Cybermedia Center

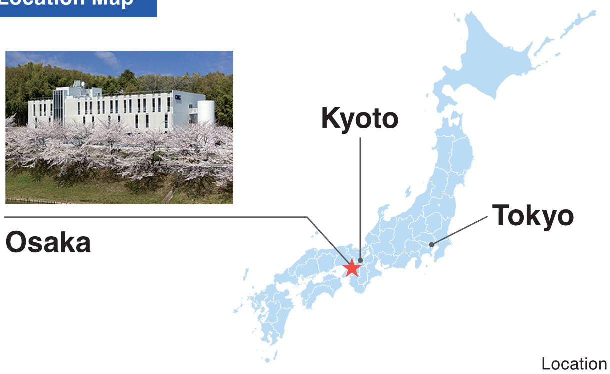
Osaka University, Japan

SC19 BOOTH 1843

About Us：Cybermedia Center, Osaka University

As a resource provider of knowledge and technology derived from advanced researches conducted in Osaka University, the Cybermedia Center (CMC) offers support in the areas of large-scale computation, information communication, multimedia content and education. The center also works closely with educational and research organizations within Osaka University, as well as with industries and institutes outside the University. By sharing its resources and encouraging local communities to use its facilities for public lectures and other events, CMC has helped to create a more internationally-oriented IT society for the region.

Location Map



Location

University-Wide Services

Large-Scale Computer System, we provide a high-performance computing environment, consisting of the NEC SX-ACE supercomputer and PC clusters, to both the academic and industrial communities. Part of the overall computer system is provided, as a computational resource, to the national High-Performance Computing Infrastructure(HPCI).

Information Media Education Multimedia Language Education, we have implemented a consistent curriculum, from the basics of computer utilization to advanced subject matter, while the Computer Assisted Language Learning System supports foreign language learning and cross-cultural understanding in accordance with each individual's language-proficiency level.

Cybermedia Commons Is an active learning space for students, exploiting a wide variety of the Cybermedia Center's information technology, to support student's active learning and research activities.

Cybermedia Commons

Digital Library provides academic information databases and remote access to electronic journals. It is equipped with multimedia terminals and public network jacks with an authentication system.

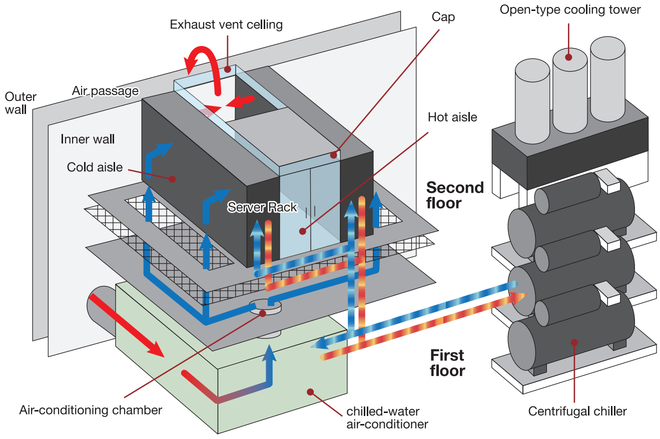
Repair and Maintenance of the Information Network, a high-speed, stable and reliable campus-wide network environment, as well as wireless access networks, as information infrastructure for supporting the educational, research, and social contribution activities of Osaka University.

Visualization Services, we maintain two types of high-resolution stereo visualization systems, as primary visualization facilities. The systems can be used for scientific visualization, information visualization, visual analytics, and other research activities.

Visualization Services

Academic Cloud improves the integration of computing resources scattered across the university. The objectives of the system are to optimize administrative operations, enhance security, and reduce costs.

IT Core Annex is a two-story steel-frame data center housing large-scale computers. The perimeter wall is designed with gently curving surface and light-permeable metal panels, to harmonize with the surrounding environment.



