JAMSTEC
Japan Agency for Marine-Earth Science and Technology
Center for Earth Information Science
and Technology (CEIST)

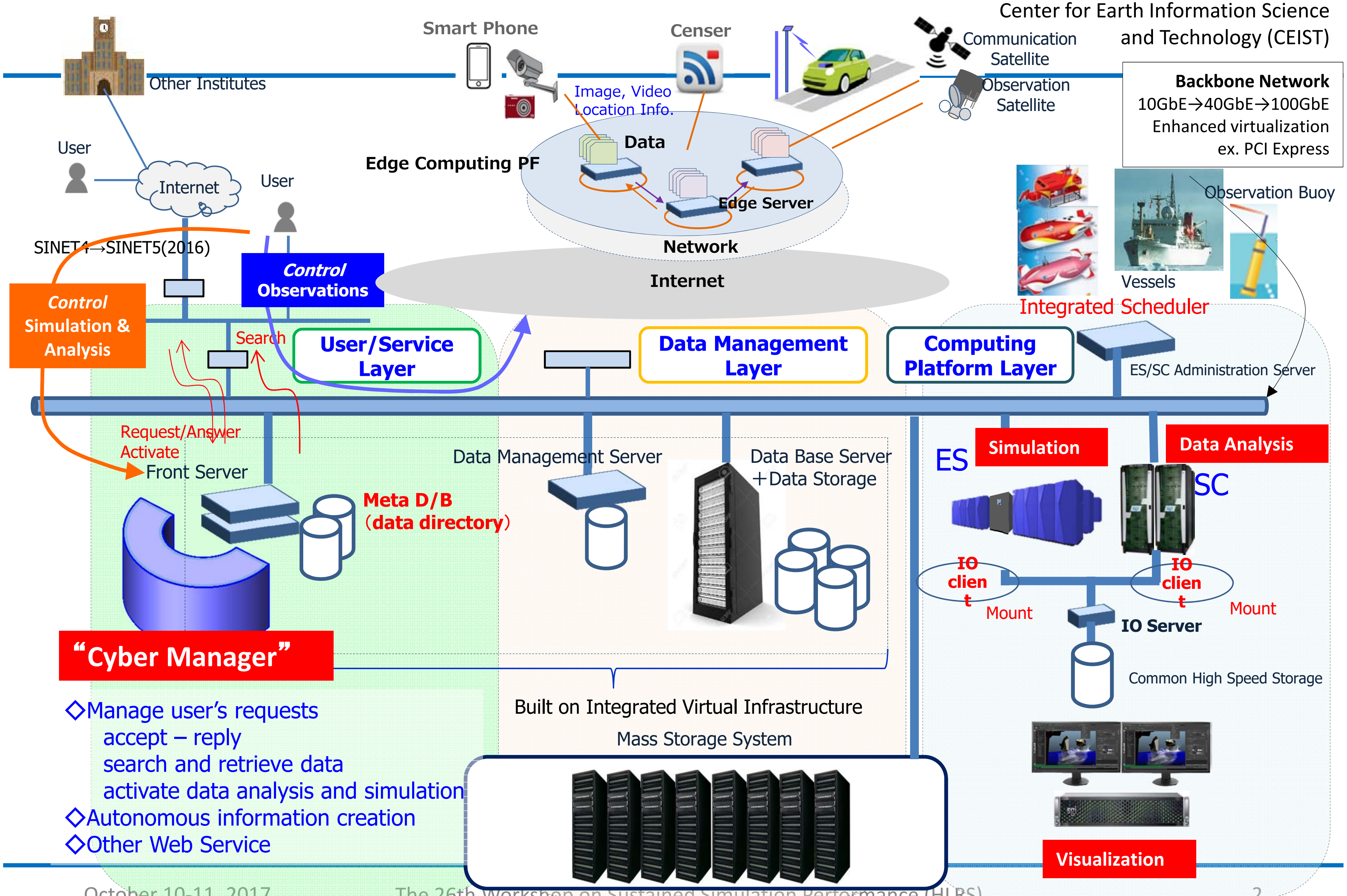
JAMSTEC Next Scalar Supercomputer System

Ken'ichi Itakura (JAMSTEC)

