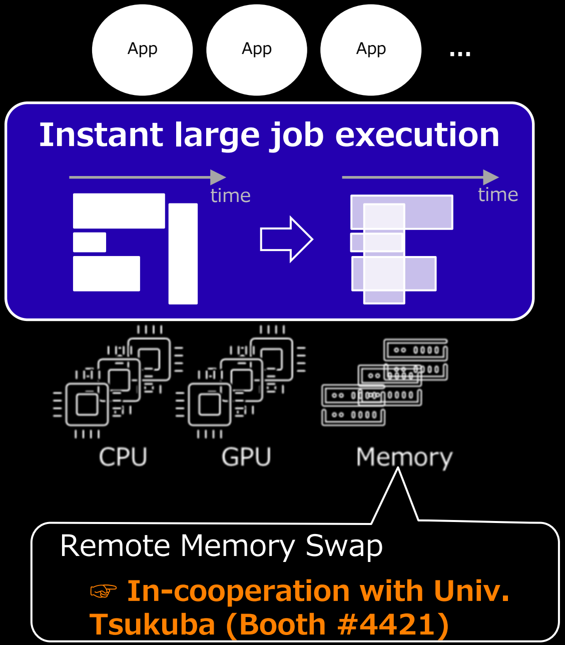
Interactive HPC

A scheduler for interactive parallel applications

- Real-time scalable processing
- Debugging large-scale programs
- Jupyter Notebook (ipython cluster)

Key Features

- Scalable sub-second global job switching
- CPU/GPU workloads are supported
- No user code modification required