Cooling mechanism in IT Core Annex

Contact : sc19@ais.cmc.osaka-u.ac.jp https://www.cmc.osaka-u.ac.jp/

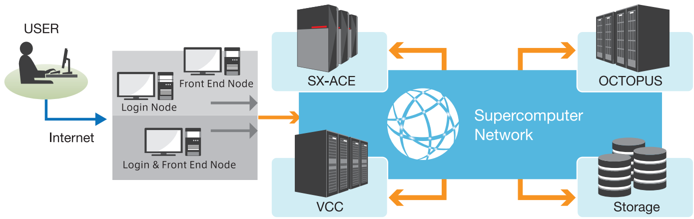
Cybermedia Center

Osaka University, Japan

SC19 BOOTH 1843

Large-scale Computing Systems at the Cybermedia Center

Overview of high-performance computing environment at the CMC



Large-scale computing systems (SX-ACE, VCC, and OCTOPUS) are deployed on CMC-Supercomputer network, a.k.a CMC-SCinet, a low-latency and wide-bandwidth network. This architectural design allows users to access to large-scale storage systems, perform large-scale high-performance computation and analysis on our large-scale computing systems.

Large-scale Computing System

OCTOPUS

OCTOPUS means Osaka university Cybermedia cenTer Over-Petascale Universal Supercomputer. OCTOPUS is a cluster system supposed to start its operation in December 2017. This system is composed of different types of 4 clusters, General purpose CPU nodes, Xeon Phi nodes, GPU nodes and Large-scale shared-memory nodes, total 319 nodes. These nodes and large-scale storage “EXAScaler” (Lustre 3.1 PB) are interconnected on InfiniBand EDR (100 Gbps) and form a cluster.

General purpose CPU nodes

CPU	Intel Xeon Skylake
OS	RHEL 7.3
# of nodes (total)	236 nodes
# of cores (total)	5,664 cores
# of memory (total)	45.3 TB
Peak performance	471.2 TFLOPS

Large-scale shared-memory nodes

CPU	Intel Xeon Skylake
OS	RHEL 7.3
# of nodes (total)	2 nodes
# of cores (total)	256 cores
# of memory (total)	12 TB
Peak performance	16.4 TFLOPS

GPU nodes

CPU	Intel Xeon Skylake
OS	RHEL 7.3
# of nodes (total)	37 nodes
# of cores (total)	888 cores
# of memory (total)	7.1 TB
Peak performance	858.3 TFLOPS
GPU	NVIDIA Tesla P100 x 148

Xeon Phi nodes

CPU	Intel Xeon Phi KNL
OS	RHEL 7.3
# of nodes (total)	44 nodes
# of cores (total)	2,816 cores
# of memory (total)	8.4 TB
Peak performance	117.1 TFLOPS

SX-ACE

CPU	NEC Vector Processor
OS	SUPER-UX
# of nodes (total)	1,536 nodes
# of cores (total)	6,144 cores
# of memory (total)	98 TB
Peak performance	423 TFLOPS

SX-ACE is a “clusterized” vector-typed supercomputer, composed of 3 clusters, each of which is composed of 512 nodes. Each node equips 4-core multi-core CPU and a 64 GB main memory. These 512 nodes are interconnected on a dedicated and specialized network switch, called IXS (Internode Crossbar Switch) and forms a cluster. Note that IXS interconnects 512 nodes with a single lane of 2-layer fat-tree structure and as a result exhibits 4 GB/s for each direction of input and output between nodes. SX-ACE will be retired on September 30, 2020. Next system will be introduced in 1Q/2021.

VCC

CPU	Intel Xeon Ivy Bridge & Broadwell
OS	Cent OS 6.8
# of nodes (total)	69 nodes
# of cores (total)	1,404 cores
# of memory (total)	4.4 TB
Peak performance	100.1 TFLOPS
GPU	NVIDIA Tesla K20 x 59

VCC is a cluster system composed of 69 nodes. These nodes are interconnected on InfiniBand FDR and form a cluster. Also, this system has introduced ExpEther, a system hardware virtualization technology. Each node can be connected with extension I/O nodes with which GPU resource, and SSD on 20 Gbps ExpEther network. A major characteristic is that this cluster system is reconfigured based on user’s usage and purpose by changing the combination of node and extension I/O node. VCC will be retired on March 31, 2020.

Application

GROMACS, LAMMPS, OpenFOAM, Relion, Quantum Espresso, VisIt , Gaussian09/16, IDL, AVS/Express (DEV/PCE/MPE), NEC Remote Debugger, NEC Ftrace viewer, Anaconda, Caffe, Theano, Chainer, TensorFlow, Digits, Torch, GAMESS, NICE Desktop Cloud Visualization, HΦ, MODYLAS, NTChem, OpenMX, SALMON, SMASH, FreeFem++, FLASH

Library (SX-ACE)

MathKeisan (BLAS, LAPACK, etc), ASL, ASLSTAT, ASLQUAD,MPI/SX, HPF/SX, XMP

Library (VCC, OCTOPUS)