JAMSTEC Earth Informatics Cyber System

JAMSTEC
Japan Agency for Marine-Earth Science and Technology

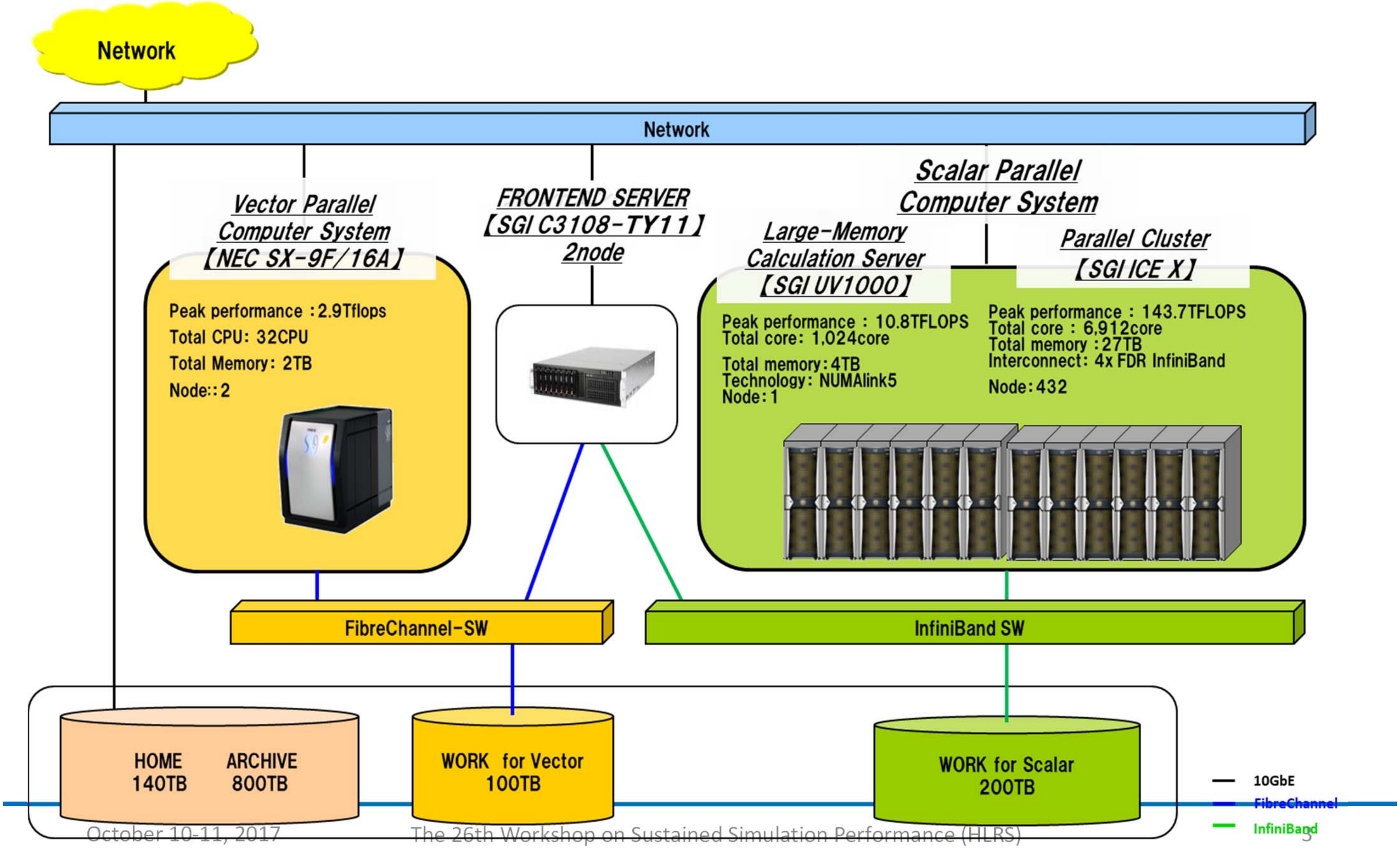
Center for Earth Information Science
and Technology (CEIST)



◇Manage user’s requests
accept – reply
search and retrieve data
activate data analysis and simulation
◇Autonomous information creation
◇Other Web Service

The Current Scalar Supercomputer System

JAMSTEC
Japan Agency for Marine-Earth Science and Technology
Center for Earth Information Science and Technology (CEIST)



Basic Concept

- We need a successor of the present super computer (SC) system.
- We plan for an upgrade and making the next term general-purpose high-performance computer system.

