Accelerated ONION based on DTN Experience

Susumu Date, Ph. D
Cybermedia Center, Osaka University, Japan

Cybermedia Center, Osaka University

CMC main building

IT core as data center

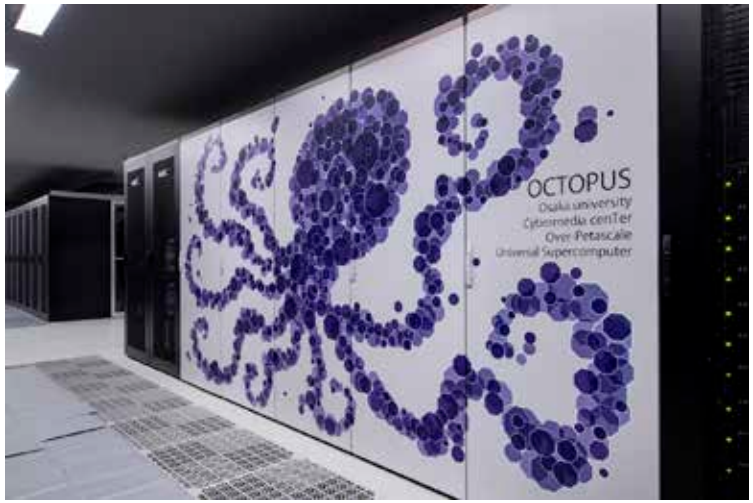
- Supercomputing center at Osaka University
  - has a responsibility of providing a powerful high-performance computing environment for university researchers across Japan as a national joint-use facility.

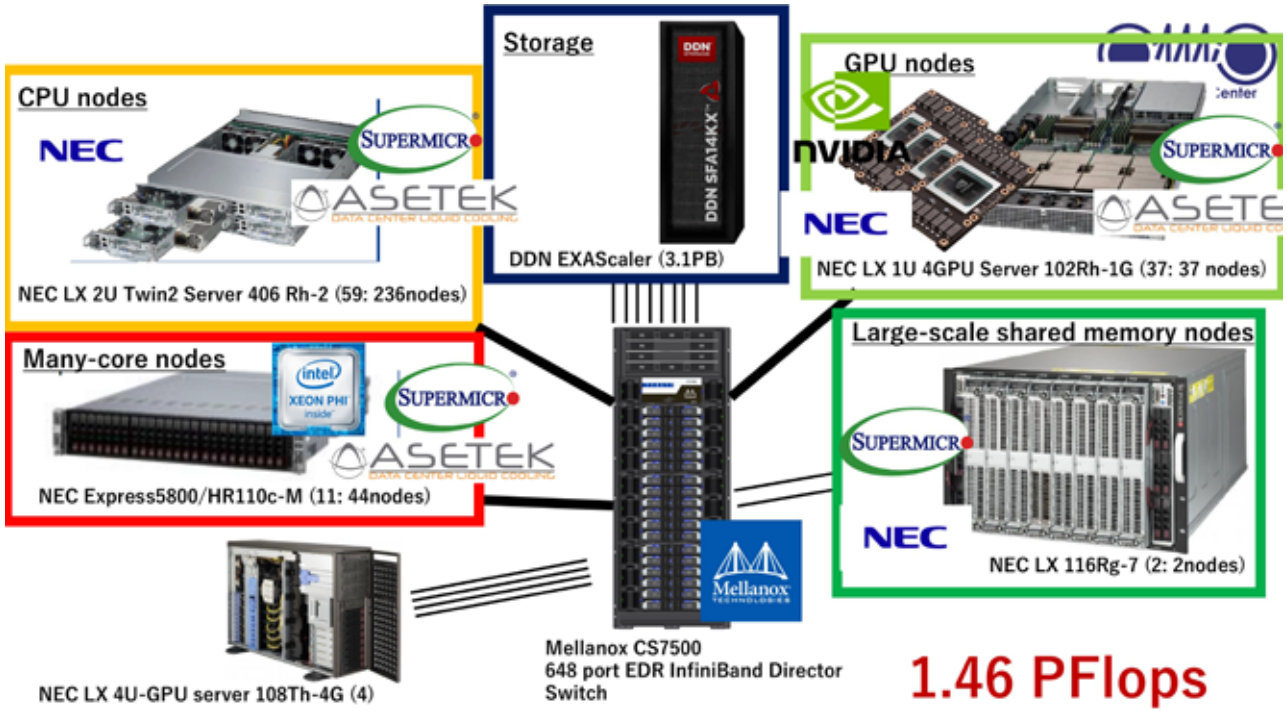
OCTOPUS since 2017

Cybermedia Center
Osaka University

➢ **OCTOPUS (Osaka university Cybermedia cenTer Over-Petascale Universal Supercomputer (Dec. 2017 ~)**

- Peak Performance 1.46 PetaFlops
- DLC Hybrid Supercomputer systems





➢ **Development of new computing needs (“Reclamation”)**

Support for computing demands and need in medical, dental and healthcare scientific areas
Expectation of DL, ML, and AI using supercomputing systems from academic research

SQUID since May 2021

Supercomputer for Quest to Unsolved Interdisciplinary Datascience

Cybermedia Center, Osaka University

- **Cloud-linked High Performance Computing and High Performance Data Analysis Supercomputer System (Supercomputer for Quest to Unsolved Interdisciplinary Datascience)**
  - **Peak Performance 16.591 PFlops**

SQUID システム構成

CPU nodes

1520 nodes x peak perf. 5.837 TFlops　8.871 PFLOPS

プロセッサ	Intel Xeon Platinum 8368 (Ice Lake / 2.40 GHz 38コア) 2 基
主記憶容量	256 GB

GPU nodes

42 nodes x peak perf. 161.836 TFlops　6.797 PFLOPS

プロセッサ	Intel Xeon Platinum 8368 (Ice Lake / 2.40 GHz 38 コア) 2 基
主記憶容量	512 GB
GPU	NVIDIA HGX A100 8 GPU ボード (Delta)

Vector nodes

36 nodes x peak perf. 25.611 TFlops　0.922 PFLOPS

プロセッサ	AMD EPYC 7402P (2.8 GHz 24コア) 1 基
主記憶容量	128 GB
Vector Engine	NEC SX-Aurora TSUBASA Type 20A 8 基

Interconnect

ノード間接続	Mellanox InfiniBand HDR (200 Gbps)

ONION data aggregation Infra.

S3-compatible Parallel File System 21.2PB

ファイルシステム	DDN EXAScaler (Lustre)
HDD	20.0 PB
SSD	1.2 PB

S3-compatible Object Storage 500TB

オブジェクトストレージ	CLOUDIAN HyperStore
HDD	500 TB

My Current Research Motivation

- To develop and provide a research platform where researchers and scientists can perform data-intensive research using our supercomputing systems (at the Cybermedia Center).
  - because I am in charge of the administration and management of supercomputing systems as associate professor of the center.
- The research platform should be the environment that allows researchers and scientists to exchange large amount of scientific data and perform large-scale computation among collaborators in the world.
  - I hope to provide a research platform that can improve research productivity of researchers who perform data intensive science using supercomputing systems.

How should we as a supercomputing center can support researchers who work on data-intensive science in this globalized research scene?