Intel MKL (BLAS, LAPACK, etc), IntelMPI, OpenMPI, MVAPICH2, XMP, OpenACC, NetCDF, HDF5, GSL

Contact : sc19@ais.cmc.osaka-u.ac.jp http://www.hpc.cmc.osaka-u.ac.jp/en/

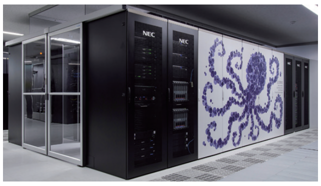
Cybermedia Center

Osaka University, Japan

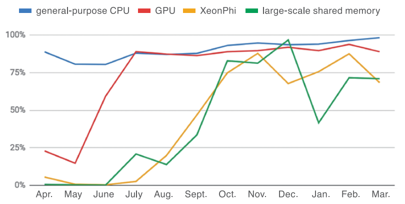
SC19 BOOTH 1843

Feasible Study of Cloud Bursting on OCTOPUS

Background



OCTOPUS is a hybrid cluster system of general-purpose CPU nodes (Skylake), many-core nodes (Knights Landing), GPU nodes (Tesla P100) and large-scale shared memory nodes with a 2PB Lustre storage (DDN EXAScaler). Since we started the operation of OCTOPUS, OCTOPUS has kept a higher utilization ratio. In particular, CPU and GPU nodes have a tendency of being demanded all year round. As a result, user waiting time is becoming longer.



Goal and Purpose

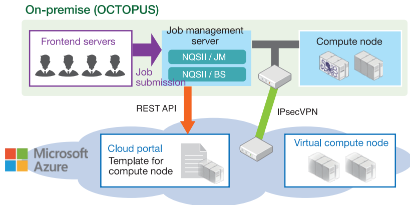
- Offloading the workload on OCTOPUS to the Cloud to alleviate the peak in hope that
  - User waiting time is reduced.
  - Higher throughput is realized.
  - We do not receive any complaint about waiting time (Higher satisfaction is achieved).
- Investigating the feasibility for the future integrated use of our supercomputing systems with the cloud.
  - For scaling out in need of compute resources.
  - For deploying and delivering the brand-new processors and accelerators to our user scientists and researchers.

OCTOPUS-Azure Environment

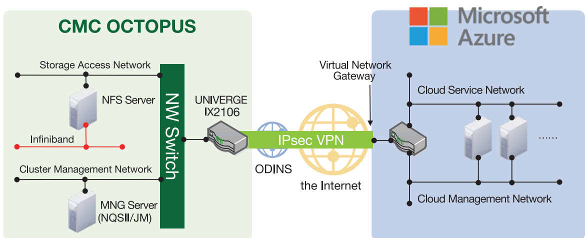
The right figure shows the overview of the first implementation of OCTOPUS-Azure environment where the workload on our on-premise environment is offloaded to the cloud when the demand for computing capacity spikes. For this study, we have introduced Microsoft Azure as an IaaS cloud to be integrated with OCTOPUS.

For realizing this environment, the following three have been considered.

1. Cloud-bridge network.
2. Virtual compute node deployment.
3. Job manager.



Cloud-bridge network



IPSec VPN has been established between on-premise and cloud.
- Lustre on OCTOPUS have to be accessed from Azure.
- NQSII/JM, the job server on OCTOPUS have to communicate with virtual compute node.

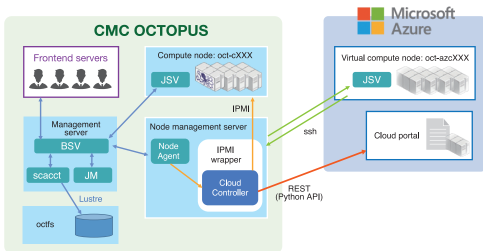
Virtual compute node deployment

Size	vCPU's	Memory:GiB
Standard_F2s_v2	2	4
Standard_F4s_v2	4	8
Standard_F8s_v2	8	16
Standard_F16s_v2	16	32
Standard_F32s_v2	32	64
Standard_F48s_v2	48	96
Standard_F64s_v2	64	128
Standard_F72s_v2	72	144