- Common platform with numerical models developed in all of the world
- Co-operating with the "Earth Simulator"
- Pre-/post-processing
  - Prediction experiment in which observation and model are integrated in real time
  - Large-scale calculation of "Earth Simulator"
- Analysis of oceanic global environment with big data analysis, machine learning, artificial intelligence, etc.
- Integer / Logical operation such as full search of parameter space by bioinformatics or statistical method

Next "SC" system – complements ES /meets growing needs

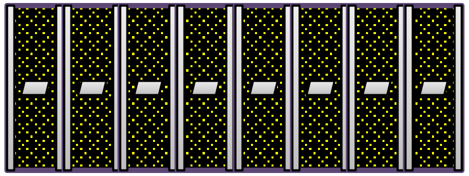
- Enhance existing functions - pre and post processing, data analysis, hosting community APs
- Flexible to meet the emerging and growing needs from various scientific approach
- Capable to host project servers in the future

ANALYSIS

Next "SC" system

[Standard Linux Cluster]

Standard CPU, Considerable computing loads, Commercial APs, Community codes



Growing & Emerging Needs

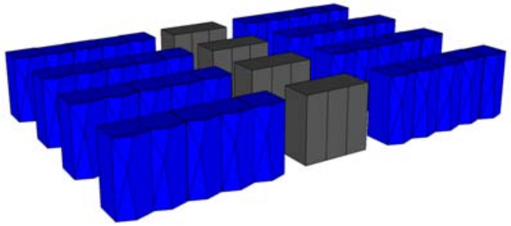
Interoperability: job, file, accounting etc.

SIMULATION

Earth Simulator

[High-end Supercomputer]

Custom CPU, High memory bandwidth, High vector performance, Proprietary codes



Used for:

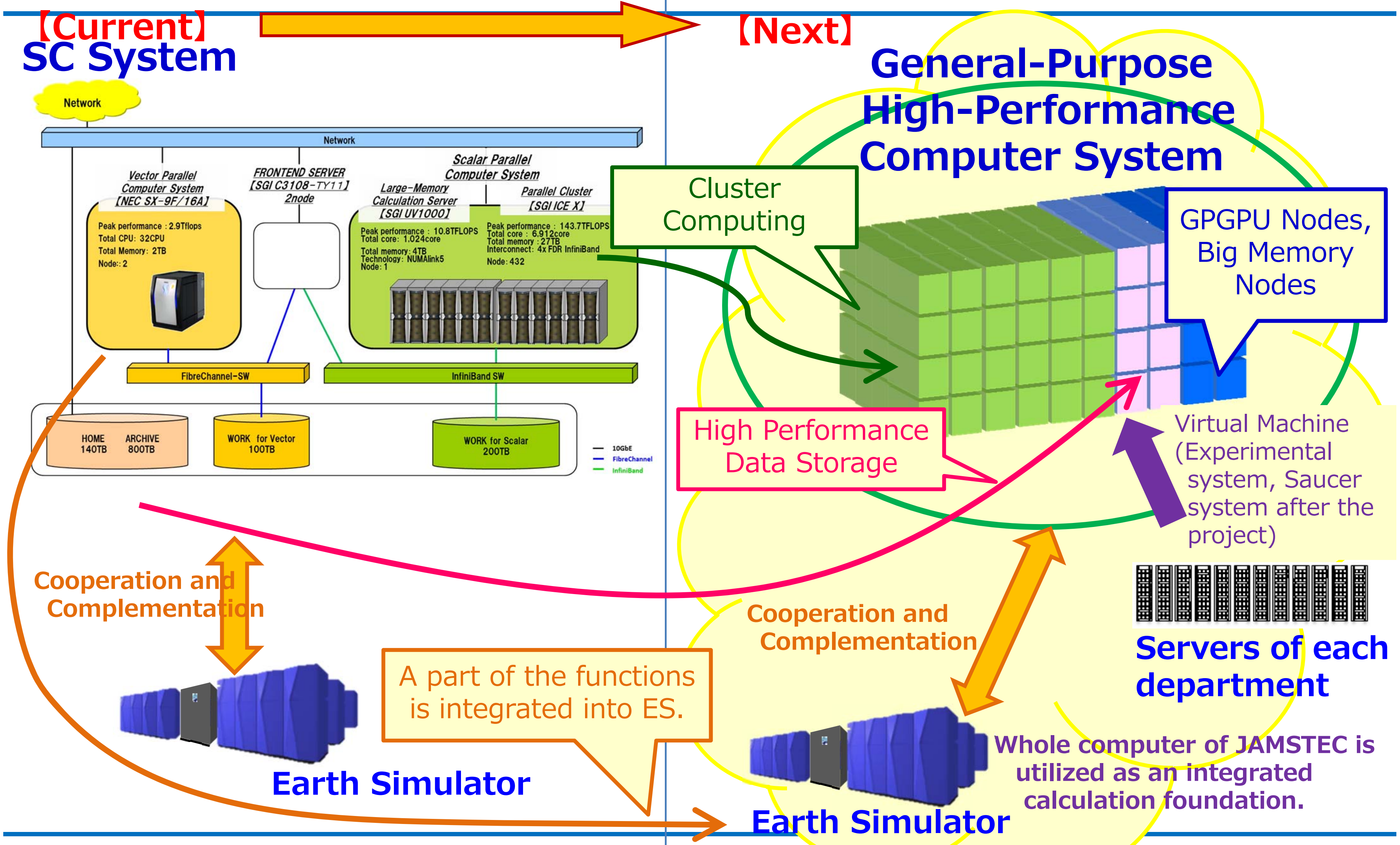
- **Community codes on standard Linux clusters.**
- **Programs of integer and Logical operation,**
- **large memory space or high i/o performance.**
- **Bio informatics.**
- **Earth informatics.**
- **Big data analysis.**
- **Real time computing for observation and ship-operation.**
- **Industry applications.**