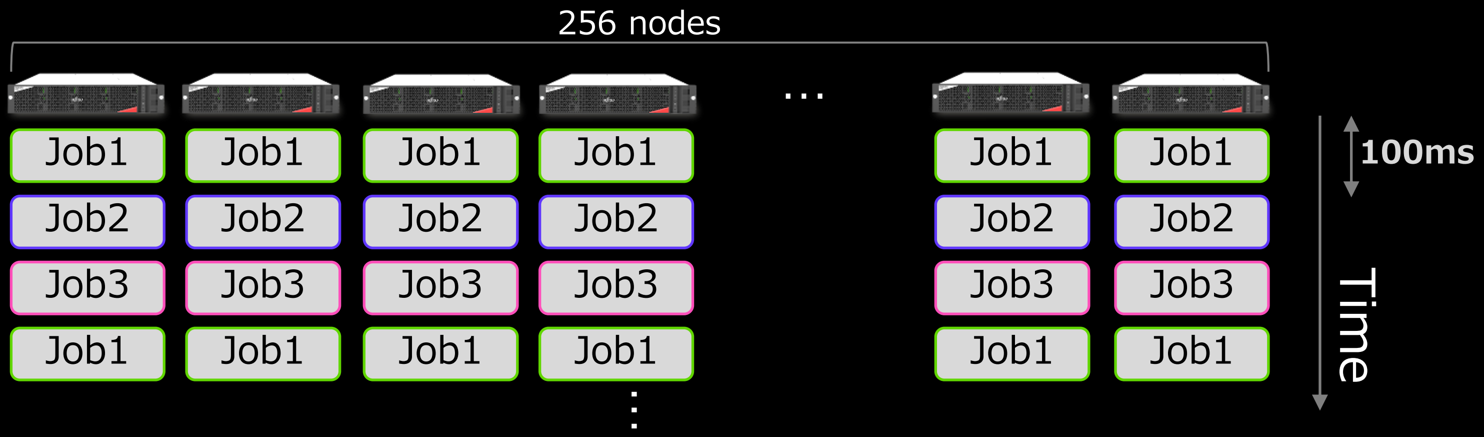


Remote Memory Swap

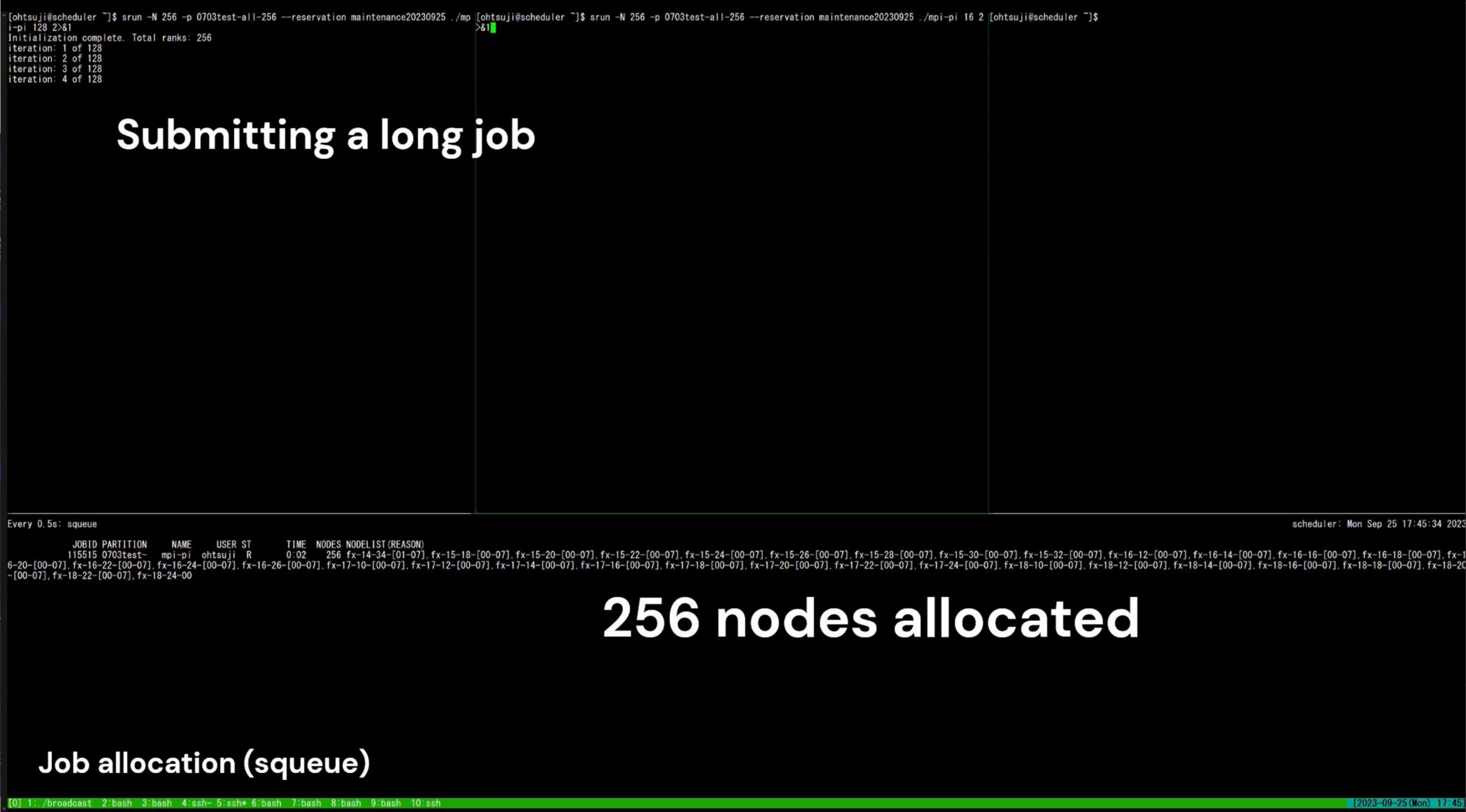
☞ **In-cooperation with Univ. Tsukuba (Booth #4421)**

Demo

- **Large-scale Parallel Application demo**
  - **Running multiple 256-node MPI-based parallel programs**
  - **100ms cluster-wide gang scheduling**



Job1	Job1	Job1	Job1		Job1	Job1
Job2	Job2	Job2	Job2		Job2	Job2
Job3	Job3	Job3	Job3		Job3	Job3
Job1	Job1	Job1	Job1		Job1	Job1



```
[ohtsuji@scheduler ~]$ srun -N 256 -p 0703test-all-256 --reservation maintenance20230925 ./mp
i-pi 128 2>&1
Initialization complete. Total ranks: 256
iteration: 1 of 128
iteration: 2 of 128
iteration: 3 of 128
iteration: 4 of 128
```

```
[ohtsuji@scheduler ~]$ srun -N 256 -p 0703test-all-256 --reservation maintenance20230925 ./mpi-pi 16 2
>&1
```

```
[ohtsuji@scheduler ~]$
```

