Okinawa Institute of Science and Technology Graduate University
5th Generation HPC Cluster

Contents

1. Background and Scope

OIST is seeking proposals for its next generation general purpose computing cluster system to provide centralized computing resources for OIST scientific research. This general-purpose computing cluster will succeed and take over the existing "Deigo" computing cluster and will be deployed in the HPC-1 room of the OIST new data center called “DC2”. This document gives OIST minimum system requirements to be considered in vendor proposals.

The proposal will include the interconnect and ethernet networks, together with the high-performance storage, the liquid cooled HPC computing system, and the HPC facility cooling for the liquid cooling. The HPC-1 room and the space for the HPC facility cooling are provided empty without access floor (bare slab) or piping, etc., therefore the proposal will include all the required additional construction and components for the implementation of the system.

The successful bid will be determined via a competitive proposal tender, for which eligibility and evaluation criteria are described in the next sections.

Each proposal shall include the total cost for the hardware, delivery and physical installation plan, OS installation, configuration, license costs, staff training, and hardware support.

2. Eligibility Criteria

The vendor must have prior experience in HPC and storage deployments in Japan at the scale of this system or larger.

The vendor must have at least three engineers having experience in installing, maintaining and supporting HPC systems with more than 1000 nodes in Japan. These engineers must be regular employees of the system vendor and cannot be outsourced. Moreover, they will be involved in the design, implementation, operation, and support of the delivered HPC system.

The vendor shall have prior experience of delivering maintenance and support in Japan.

OIST will evaluate the entire proposition using the criteria in the section below.

3. Evaluation Criteria

Together with the scoring and evaluation criteria in Appendix 12-1, proposals will be evaluated against each requirement in this specification.

4. The matters listed in the proposal

The following must be provided as part of the response to the tender submissions (documentation clearness, format and completeness are taken into consideration during the evaluation):

- Evidence for eligibility
    - Relevant experience and demonstrated ability to design, deliver and support HPC system (computing and storage) in Japan.
    - At least three names of engineers having experience installing, maintaining and supporting HPC systems with a total of more than 1000 nodes in Japan.
- A complete set of quotes that include the unit cost for each item in the system
    - A quote must be provided for the whole system.

    - Unit cost shall be at offer price (not list price)
    - Warranty, maintenance and support for at least 5 years (included in the total price, not separated).

- Evidence that the storage system can fulfill the requirements detailed in the specification
- Provide expected (best estimation) SPECfp/SPECfp_rate and SPECint/SPECint_rate performance values (using version 2017) for each type of CPU present in the compute nodes 8-1 and 8-2 of the proposed system, considering the BIOS setting in 9-1.
- Detailed faceplate and an estimated (at peak performance) maximum power consumption of the system. Total (estimated) power consumption at peak performance should not exceed the 1150kW (1100kW commercial + 50kW UPS) power capacity limit.
- A basic acceptance testing procedure for the system that includes a stress check, and the (best effort) estimated performance values for the PUE, WUE (also CUE, ERF) and peak FLOPS/W of the system under the stress check (HPL).
- Evidence of having provided support response time within the next two business days for customers in Japan (from first contact to problem resolution including part replacement lead-time).
- Detailed flowchart of the support workflow and maintenance structure which includes manufacturers, subcontractors, and all other involved parties.
- Detailed proposal of the HPC system and the HPC facility cooling solution that will be provided which includes all schematics, and all the components' specifications (when custom made) or catalog information
- Detailed plan and schedule for delivery, acceptance test, phased deliverables and deliverable prices

5. Warranty, Maintenance and Support

All systems must be covered by a 5 years warranty starting from the end of acceptance that can optionally be extended, on OIST request. The warranty should cover everything required for the normal operation of the whole system, including inspection, onsite maintenance, RMA, cleaning, etc.

The vendor must provide a minimum spare part stock, to be included in the proposal, for the items in the following table. Below is an example of an expected list of parts; the proposal must provide their recommended list.

Component	**Minimum number of spare parts per component type [The vendor proposed number ## must be, when possible, based on annualized failure rate (AFR) or MTBF, but can not be zero. The component list can also be adjusted according to the proposal composition]**
Blade (if applicable))	##
Memory (if not included in blades)	##

SSD (if not included in blades)	##
NVMe (if not included in blades)	##
Power supply	## (for each type, including switches)
PCI card (if not included in blades)	## (of each type)
InfiniBand HCA (if not integrated or not included in blades)	##
NIC card (if not integrated or not included in blades)	##
HPC facility cooling items: pump, CDUs, fan, etc.	##

Server blades and other (hot-)swappable devices must be fully replaceable over any failure, and the faulty device must be sent back for troubleshooting (RMA). The vendor can keep replacement spare parts on the OIST site.