Integrated use with Earth Simulator: (example)

- **run reginal model downscaling from global model on ES**
- **run short term forecast while ES runs seasonal forecast**

The Next Term General-Purpose High-Performance Computer System

Center for Earth Information Science and Technology (CEIST)



【Current】

SC System

【Next】

General-Purpose High-Performance Computer System

Cluster Computing

GPGPU Nodes, Big Memory Nodes

High Performance Data Storage

Virtual Machine (Experimental system, Saucer system after the project)

Servers of each department

Cooperation and Complementation

Cooperation and Complementation

A part of the functions is integrated into ES.

Earth Simulator

Earth Simulator

Whole computer of JAMSTEC is utilized as an integrated calculation foundation.

Network
Network
Vector Parallel Computer System 【NEC SX-9F/16A】
Peak performance : 2.9Tflops
Total CPU: 32CPU
Total Memory: 2TB
Node:: 2
FRONTEND SERVER 【SGI C3108-TY11】 2node
Scalar Parallel Computer System
Large-Memory Calculation Server 【SGI UV1000】
Peak performance : 10.8TFLOPS
Total core: 1,024core
Total memory: 4TB
Technology: NUMAlink5
Node: 1
Parallel Cluster 【SGI ICE X】
Peak performance : 143.7TFLOPS
Total core : 6,912core
Total memory : 27TB
Interconnect: 4x FDR InfiniBand
Node: 432
FibreChannel-SW
InfiniBand SW
HOME 140TB
ARCHIVE 800TB
WORK for Vector 100TB
WORK for Scalar 200TB
10GbE
FibreChannel
InfiniBand

The Next Term General-Purpose High-Performance Computer System

- The new system is used from February, 2018.
- Successor of a General-Purpose Cluster System