Submitting a long job

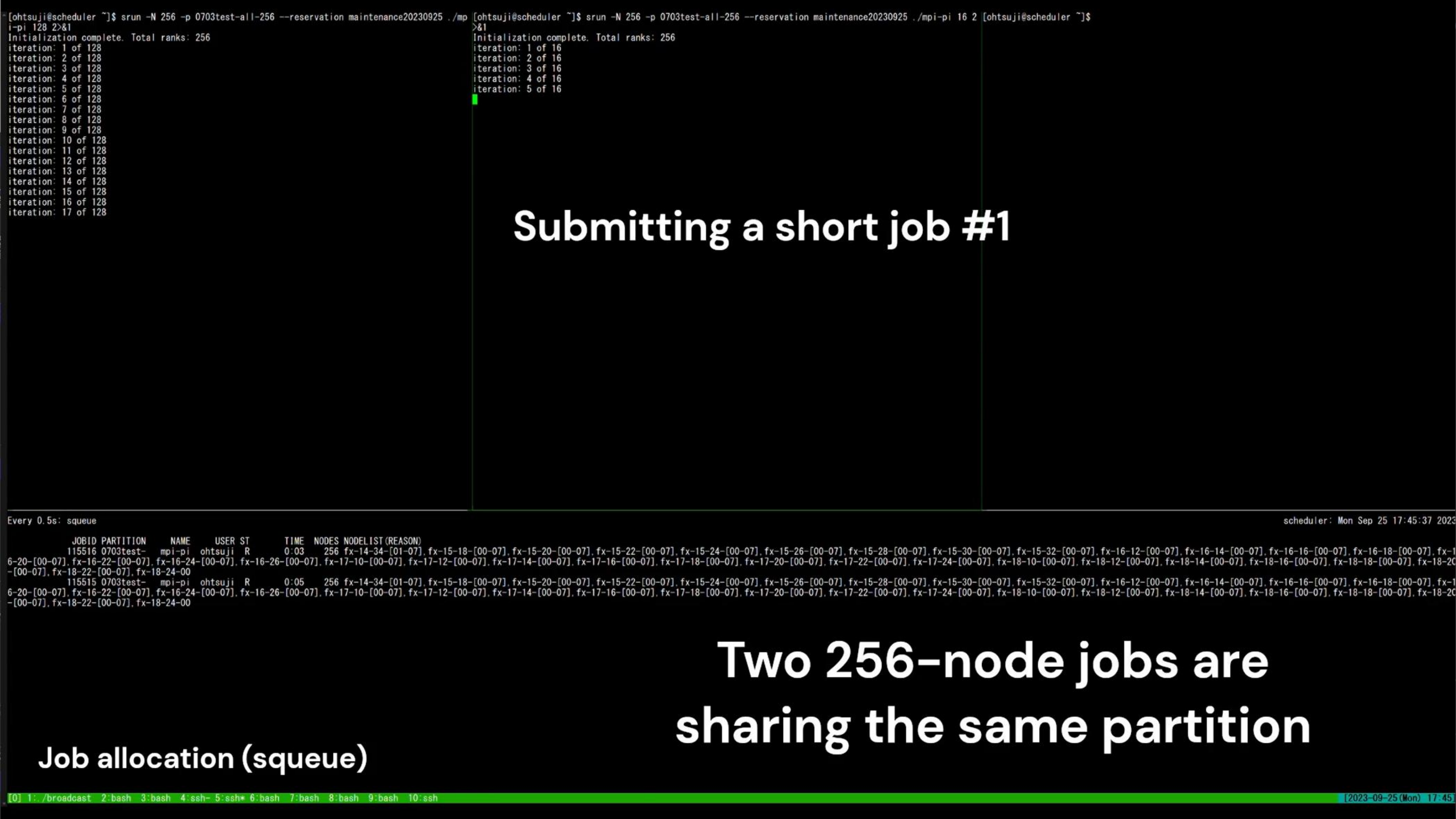
```
Every 0.5s: squeue                                                                                                                                                                                                          scheduler: Mon Sep 25 17:45:34 2023

             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
            115515 0703test-   mpi-pi  ohtsuji  R       0:02    256 fx-14-34-[01-07],fx-15-18-[00-07],fx-15-20-[00-07],fx-15-22-[00-07],fx-15-24-[00-07],fx-15-26-[00-07],fx-15-28-[00-07],fx-15-30-[00-07],fx-15-32-[00-07],fx-16-12-[00-07],fx-16-14-[00-07],fx-16-16-[00-07],fx-16-18-[00-07],fx-1
6-20-[00-07],fx-16-22-[00-07],fx-16-24-[00-07],fx-16-26-[00-07],fx-17-10-[00-07],fx-17-12-[00-07],fx-17-14-[00-07],fx-17-16-[00-07],fx-17-18-[00-07],fx-17-20-[00-07],fx-17-22-[00-07],fx-17-24-[00-07],fx-18-10-[00-07],fx-18-12-[00-07],fx-18-14-[00-07],fx-18-16-[00-07],fx-18-18-[00-07],fx-18-20
-[00-07],fx-18-22-[00-07],fx-18-24-00
```