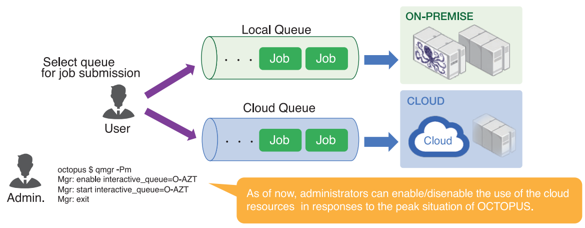
Theoretical Computing Speed	1.463 PFLOPS	
Compute Node	General purpose CPU nodes 236 nodes (471.24 TFLOPS)	CPU : Intel Xeon Gold 6126 (Skylake / 2.6 GHz 12 cores) 2 CPUs Memory : 192 GB
	GPU nodes 37 nodes (858.28 TFLOPS)	CPU : Intel Xeon Gold 6126 (Skylake / 2.6 GHz 12 cores) 2 CPUs GPU : NVIDIA Tesla P100 (NV-Link) 4 units Memory : 192 GB
	Xeon Phi nodes 44 nodes (117.14 TFLOPS)	CPU : Intel Xeon Phi 7210 (Knights Landing / 1.3 GHz 64 cores) 1 CPU Memory : 192 GB
	Large-scale shared-memory nodes 2 nodes (16.38 TFLOPS)	CPU : Intel Xeon Platinum8153 (Skylake / 2.0 GHz 16 cores) 8 CPUs Memory : 6 TB
Interconnect	InfiniBand EDR (100Gbps)	
Stroage	DDN EXAScaler (Lustre / 3.1PB)	

Taking it into consideration that we forward job requests onto OCTOPUS to Azure, virtual compute node should have more CPU cores and memory than OCTOPUS.

Job manager



IPMI wrapper and cloud controller have been developed so that
- We take advantage of NEC NQS II/JM's energy-saving functionality on the cloud.
- NQSII/JM can handle multiple cloud services simultaneously.



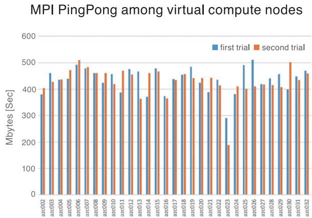
Ideally, these two queues should be integrated to a single queue so that jobs are processed on on-premise or the cloud in a user-transparent manner.

Evaluation

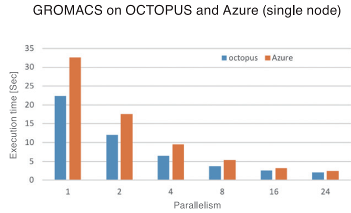
The following criteria have been used for evaluation of this OCTOPUS-Azure environment.

1. On-demand
   - Cloud resources should become available/unavailable in an on-demand way.
2. Transparency
   - Job submission to on-premise and cloud resources should not be different.
3. Selectivity
   - Users have to be able to specify whether they prefer the use of the cloud or not.
4. Equality
   - Computing results should be the "same" as in the cloud
5. High throughput
   - Throughput should be increased and then user waiting time should be reduced.

MPI PingPong among virtual compute nodes



GROMACS on OCTOPUS and Azure (single node)



Cybermedia Center

Osaka University, Japan

SC19 BOOTH 1843

Cloud Bursting with Secure Staging / GPU Burst Buffer with GPU/NVMe Direct

1. Cloud Bursting with Secure Staging

Problem

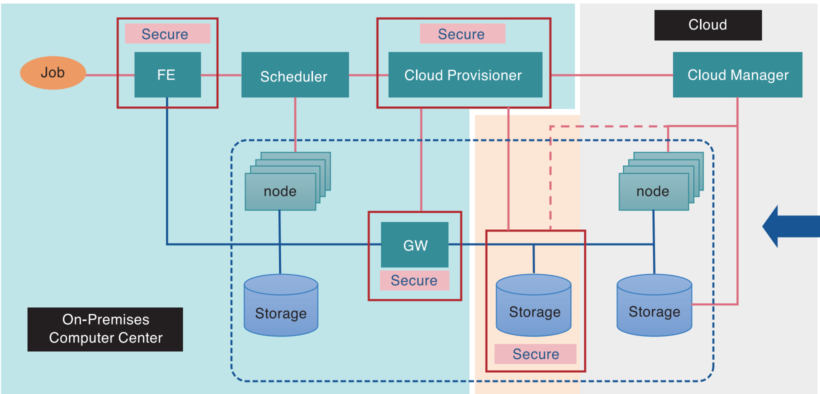
- Data must be shared between cloud and on-premises computer center.
  - Sharing on-premises storage may degrade performance and security level.
  - Some users have security concern about leaving data on shared storage in cloud.