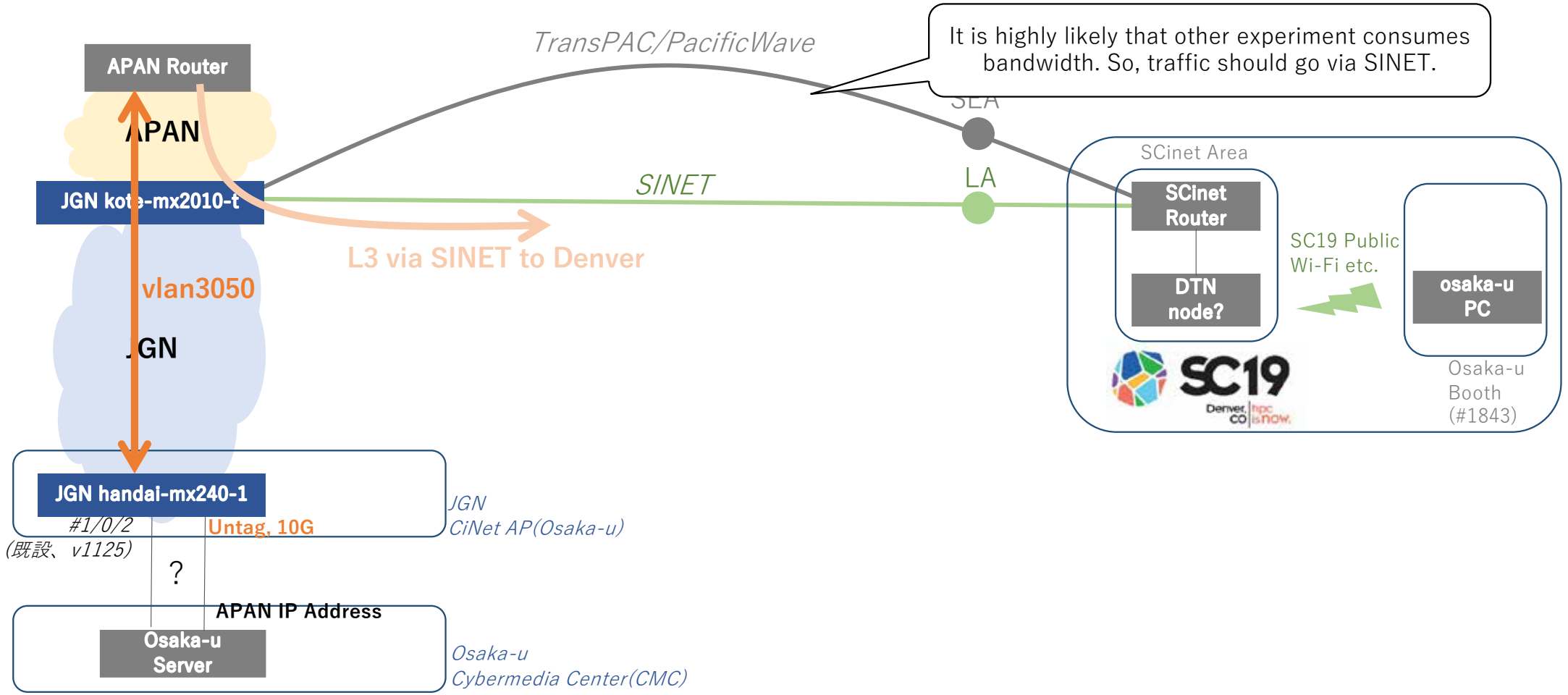
History of our DTN project with Jim Chen/StarLight

- In PRAGMA, CENTRA, SEAIP, and GRP held around 2018 and 2019 (before COVID-19), we discussed about the possibility of research collaboration and Jim asked the possibility of setting up a DTN node at Osaka University.
  - I guess, the trigger was for SCAsia Data Mover Challenge 2019.
- Jim shipped a DTN node to Osaka University and we connected the node to JGN 10G network thanks to NICT.

Intel® NUC 8 Mainstream-G mini PC with QNAP QNA-T310G1S

History2 of our DTN project with Jim Chen/StarLight

- SC19 Experiment conducted between Osaka University and SC venue.



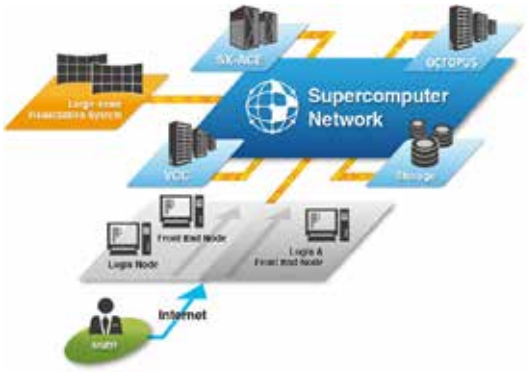
Result of performance measurement in SC19
- CMC to SC venue -

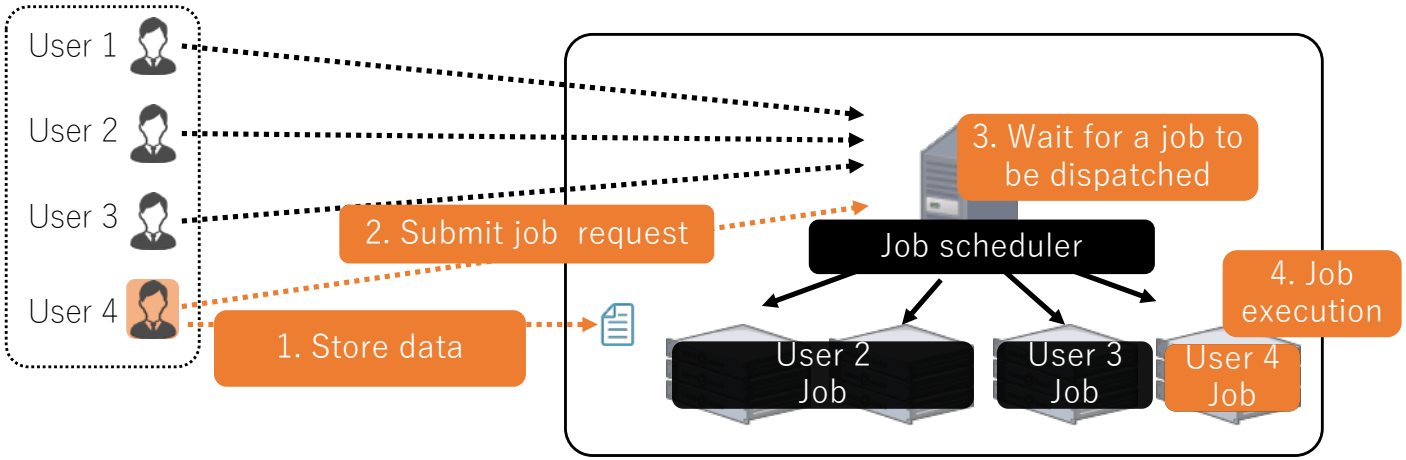
Memory to Memory

NVMe to NVMe

Motivation of an On-Time Data Transfer Framework in Cooperation with Scheduler System