256 nodes allocated

Job allocation (squeue)

```
[0] 1:./broadcast  2:bash  3:bash  4:ssh- 5:ssh* 6:bash  7:bash  8:bash  9:bash  10:ssh                                                                                                                                                    [2023-09-25(Mon) 17:45
```



```
[ohtsuji@scheduler ~]$ srun -N 256 -p 0703test-all-256 --reservation maintenance20230925 ./mp
i-pi 128 2>&1
Initialization complete. Total ranks: 256
iteration: 1 of 128
iteration: 2 of 128
iteration: 3 of 128
iteration: 4 of 128
iteration: 5 of 128
iteration: 6 of 128
iteration: 7 of 128
iteration: 8 of 128
iteration: 9 of 128
iteration: 10 of 128
iteration: 11 of 128
iteration: 12 of 128
iteration: 13 of 128
iteration: 14 of 128
iteration: 15 of 128
iteration: 16 of 128
iteration: 17 of 128
```

```
[ohtsuji@scheduler ~]$ srun -N 256 -p 0703test-all-256 --reservation maintenance20230925 ./mpi-pi 16 2
>&1
Initialization complete. Total ranks: 256
iteration: 1 of 16
iteration: 2 of 16
iteration: 3 of 16
iteration: 4 of 16
iteration: 5 of 16
```

```
[ohtsuji@scheduler ~]$
```