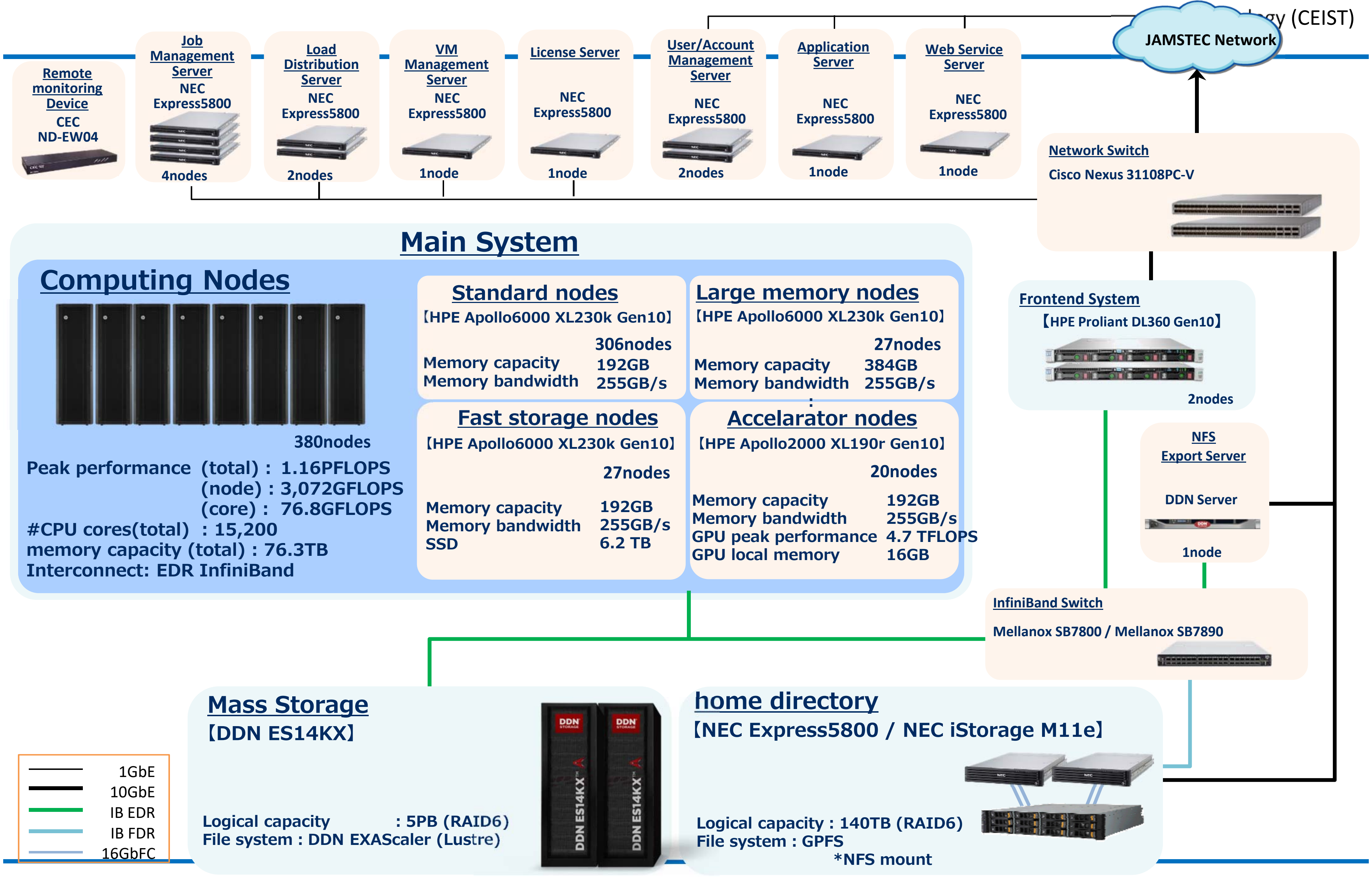
  (Minimum composition)
  - 240 nodes
  - Peak Performance 529TFLOPS (2.24TFlOPS/node)
  - Total Main Memory 46TB (192GB/node)
  - Global Storage 5PB
- Special Nodes
  - GPGPU 4.7TFLOPS/node (20 nodes)
  - Big Main Memory 384GB/node (20 nodes)
  - Big Local Storage 160GB/node (20 nodes)
- Support Virtual Machines
- Cooperation and Complementation to Earth Simulator

System Spec

		Spec	Required specifications
Model Name		HPE appllo6000	
#nodes		**380**	**240**
#CPUs		760	
#Cores		15,200	
Node	CPU	Intel Xeon (Skyake 14nm)	
	CPU Speed	2.4GHz	
	CPU #Cores	20	
	Peak Performance (Core)	76.8GFLOPS	
	Peak Performance (CPU)	1,536GFLOPS	
	Memory Capacity (Node)	192GB	
	Memory Bandwidth (Node)	255GB/sec	
Storage	Capacity (HOME)	140TB	
	Capacity (WORK)	**5,000TB**	**5,000TB**
	I/O Bandwidth (WORK)	60GB/sec	
Network	Interface	EDR Infiniband 100Gbps	
	Bandwidth	25GB/sec (bidirection)	
	Topology	Fat tree	
	Bandwidth	4,750GB/sec	
System	**CPU Peak Performance**	**1,167.36TFLOPS**	**529TFLOPS**
	GPGPU Peak Performance	94.0TFLOPS	
	Memory Capacity	**76.3TB**	**46TB**
	Power consumption	258kVA	
	#racks (Nodes)	13	