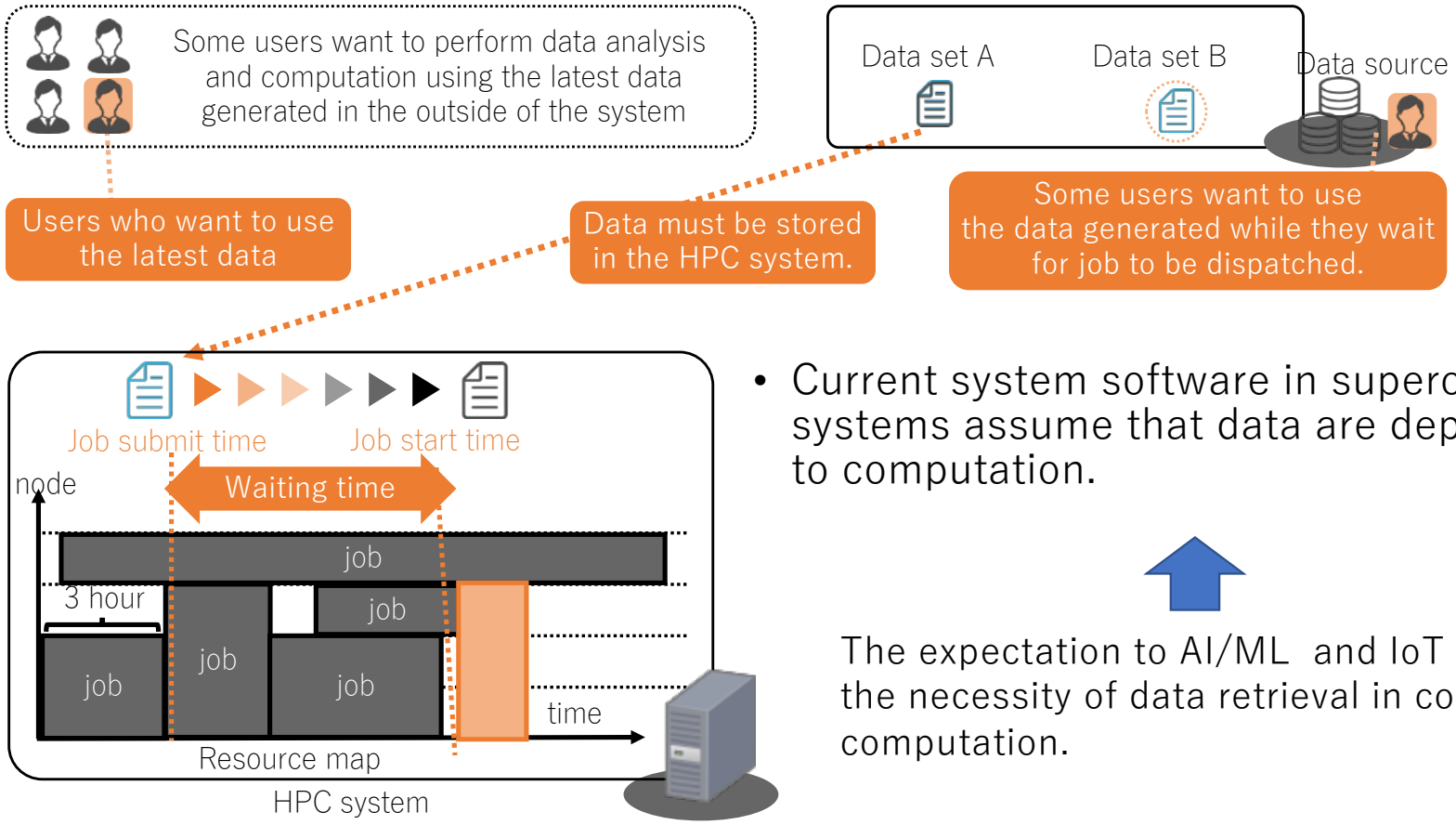
- Supercomputing systems are used in an “isolated” manner.
  - Recent academic research is globalized and requires the aggregation of scientific knowledge and data for problem solving, meaning that data and computing results should be immediately exchanged and shared for collaboration.
- Most of supercomputing systems assume that the data to be used for computation are stored in the inside of supercomputing systems.
  - For example, scheduler system assumes that users submit job requests after storing their data on supercomputing systems.
- SC19 results was interesting!





Motivation of an On-Time Data Transfer Framework in Cooperation with Scheduler System

- Waiting time makes data outdated.

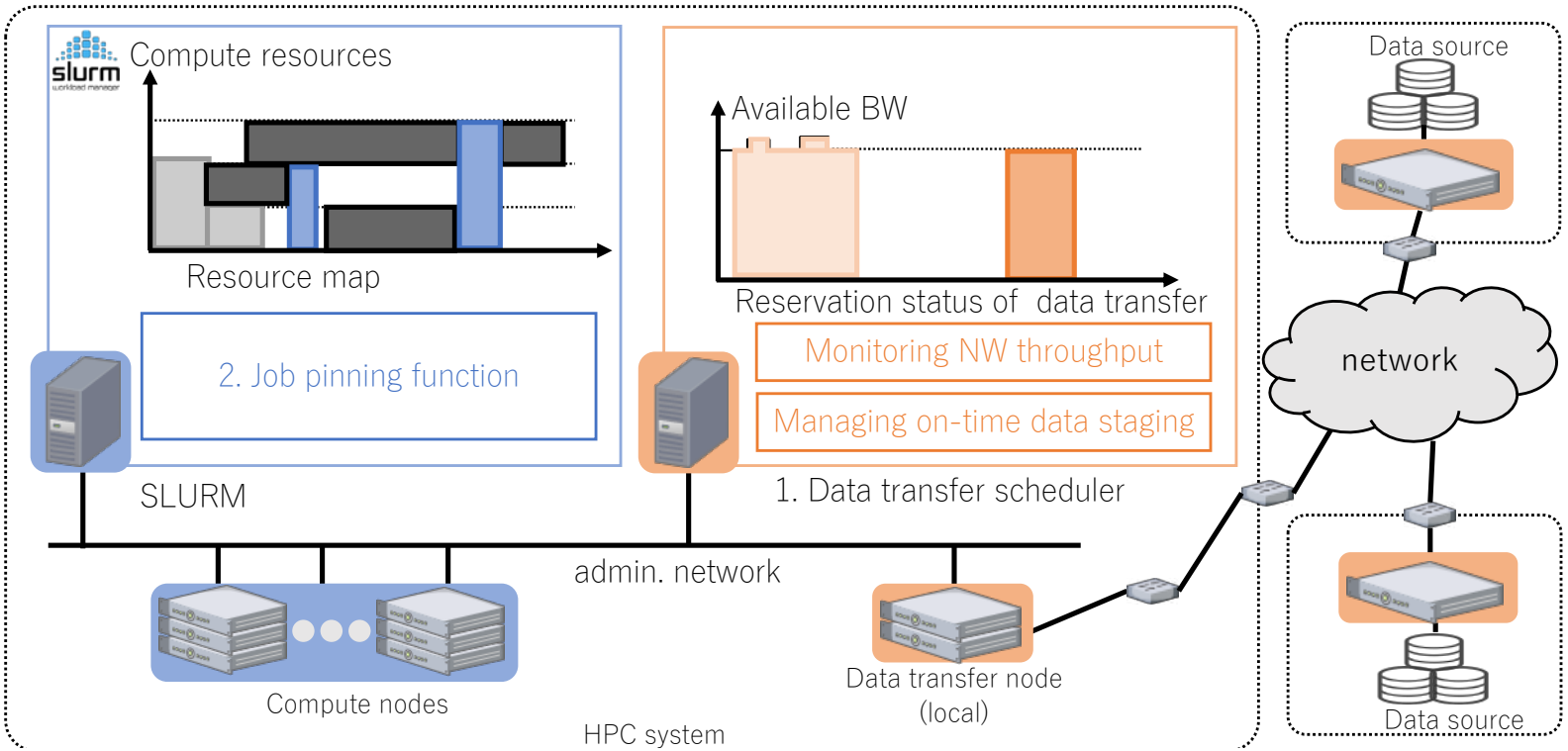


- Current system software in supercomputing systems assume that data are deployed in prior to computation.