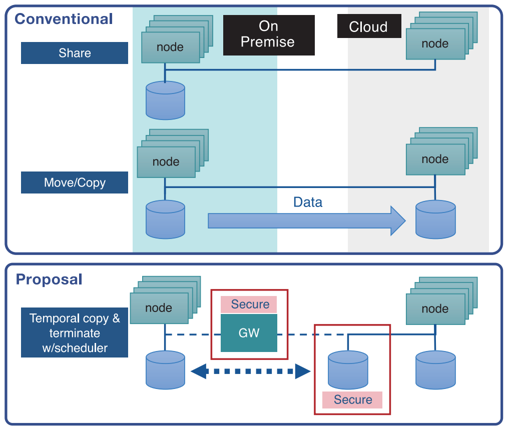
Proposal : Staging data just in time for Cloud Bursting

Operation

1. Job input
2. Resource reservation (node, storage, network)
3. Secure stage in
4. Job execution
5. Secure stage out
6. Resource release

System Architecture





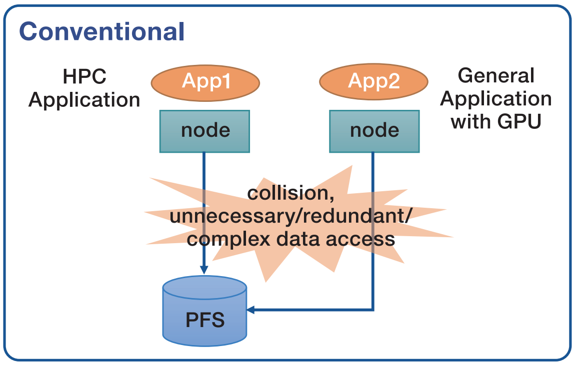
2. GPU Direct Burst Buffer

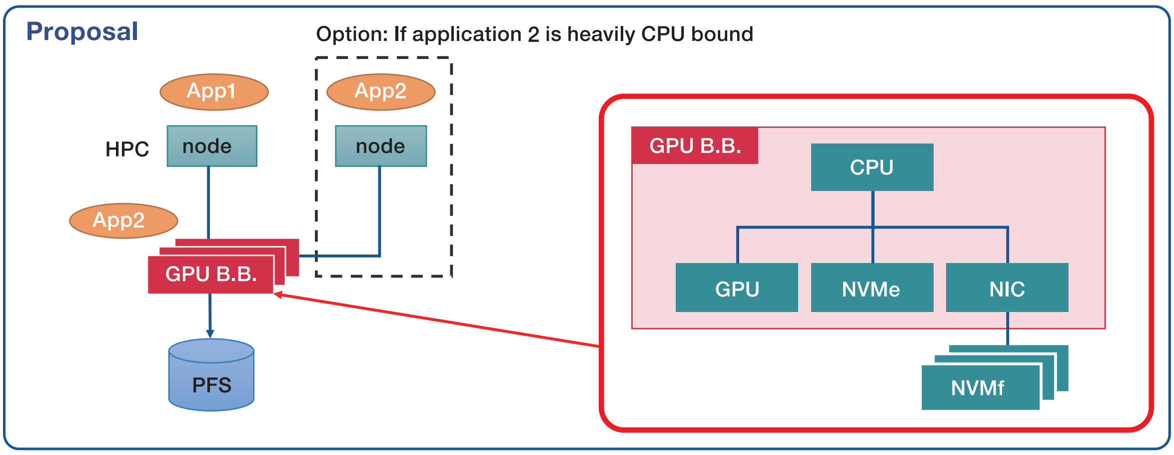
Problem

- Delay in cascading/interacting multiple application/process.
  e.g. visualization/application integration

Proposal : GPU Burst Buffer

- Cache with direct transfer between GPU and NVMe/NVMf.
- 2nd application executed on GPU Burst Buffer.