JAMSTEC
Japan Agency for Marine-Earth Science and Technology
Center for Earth Information Science
gy (CEIST)

Next “SC system” – system configuration



JAMSTEC Network

Remote monitoring Device
CEC
ND-EW04

Job Management Server
NEC
Express5800
4nodes

Load Distribution Server
NEC
Express5800
2nodes

VM Management Server
NEC
Express5800
1node

License Server
NEC
Express5800
1node

User/Account Management Server
NEC
Express5800
2nodes

Application Server
NEC
Express5800
1node

Web Service Server
NEC
Express5800
1node

Network Switch
Cisco Nexus 31108PC-V

Main System

Computing Nodes

380nodes

Peak performance (total) : 1.16PFLOPS
(node) : 3,072GFLOPS
(core) : 76.8GFLOPS
#CPU cores(total) : 15,200
memory capacity (total) : 76.3TB
Interconnect: EDR InfiniBand

Standard nodes

【HPE Apollo6000 XL230k Gen10】

	306nodes
Memory capacity	192GB
Memory bandwidth	255GB/s

Large memory nodes

【HPE Apollo6000 XL230k Gen10】

	27nodes
Memory capacity	384GB
Memory bandwidth	255GB/s

Fast storage nodes

【HPE Apollo6000 XL230k Gen10】

	27nodes
Memory capacity	192GB
Memory bandwidth	255GB/s
SSD	6.2 TB

Accelarator nodes

【HPE Apollo2000 XL190r Gen10】

	20nodes
Memory capacity	192GB
Memory bandwidth	255GB/s
GPU peak performance	4.7 TFLOPS
GPU local memory	16GB

Frontend System
【HPE Proliant DL360 Gen10】
2nodes

NFS Export Server
DDN Server
1node

InfiniBand Switch
Mellanox SB7800 / Mellanox SB7890

Mass Storage

【DDN ES14KX】

Logical capacity : 5PB (RAID6)
File system : DDN EXAScaler (Lustre)

home directory

【NEC Express5800 / NEC iStorage M11e】

Logical capacity : 140TB (RAID6)
File system : GPFS
*NFS mount

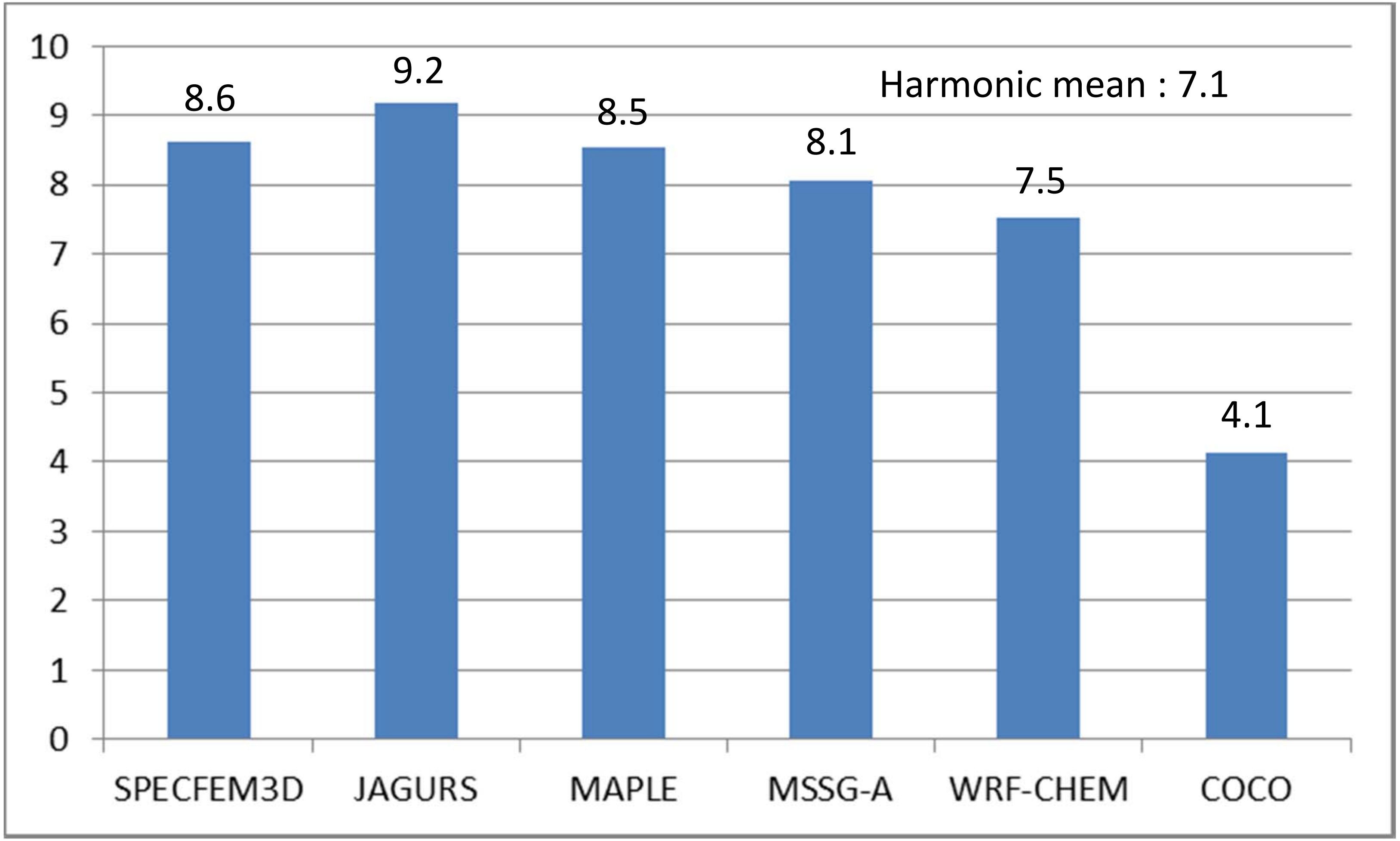
─	1GbE
━	10GbE
━	IB EDR
━	IB FDR
━	16GbFC