The expectation to AI/ML and IoT increases the necessity of data retrieval in cooperation with computation.

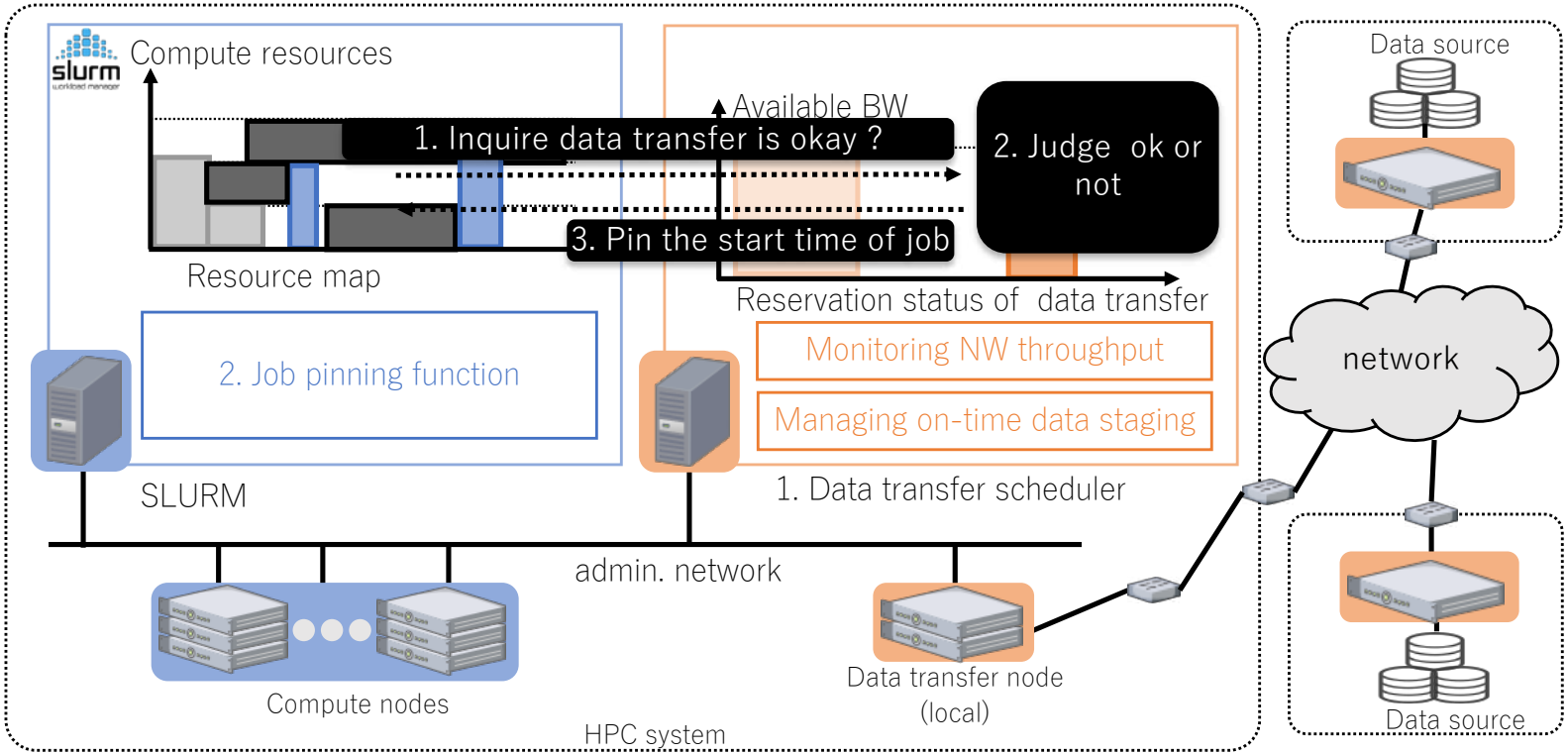
On-Time Data Transfer Framework in Cooperation with Scheduler System

- We have prototyped on-time data transfer framework so that data can be delivered just before job execution.
  - keep utilization high and data fresh even when the job requesting the data transfer is submitted
  - Data transfer scheduling function and job pinning function as on-time data staging module have been integrated into Slurm scheduler.



On-Time Data Transfer Framework in Cooperation with Scheduler System

- Current implementation simply compares the data amount to be transferred and available BW under the interaction of job pinning and data transfer scheduler.