Contact: sc19@ais.cmc.osaka-u.ac.jp

Cybermedia Center

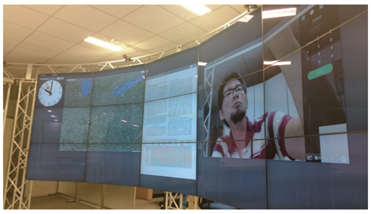
Osaka University, Japan

SC19 BOOTH 1843

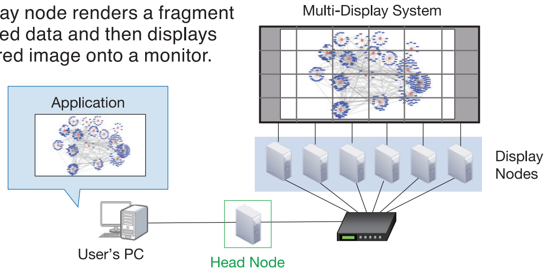
Novel Mechanisms to Support Scientific Visualization on Multi-Display

Multi-Display System

- Multi-Display System (MDS) is a scalable visualization system, which provides **a virtual high-resolution screen** by combining multiple sets of computers and monitors.
- An implementation of MDS is now utilized for scientific visualization.
  - MDS can visualize different types of scientific data without a lack of information. (e.g. simulation results, network graph etc.)
  - A lot of researchers can observe visualized data simultaneously and exchange ideas with each other on the spot.

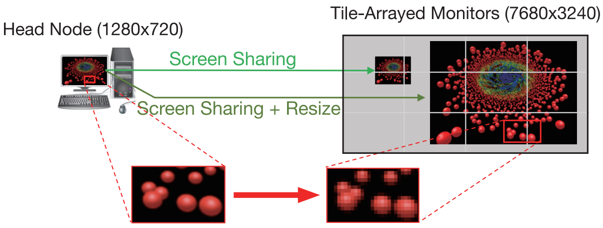


- In general, MDS has a **cluster-based architecture.**
  - The head node and the display nodes are cooperated by dedicated visualization software. (e.g. SAGE2, ParaView, COVISE etc.)
  - The head node provides to allow users to move/resize the window on the MDS.
  - Each display node renders a fragment of visualized data and then displays the rendered image onto a monitor.

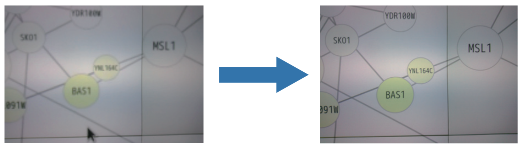


High-Resolution Streaming Functionality in SAGE2 Screen Sharing

- SAGE2 (popular visualization middleware) provides a screen sharing functionality, which is the function to stream user's desktop contents to a multi-display.
  - A screen sharing functionality that allows users to display their own application on their own PC onto the MDS.
- Problem: Resolution constraint
  - The desktop contents are displayed at the same resolution as the monitor of the head node.
  - Large difference in the screen resolution will deteriorate the visibility of desktop applications.

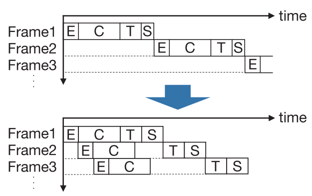


- Proposed method: Virtual screen and pipeline streaming
  - Xvnc creates the virtual screen at an arbitrary resolution on the head node regardless of the specifications of its monitor.



Without our approach (Jaggy) — With our approach (Clearness)

- To improve the frame rate in the high-resolution streaming, the streaming process is pipelined.



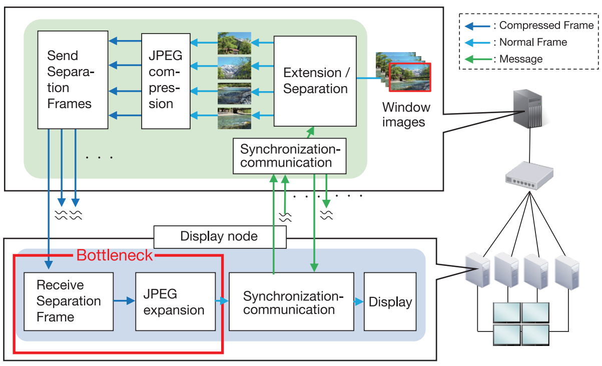
- At 4K resolution 3.55fps → 19.20fps
- At 8K resolution 1.08fps → 5.89fps