For all other failed hardware components, the vendor must provide the necessary support (spares, periodic inspection, onsite repair, preventive maintenance, etc.) to maintain the system in proper operation.

Technical support in either English or Japanese (whichever is available) must be available by telephone and by email during business hours (weekdays, 9:00-17:00 minimum core time, Japan local time, excluding national holidays and end of year holidays).

Maintenance and support for the storage for this new system should be at least equivalent (quality, response time, implementation, etc.) to the maintenance and support provided to actual OIST tiered storage systems (See the attached document “OIST_Storage Maintenance_Spec_example.docx” for an example maintenance and support specification).

6. Network

The HPC system will have four different networks, for which all the switch components required to implement the four networks will be included in the system. The campus network, the management network, and the service network will be ethernet networks. The fourth network will be a high bandwidth private interconnect network. The network of the HPC system should stay operational even in the event of a loss of connectivity with the OIST network. Fig 1 shows the campus network topology and the expected campus network part for the HPC system, called HPC-1. Fig 2 shows the management and service networks for HPC-1, and Fig 3 shows an example of the network topology to be used in the HPC-1 management and HPC-1 service network.
All the networking components must be fully operational without the need for an Internet or cloud connection, or subscription license, and have at least 6 years of support lifetime from the delivery date. Moreover, all the switches must have redundant hot swappable power supply and FAN units.

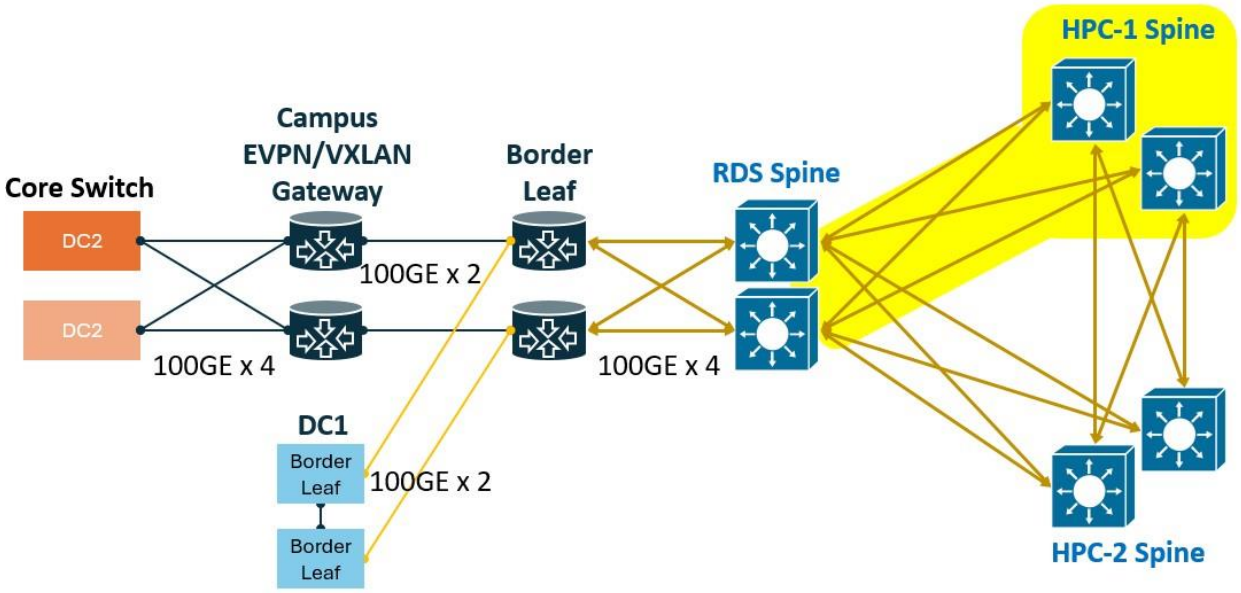


Fig. 1 OIST DC2 HPC-1 Campus network layout. All the switches up to the RDS spine are located in the RDS room and provided by OIST. The HPC-1 spine and leaf switches for the new HPC system (yellow background, including cables) will be provided in this proposal and locates in the HPC-1 room