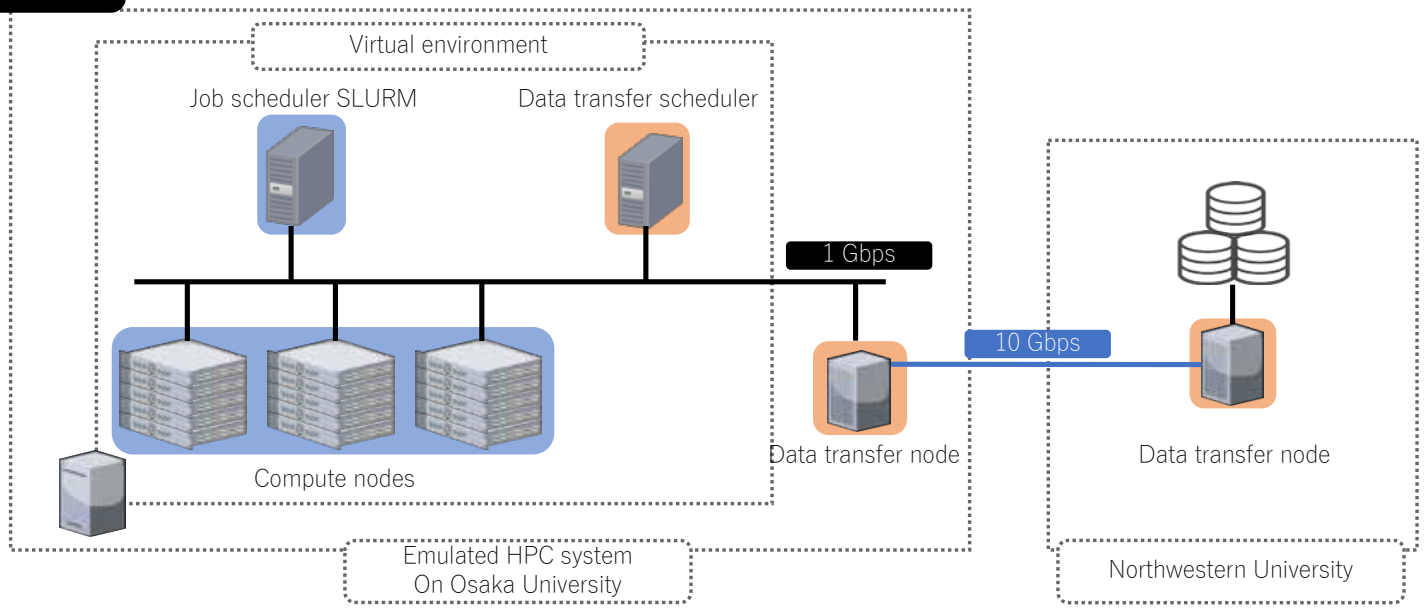


Evaluation

Evaluation goal

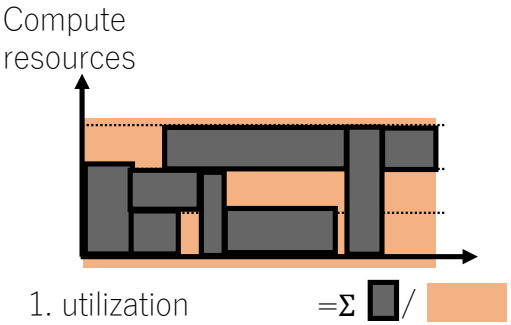
Investigate whether the data transfer is completed before job execution start as well as whether the proposed framework keep the utilization of computing resources

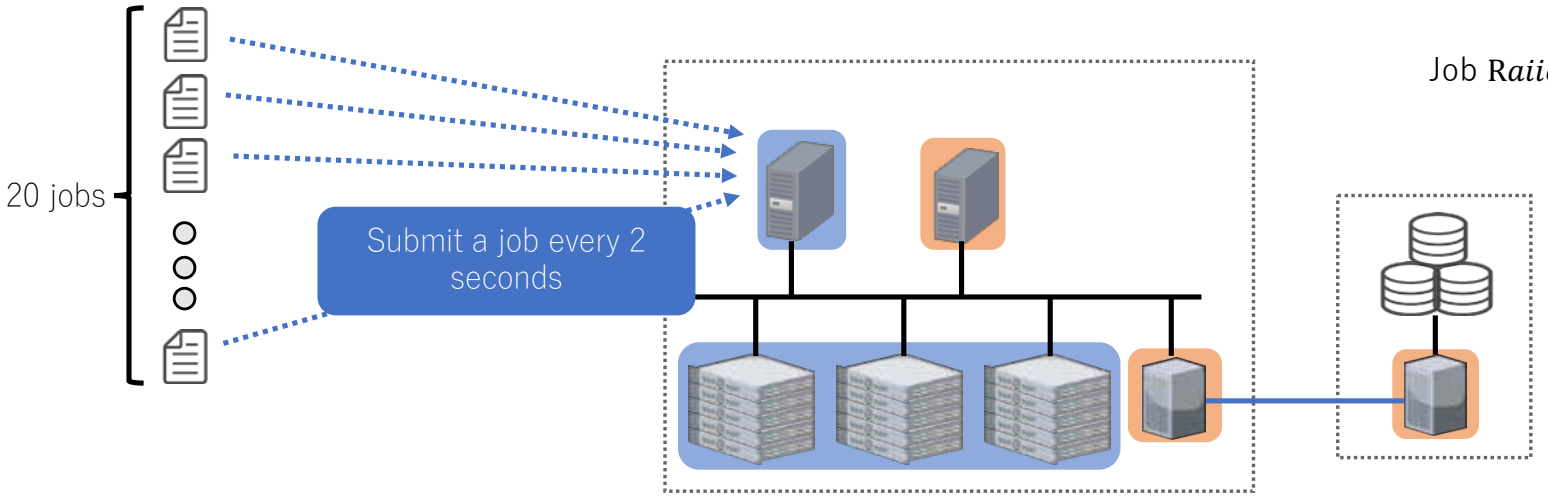
Environment



Evaluation method

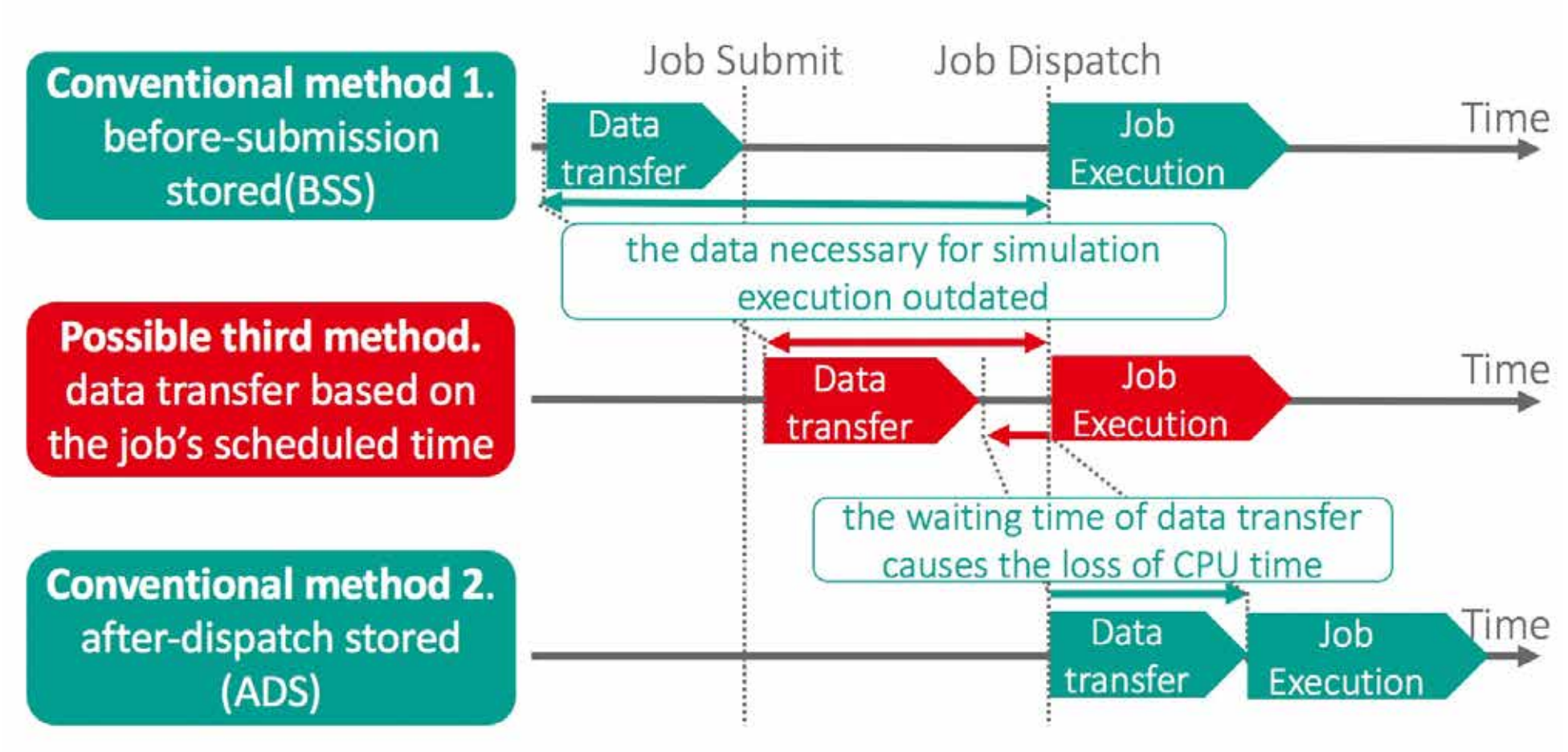
- Submit jobs assuming multiple users
- Compare the following criteria w/ and w/o the proposed framework by changing the job ratio.
  1. Utilization of compute resources
  2. Waiting time from submission until job execution start.