Submitting a short job #1

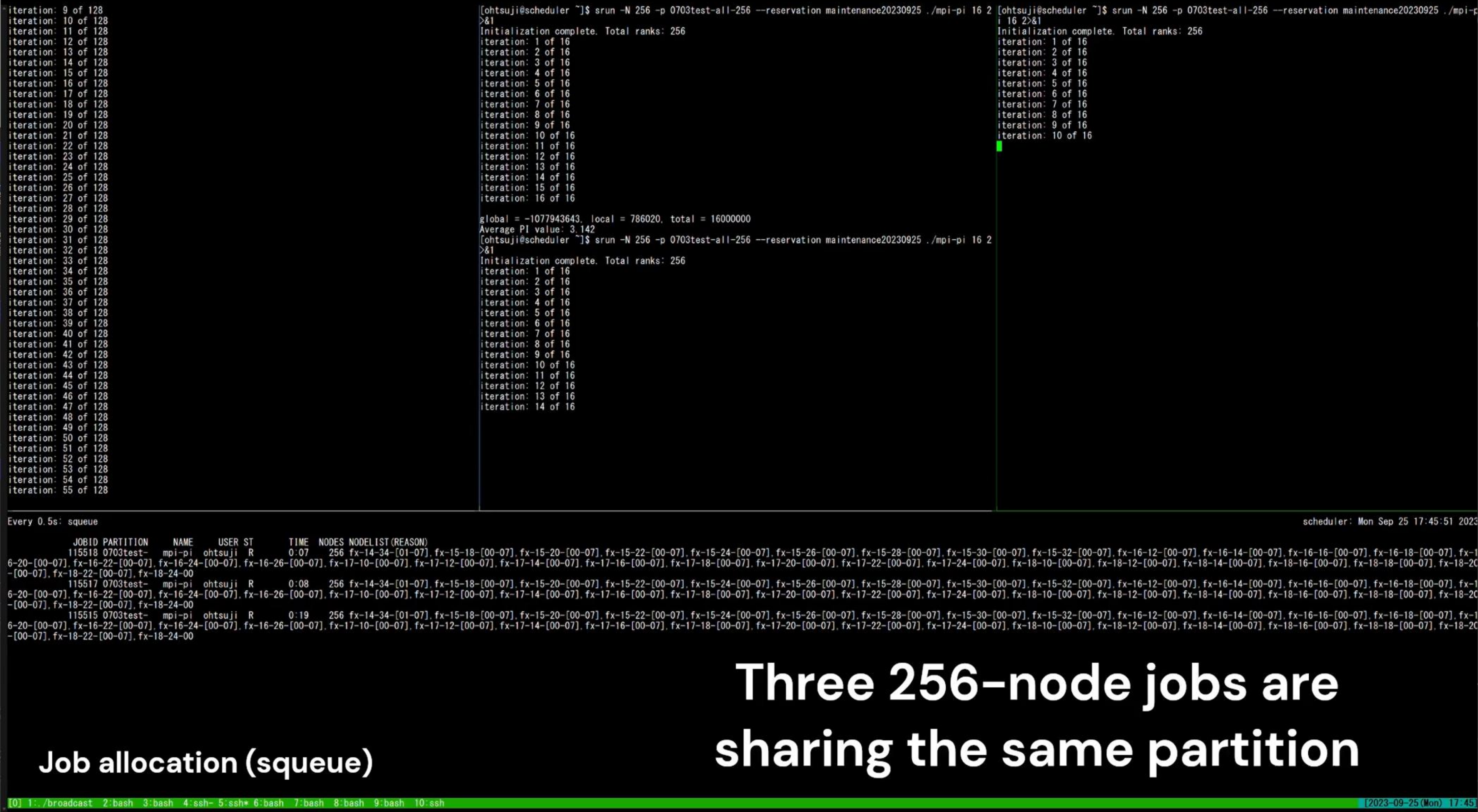
```
Every 0.5s: squeue                                                                                                                                                                                                                                                                                         scheduler: Mon Sep 25 17:45:37 2023

             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
            115516 0703test-   mpi-pi  ohtsuji  R       0:03    256 fx-14-34-[01-07],fx-15-18-[00-07],fx-15-20-[00-07],fx-15-22-[00-07],fx-15-24-[00-07],fx-15-26-[00-07],fx-15-28-[00-07],fx-15-30-[00-07],fx-15-32-[00-07],fx-16-12-[00-07],fx-16-14-[00-07],fx-16-16-[00-07],fx-16-18-[00-07],fx-1
6-20-[00-07],fx-16-22-[00-07],fx-16-24-[00-07],fx-16-26-[00-07],fx-17-10-[00-07],fx-17-12-[00-07],fx-17-14-[00-07],fx-17-16-[00-07],fx-17-18-[00-07],fx-17-20-[00-07],fx-17-22-[00-07],fx-17-24-[00-07],fx-18-10-[00-07],fx-18-12-[00-07],fx-18-14-[00-07],fx-18-16-[00-07],fx-18-18-[00-07],fx-18-20
-[00-07],fx-18-22-[00-07],fx-18-24-00
            115515 0703test-   mpi-pi  ohtsuji  R       0:05    256 fx-14-34-[01-07],fx-15-18-[00-07],fx-15-20-[00-07],fx-15-22-[00-07],fx-15-24-[00-07],fx-15-26-[00-07],fx-15-28-[00-07],fx-15-30-[00-07],fx-15-32-[00-07],fx-16-12-[00-07],fx-16-14-[00-07],fx-16-16-[00-07],fx-16-18-[00-07],fx-1
6-20-[00-07],fx-16-22-[00-07],fx-16-24-[00-07],fx-16-26-[00-07],fx-17-10-[00-07],fx-17-12-[00-07],fx-17-14-[00-07],fx-17-16-[00-07],fx-17-18-[00-07],fx-17-20-[00-07],fx-17-22-[00-07],fx-17-24-[00-07],fx-18-10-[00-07],fx-18-12-[00-07],fx-18-14-[00-07],fx-18-16-[00-07],fx-18-18-[00-07],fx-18-20
-[00-07],fx-18-22-[00-07],fx-18-24-00
```