Application Benchmarks

Program Name	Details
SPECFEM3D	SPECFEM3D Cartesian simulates acoustic (fluid), elastic (solid), coupled acoustic/elastic, poroelastic or seismic wave propagation in any type of conforming mesh of hexahedra. [Community code]
JAGURS	JAGURS is a tsunami simulation code solving linear/nonlinear long-wave/Boussinesq equations with/without effects of elastic deformation of the Earth due to tsunami load and vertical profile of seawater density stratification. [Community code]
MAPLE	MAPLE (Metabolic And Physiological potentiaL Evaluator) is an automatic system for mapping genes in an individual genome and metagenome to the functional module and for calculating the module completion ratio (MCR) in each functional module defined by Kyoto Encyclopedia of Genes and Genomes (KEGG). [Community code]
MSSG-A	A non-hydrostatic atmospheric general circulation model, a marine general circulation model that can correspond to each of non-hydrostatic and hydrostatic, and a new bonded model combining land and sea ice models [Original Code]
WRF-CHEM	Weather Research and Forecasting model coupled to Chemistry [Community code]
COCO	COCO is the ocean general circulation model developed jointly by AORI ocean modeling group and JAMSTEC RIGC Advanced Ocean Modeling Research Team. It's also the oceanic part of the coupled general circulation model MIROC. [Original code]

Benchmark Result

Throughput performance
(magnification with current SC as 1)



Really? “x 7.1 Speed up”