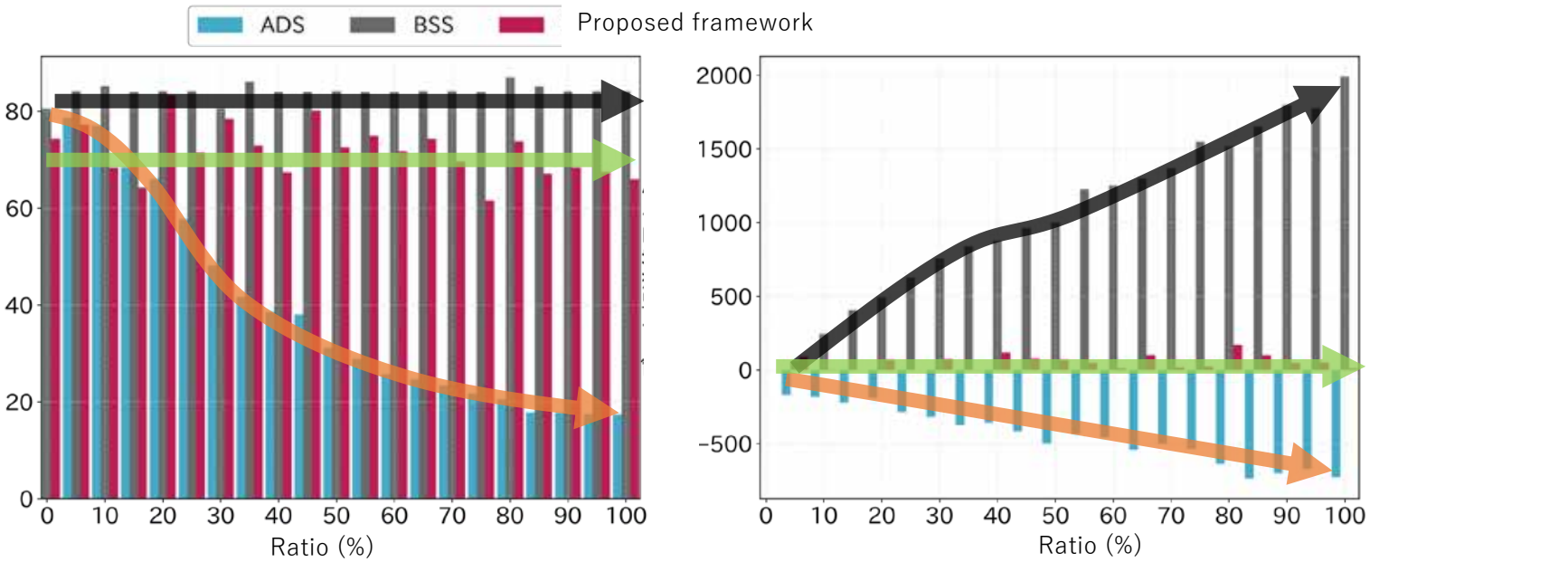


$$\text{Job } Raiio = \frac{number\ of\ jobs\ requiring\ the\ data\ located\ not\ in\ the\ system}{number\ of\ all\ jobs} * 100$$

Evaluation method



Evaluation Result



(a) Utilization of compute resources

(b) Waiting time from data transfer completion to job execution time

- BSS(data staging before job submission): utilization is high, but waiting time increase with the increase of the job ratio
- ADS(data staging after dispatch): utilization decreases but wating time is short .
- Proposed framework keeps the utilization irrespective of job and waiting time is kept short.

ONION and DTN

SQUID since May 2021

Supercomputer for Quest to Unsolved Interdisciplinary Datascience

Cybermedia Center, Osaka University

- **Cloud-linked High Performance Computing and High Performance Data Analysis Supercomputer System (Supercomputer for Quest to Unsolved Interdisciplinary Datascience)**
  - **Peak Performance 16.591 PFlops**

SQUID システム構成

CPU nodes

1520 nodes x peak perf. 5.837 TFlops　8.871 PFLOPS

プロセッサ	Intel Xeon Platinum 8368 (Ice Lake / 2.40 GHz 38コア) 2 基
主記憶容量	256 GB

GPU nodes

42 nodes x peak perf. 161.836 TFlops　6.797 PFLOPS

プロセッサ	Intel Xeon Platinum 8368 (Ice Lake / 2.40 GHz 38 コア) 2 基
主記憶容量	512 GB
GPU	NVIDIA HGX A100 8 GPU ボード (Delta)

Vector nodes

36 nodes x peak perf. 25.611 TFlops　0.922 PFLOPS

プロセッサ	AMD EPYC 7402P (2.8 GHz 24コア) 1 基
主記憶容量	128 GB
Vector Engine	NEC SX-Aurora TSUBASA Type 20A 8 基

Interconnect

ノード間接続	Mellanox InfiniBand HDR (200 Gbps)

ONION data aggregation Infra.