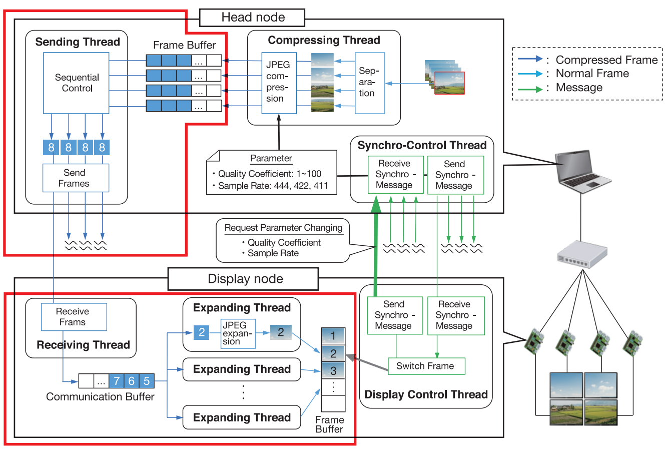
E : Frame Extraction
C : Frame Compression
T : Frame Transmission
S : Synchronization

K. Ishida, et al., "High-Resolution Streaming Functionality in SAGE2 Screen Sharing," Advances in Information and Communication, Proceedings of the 2019 Future of Information and Communications Conference (FICC2019), Lecture Notes in Networks and Systems, vol. 70, pp.384-399, Mar. 2019. [DOI:10.1007/978-3-030-12385-7_30]

High Frame Rate MDS on Low-Spec Computers

- To construct MDS by using high-spec computers, the cost is very high. On the other hand, a low-spec computer like Single Board Computer (SBC: e.g. Raspberry Pi, NVIDIA Jetson Nano) does not cost much but also have graphics performance sufficient for a single monitor.
- Problem: Existing MDS middleware require more powerful computer.
  - SBC's CPU performance is not enough for exiting MDS middleware. Receiving frame packets and JPEG expansion are the major bottlenecks at decreasing frame rate.
  - Using SAGE2 in Raspberry Pi 3, the frame rate is 1-5fps.





- Evaluation and comparison with existing middleware on Raspberry Pi3 (4 display nodes)

- SAGE2 on Raspberry Pi3 : 1.2 fps
- Display Cluster : 2.1 fps
- Proposal Method: 23.2 fps

This research work was supported by the JSPS KAKENHI Grant Number JP18K11355, the "Joint Usage/Research Center for Interdisciplinary Largescale Information Infrastructures" in Japan (Project ID: jh170056-ISJ, jh180077-ISJ) and ROIS NII Open Collaborative Research2019-(19S0104)

Contact: sc19@ais.cmc.osaka-u.ac.jp

Cybermedia Center

Osaka University, Japan

SC19 BOOTH 1843

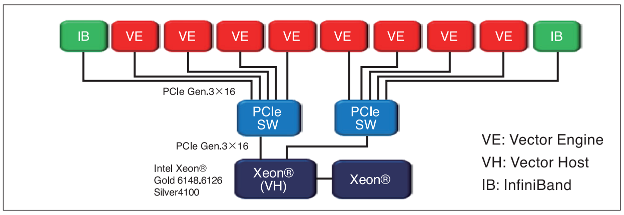
Towards the Future Supercomputing Services at the Cybermedia Center

Tuning on SX-Aurora TSUBASA

・The current flagship supercomputing system at the Cybermedia Center is NEC SX-ACE system. Now we are in the process of investigating SX-Aurora TSUBASA as a candidate processor for the next supercomputing system.

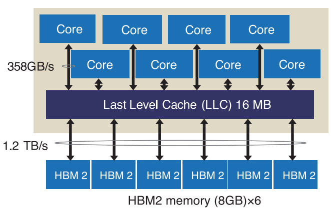
* SX-Aurora TSUBASA
  - a vector processor from NEC.
  - deal with 256 elements with 1 directive.

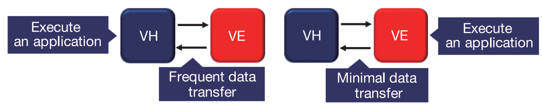
Architecture of SX-Aurora TSUBASA (A300-8)



Architecture of VE

Vector length	256
Core	8
Frequency	1.4 GHz
Performance (Core)	268.8 GF (DP) 537.6 GF (SP)
Performance (VE)	2150.4 GF (DP) 4300.8 GF (SP)
Capacity of LLC	16 MB
Memory bandwidth	1.2 TB/s
Capacity of memory	48 GB





Lesson learnt: Things to keep in mind to improve the performance of applications on SX-Aurora TSUBASA

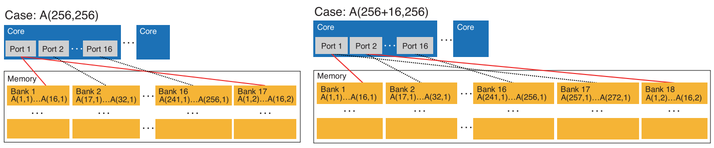
The following three considerations affect on application performance.
- 256 or multiples of 256 iterations: SX-Aurora TSUBASA can deal with 256 elements.
- sufficient margins: avoiding of cpu port conflict increases the performance for non sequential access.
- sequential access: avoiding of non sequential access increases the performance.