Two 256-node jobs are sharing the same partition

Job allocation (squeue)

[0] 1: /broadcast 2:bash 3:bash 4:ssh- 5:ssh* 6:bash 7:bash 8:bash 9:bash 10:ssh
[2023-09-25(Mon) 17:45



```
iteration: 9 of 128
iteration: 10 of 128
iteration: 11 of 128
iteration: 12 of 128
iteration: 13 of 128
iteration: 14 of 128
iteration: 15 of 128
iteration: 16 of 128
iteration: 17 of 128
iteration: 18 of 128
iteration: 19 of 128
iteration: 20 of 128
iteration: 21 of 128
iteration: 22 of 128
iteration: 23 of 128
iteration: 24 of 128
iteration: 25 of 128
iteration: 26 of 128
iteration: 27 of 128
iteration: 28 of 128
iteration: 29 of 128
iteration: 30 of 128
iteration: 31 of 128
iteration: 32 of 128
iteration: 33 of 128
iteration: 34 of 128
iteration: 35 of 128
iteration: 36 of 128
iteration: 37 of 128
iteration: 38 of 128
iteration: 39 of 128
iteration: 40 of 128
iteration: 41 of 128
iteration: 42 of 128
iteration: 43 of 128
iteration: 44 of 128
iteration: 45 of 128
iteration: 46 of 128
iteration: 47 of 128
iteration: 48 of 128
iteration: 49 of 128
iteration: 50 of 128
iteration: 51 of 128
iteration: 52 of 128
iteration: 53 of 128
iteration: 54 of 128
iteration: 55 of 128
```

```
[ohtsuji@scheduler ~]$ srun -N 256 -p 0703test-all-256 --reservation maintenance20230925 ./mpi-pi 16 2
>&1
Initialization complete. Total ranks: 256
iteration: 1 of 16
iteration: 2 of 16
iteration: 3 of 16
iteration: 4 of 16
iteration: 5 of 16
iteration: 6 of 16
iteration: 7 of 16
iteration: 8 of 16
iteration: 9 of 16
iteration: 10 of 16
iteration: 11 of 16
iteration: 12 of 16
iteration: 13 of 16
iteration: 14 of 16
iteration: 15 of 16
iteration: 16 of 16

global = -1077943643, local = 786020, total = 16000000
Average PI value: 3.142
[ohtsuji@scheduler ~]$ srun -N 256 -p 0703test-all-256 --reservation maintenance20230925 ./mpi-pi 16 2
>&1
Initialization complete. Total ranks: 256
iteration: 1 of 16
iteration: 2 of 16
iteration: 3 of 16
iteration: 4 of 16
iteration: 5 of 16
iteration: 6 of 16
iteration: 7 of 16
iteration: 8 of 16
iteration: 9 of 16
iteration: 10 of 16
iteration: 11 of 16
iteration: 12 of 16
iteration: 13 of 16
iteration: 14 of 16
```

```
[ohtsuji@scheduler ~]$ srun -N 256 -p 0703test-all-256 --reservation maintenance20230925 ./mpi-p
i 16 2>&1
Initialization complete. Total ranks: 256
iteration: 1 of 16
iteration: 2 of 16
iteration: 3 of 16
iteration: 4 of 16
iteration: 5 of 16
iteration: 6 of 16
iteration: 7 of 16
iteration: 8 of 16
iteration: 9 of 16
iteration: 10 of 16
```

```
Every 0.5s: squeue                                                                                                                                                                                   scheduler: Mon Sep 25 17:45:51 2023

             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
            115518 0703test-   mpi-pi  ohtsuji  R       0:07    256 fx-14-34-[01-07],fx-15-18-[00-07],fx-15-20-[00-07],fx-15-22-[00-07],fx-15-24-[00-07],fx-15-26-[00-07],fx-15-28-[00-07],fx-15-30-[00-07],fx-15-32-[00-07],fx-16-12-[00-07],fx-16-14-[00-07],fx-16-16-[00-07],fx-16-18-[00-07],fx-1
6-20-[00-07],fx-16-22-[00-07],fx-16-24-[00-07],fx-16-26-[00-07],fx-17-10-[00-07],fx-17-12-[00-07],fx-17-14-[00-07],fx-17-16-[00-07],fx-17-18-[00-07],fx-17-20-[00-07],fx-17-22-[00-07],fx-17-24-[00-07],fx-18-10-[00-07],fx-18-12-[00-07],fx-18-14-[00-07],fx-18-16-[00-07],fx-18-18-[00-07],fx-18-20
-[00-07],fx-18-22-[00-07],fx-18-24-00
            115517 0703test-   mpi-pi  ohtsuji  R       0:08    256 fx-14-34-[01-07],fx-15-18-[00-07],fx-15-20-[00-07],fx-15-22-[00-07],fx-15-24-[00-07],fx-15-26-[00-07],fx-15-28-[00-07],fx-15-30-[00-07],fx-15-32-[00-07],fx-16-12-[00-07],fx-16-14-[00-07],fx-16-16-[00-07],fx-16-18-[00-07],fx-1
6-20-[00-07],fx-16-22-[00-07],fx-16-24-[00-07],fx-16-26-[00-07],fx-17-10-[00-07],fx-17-12-[00-07],fx-17-14-[00-07],fx-17-16-[00-07],fx-17-18-[00-07],fx-17-20-[00-07],fx-17-22-[00-07],fx-17-24-[00-07],fx-18-10-[00-07],fx-18-12-[00-07],fx-18-14-[00-07],fx-18-16-[00-07],fx-18-18-[00-07],fx-18-20
-[00-07],fx-18-22-[00-07],fx-18-24-00
            115515 0703test-   mpi-pi  ohtsuji  R       0:19    256 fx-14-34-[01-07],fx-15-18-[00-07],fx-15-20-[00-07],fx-15-22-[00-07],fx-15-24-[00-07],fx-15-26-[00-07],fx-15-28-[00-07],fx-15-30-[00-07],fx-15-32-[00-07],fx-16-12-[00-07],fx-16-14-[00-07],fx-16-16-[00-07],fx-16-18-[00-07],fx-1
6-20-[00-07],fx-16-22-[00-07],fx-16-24-[00-07],fx-16-26-[00-07],fx-17-10-[00-07],fx-17-12-[00-07],fx-17-14-[00-07],fx-17-16-[00-07],fx-17-18-[00-07],fx-17-20-[00-07],fx-17-22-[00-07],fx-17-24-[00-07],fx-18-10-[00-07],fx-18-12-[00-07],fx-18-14-[00-07],fx-18-16-[00-07],fx-18-18-[00-07],fx-18-20
-[00-07],fx-18-22-[00-07],fx-18-24-00
```