S3-compatible Parallel File System 21.2PB

ファイルシステム	DDN EXAScaler (Lustre)
HDD	20.0 PB
SSD	1.2 PB

S3-compatible Object Storage 500TB

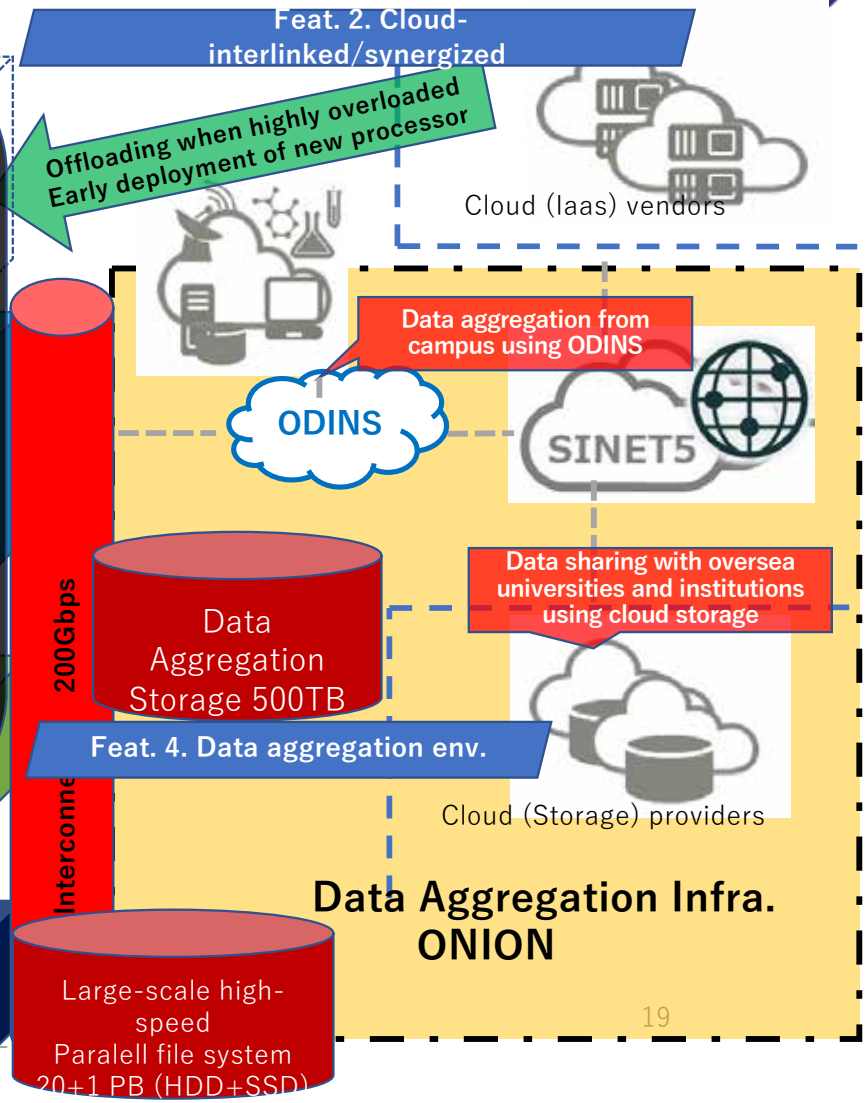
オブジェクトストレージ	CLOUDIAN HyperStore
HDD	500 TB

Five Features in SQUID

ONION, data aggregation infrastructure

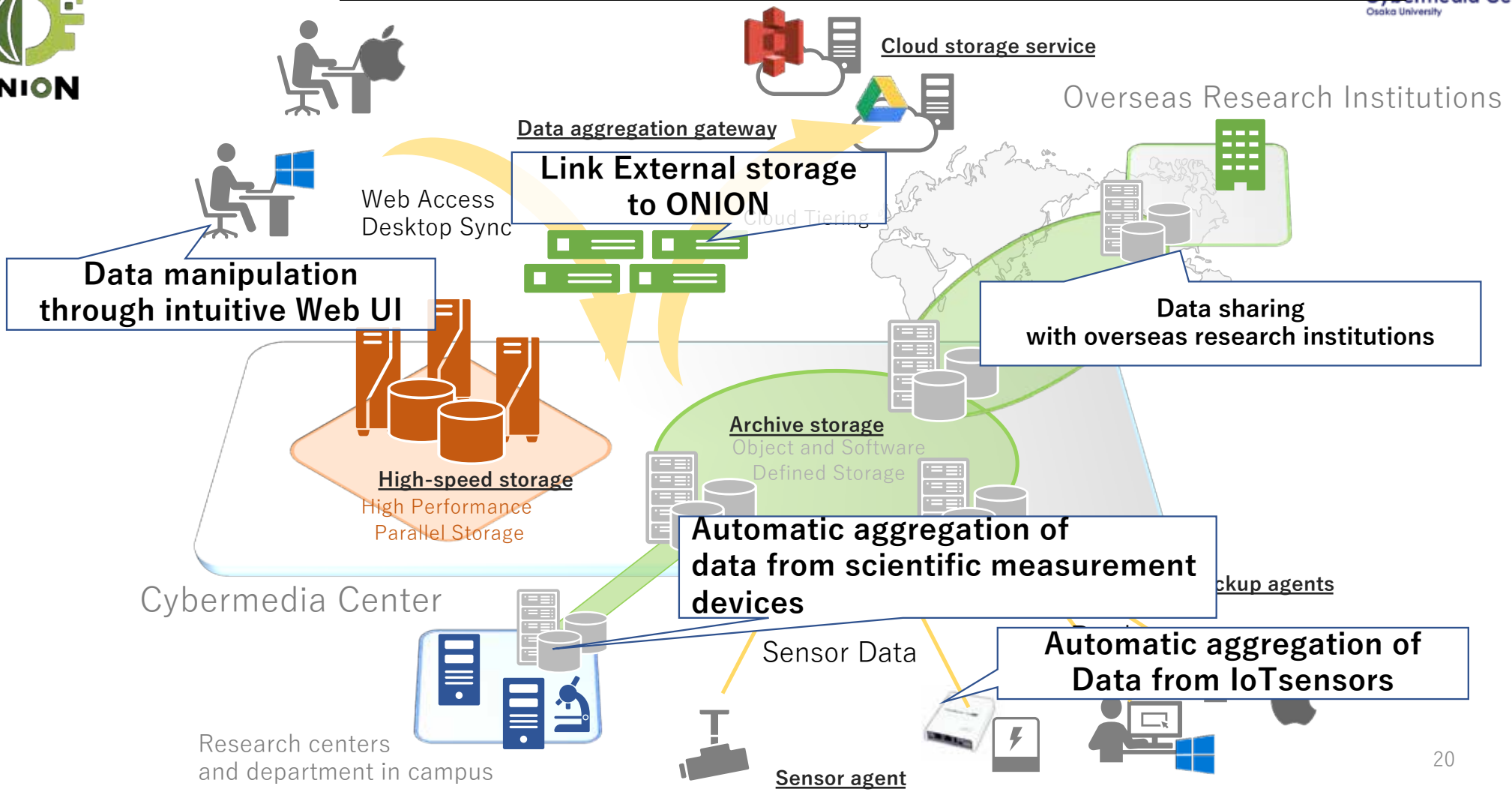
- Osaka university Next-generation Infrastructure for Open research and open innovatioN
- Data aggregation infrastructure that not only enables the sustainable handling of ”super big data” generated in Osaka University in a responsible manner while ensuring the sustainability of such data into the future but also facilitates utilization of research data for “co-creation between academia and industries” and “international research collaboration”.
- introduce as PoC (Proof-of Concept) implementation in the procurement of SQUID on a trial basis.

(1) The primary purpose of procurement is supercomputing system, not for data storage.
(2) Our designed ONION might not be useful and thus not be used.