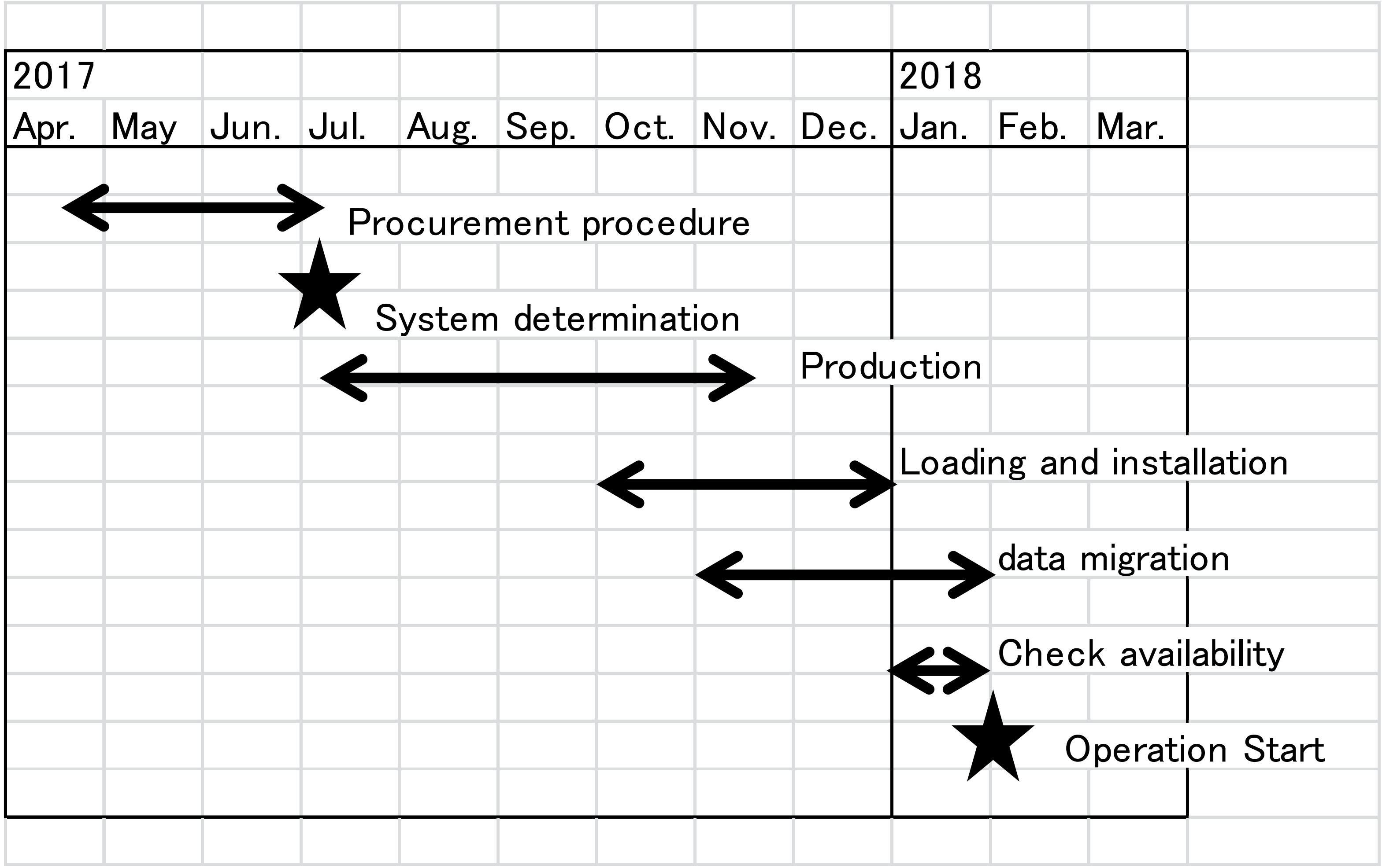
	Next SC system	Ratio with ES=1	Ratio with Current SC(ICE-X）=1
Theoretical peak performance	1.26PFLOPS	×0.96	×8.75
Total memory bandwidth	92.8TiByte/sec	×0.07	×2.1
Total memory capacity	76.3TB	×0.24	×2.8
Total #transistor	5.5 trillion	×0.53	×6.3

- CPU “Skylake” effective performance
  - Peak Performance: 20core, full VFMADD operations
  - Is the clock going down due to thermal problems?
  - How is the ratio of peak performance and effective performance in Linpack?

- How to use special node
  - The basic node configuration is uniform.
  - Adjustment between individual use of special node (GPGPU, large-scale memory, high-speed local storage) and large-scale job pool operation

- User of the current vector (SX-9)
  - Conversion to ES
  - User management, resource allocation

Schedule



2017
2018
Apr.
May
Jun.
Jul.
Aug.
Sep.
Oct.
Nov.
Dec.
Jan.
Feb.
Mar.
Procurement procedure
System determination
Production
Loading and installation
data migration
Check availability
Operation Start

Check availability

- It is inspected whether the system can be stably operated for a long time.
- We conduct two kinds of "continuous operation" inspection and "availability" inspection.
- It is necessary to pass inspections by the day before the start date of performance. (1st Feb.)
- It is possible to redo the test many times.

- Continuous operation: For 168 hours (7 days), more than 50% of nodes should not be available.
- Availability: For 168 hours, node time of 90% or more is available.

- During the test, benchmark execution and actual user test execution are performed.
- Test not only compute nodes but also batch systems and storage systems are available.

Thank you for your attention.