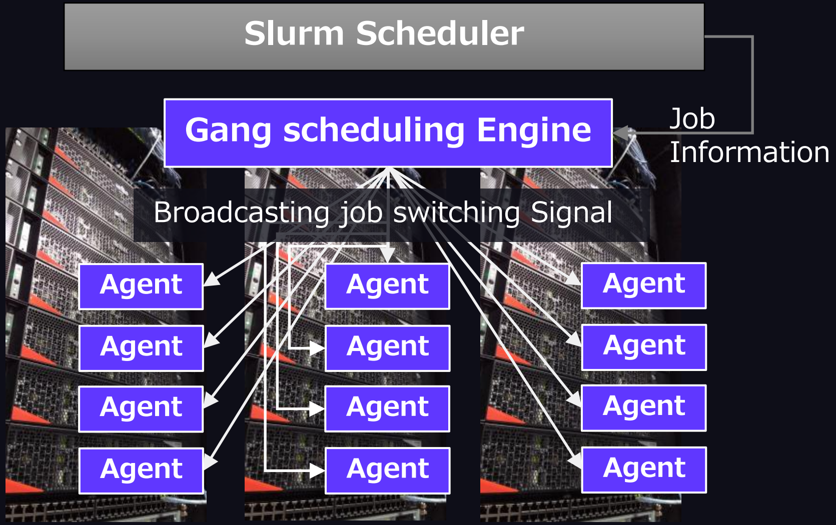
```
[0] 1:./broadcast  2:bash  3:bash  4:ssh-  5:ssh* 6:bash  7:bash  8:bash  9:bash  10:ssh                                                                                                             [2023-09-25(Mon) 17:45]
```