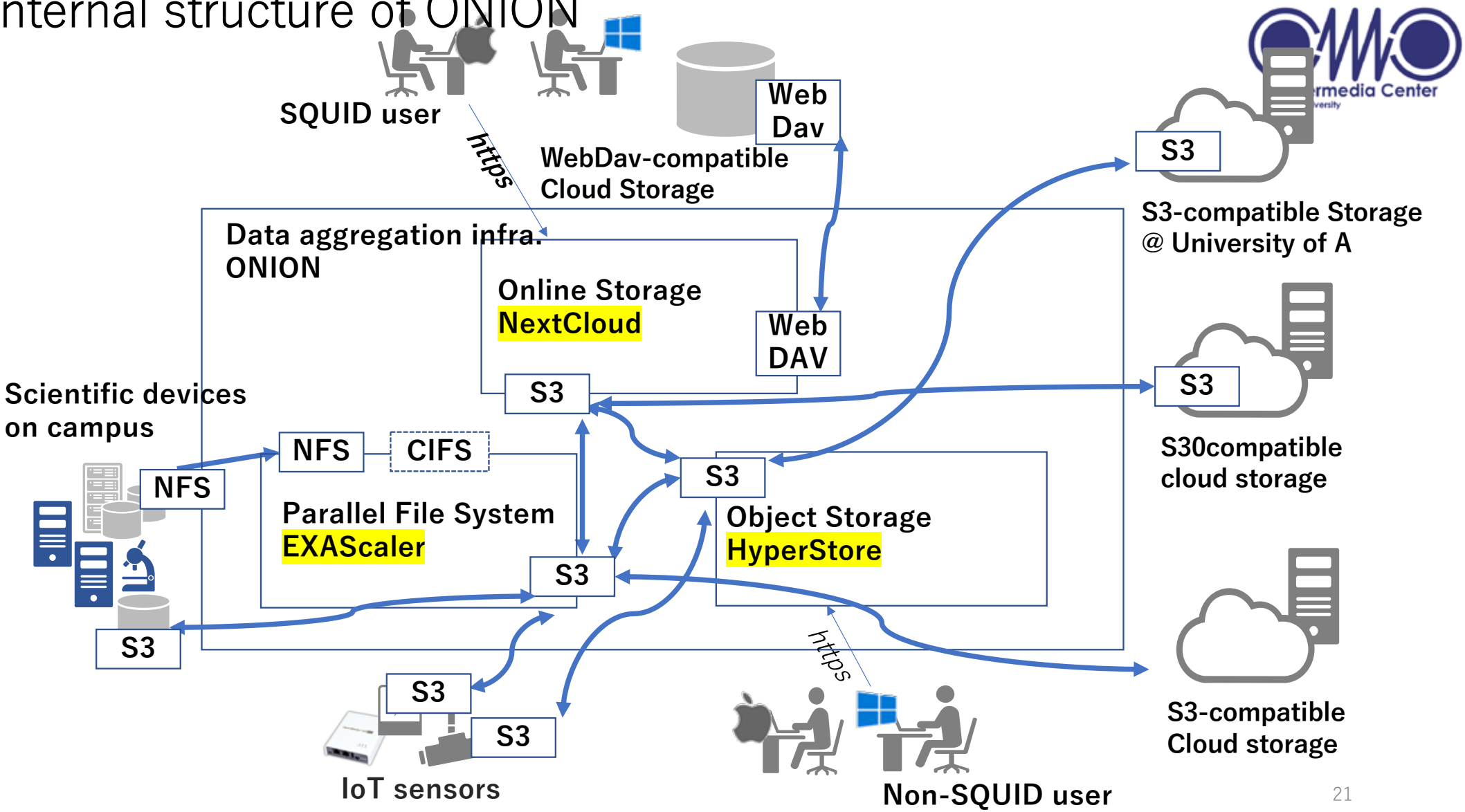


Overview of ONION

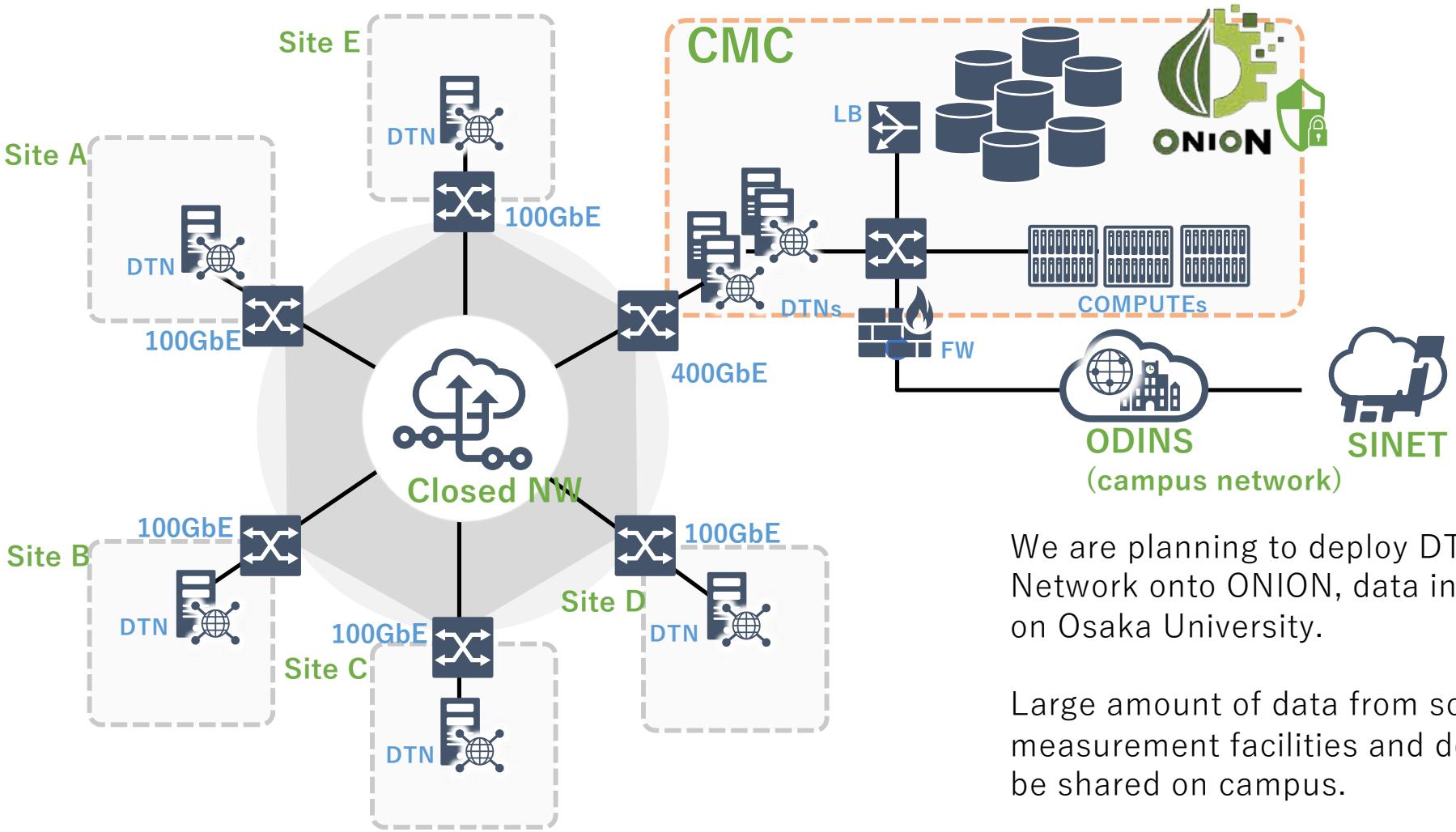
3 storage solutions are seamlessly combined in a synergic manner



Internal structure of ONION



DTN- ONION On Campus (towards Science DMG)



We are planning to deploy DTN and high-speed Network onto ONION, data infrastructure on Osaka University.

Large amount of data from scientific measurement facilities and devices can be shared on campus.

Summary

Cybermedia Center
Osaka University

- We enjoyed DTN performance for on-time data transfer in cooperation with scheduler system.
- Also, based on my experience and expectation to DTN, I personally feel that our supercomputing systems can be changed by leveraging DTN so that our systems can accommodate the requirements and needs from researchers who work on academic research through the global collaboration.