Example code

```
REAL, DIMENSION(256+16, 256)::A, B
  DO j=1,256
    DO i=1,256
      A(j, i) = B(i, j)
```

sufficient margins / 256 or multiples of 256 iterations / sequential access

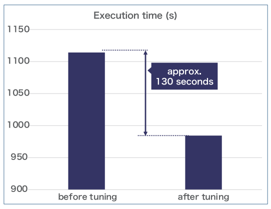
If array A doesn't have sufficient margins in the example code, load/store requests are concentrated on a specific port.



An application example

Radiation Fluid simulation code running on SX-Aurora TSUBASA was accelerated as follows.

	Execution time (s)
before tuning	1114.085353
after tuning	984.334410

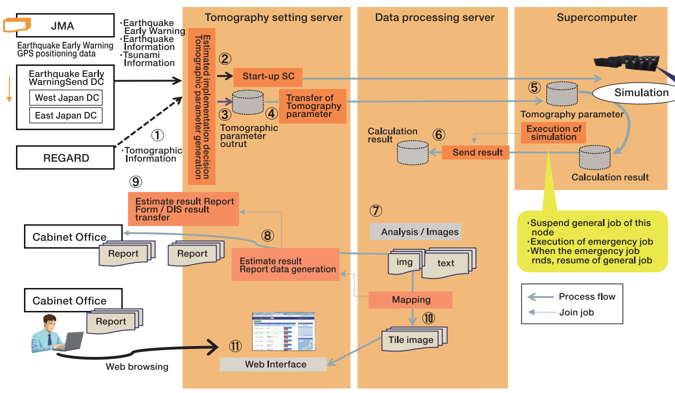


Reproduction of a system that executes emergency jobs

Tsunami inundation damage estimation system is now operated using two SX-ACE systems at Tohoku University and Osaka University.
The system completes Tsunami inundation damage estimation in 30 minutes after a large-scale earthquake that may trigger Tsunami happens.
The current coverage of the system is from the Izu Peninsula to Osumi Peninsula.

RTi-cast

Tsunami inundation damage estimation system



・OS: SUPER-UX (UNIX SystemV + NEC Extension)
RHEL7 (Red Hat Enterprise Linux7)
・Batch system: NQSII

The functionality with which the system stops running jobs, runs this emergency job and then recovers the previous running jobs to the original status heavily depends on NEC NQSII/JM, the proprietary job scheduler proprietary from NEC.

A. Musa, et al., "Real-time Tsunami Inundation Forecast System for Tsunami Disaster Prevention and Mitigation," The Journal of Supercomputing, vol. 74, pp.3093-3113, 2018. [DOI:10.1007/s11227-018-2363-0]

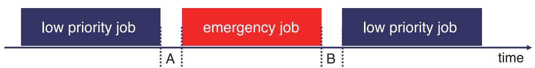
Investigation of Slurm

Currently, Slurm supports the following three modes for suspending, executing and recovering of jobs. We are investigating whether the upper two can be used as an alternative of NQSII.

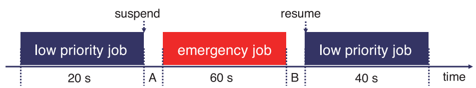
- Suspend low priority job, Execute emergency job, Resume low priority job
- Cancel low priority job, Execute emergency job, Requeue low priority job
- Stop at the checkpoint of the low priority job, Execute the emergency job, Restart from the checkpoint of the low priority job

Experiment

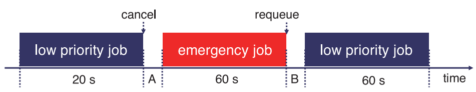
We conducted an experiment to investigate whether Slurm, OSS scheduler, can be used as an alternative to NQSII or not. In the experiment the following time was measured when we switched a running job to an emergency job.
- A: Switching time from a low priority job to an emergency job
- B: Switching time from the emergency job to the low priority job



Result



suspend/resume	
A	0.4 s
B	0.5 s



requeue	
A	0.4 s
B	100 s

Contact: sc19@ais.cmc.osaka-u.ac.jp