Job allocation (squeue)

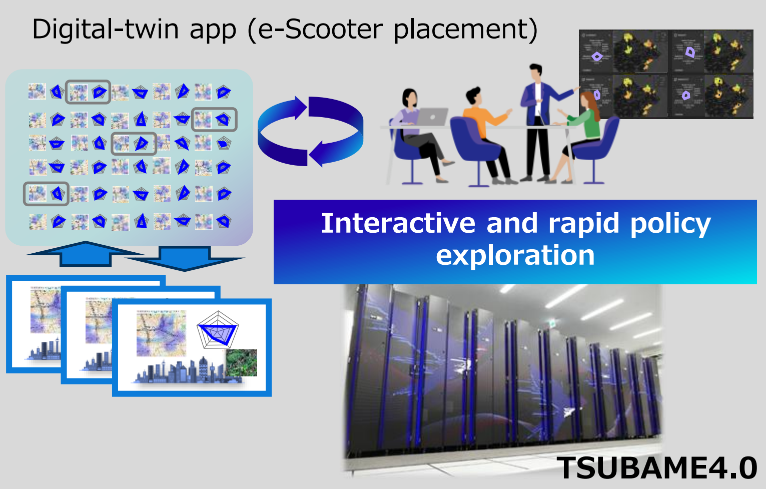
Three 256-node jobs are sharing the same partition

Deployment Case

- 1056-node ARM-based HPC Cluster (System-wide)



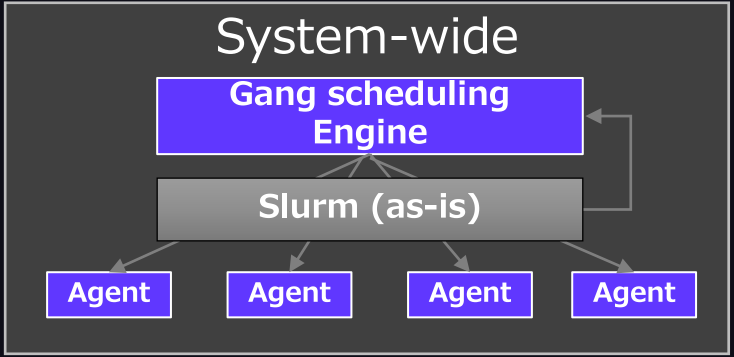
- TSUBAME 3.0/4.0 in Science Tokyo (User space)


Digital-twin app (e-Scooter placement)
Interactive and rapid policy exploration
TSUBAME4.0

☞ In-cooperation with Science Tokyo (Booth #4109)

Trial Kit is ready

- Add-on for Slurm and other schedulers
- User-mode trial kit is also available
  - No change to existing schedulers



System-wide

Gang scheduling Engine

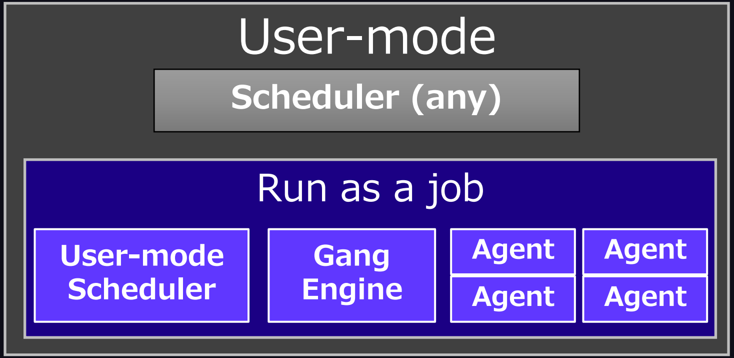
Slurm (as-is)

Agent

Agent

Agent

Agent



User-mode

Scheduler (any)

Run as a job

User-mode Scheduler

Gang Engine

Agent

Agent

Agent

Agent

Thank you

For more details:
https://www.fujitsu.com/global/products/computing/servers/supercomputer/topics/sc24/