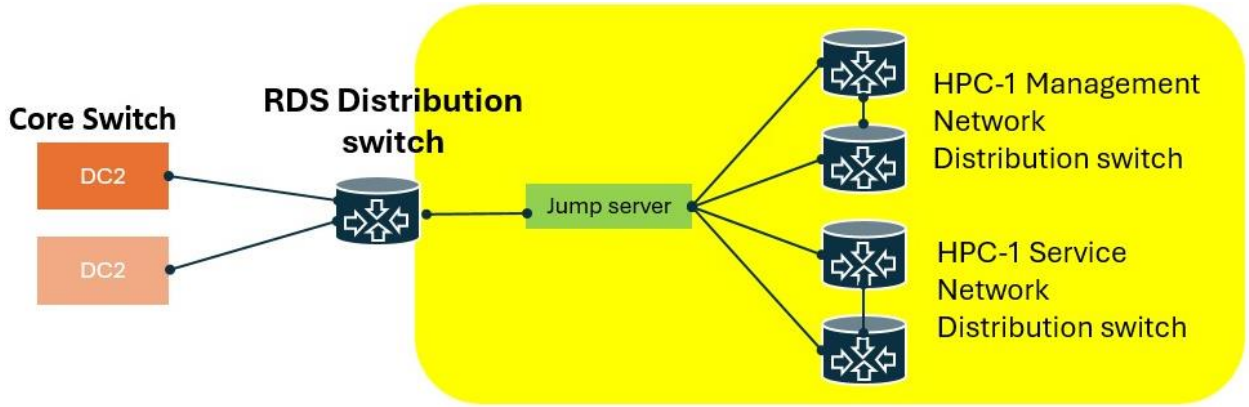


Fig. 2 OIST DC2 HPC-1 management and service networks. The core and RDS distribution switches are located in the RDS room and provided by OIST. The proposal for HPC-1 room

will include the "jump server", the HPC-1 management and service distribution switches and all the required cables.

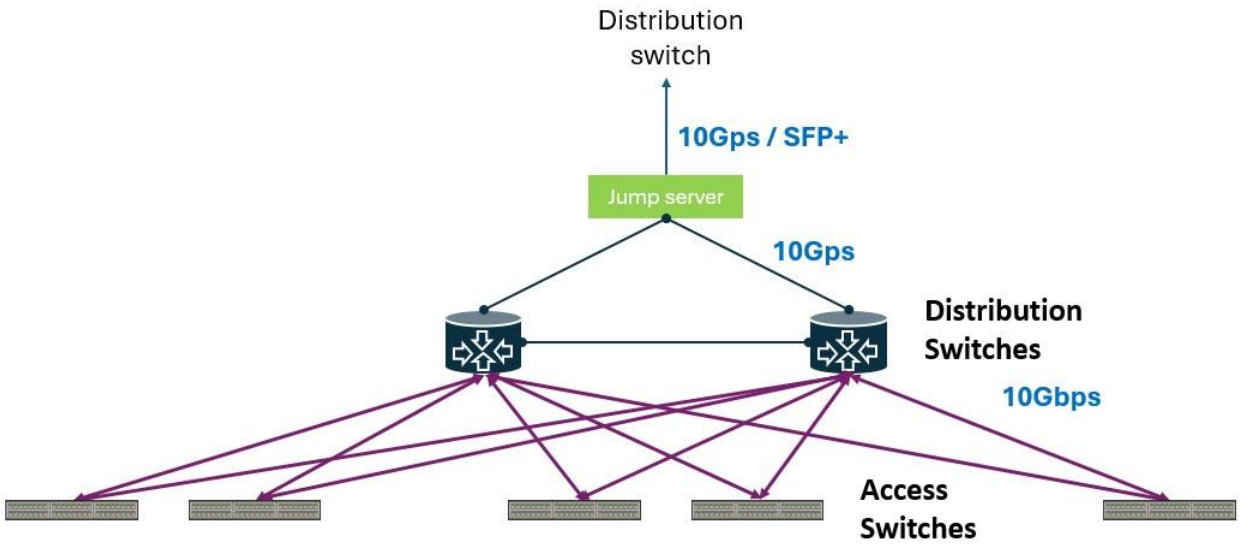


Fig. 3 Example of management or service network topology

6-1. Ethernet network

The campus network will provide 25Gb/s network connection to the servers through its leaf switches, which will have 2 x 100Gb/s uplinks to the HPC-1 spine switches, each HPC-1 spine switch having 3 x 400Gb/s uplinks to each of the RDS spine switches. The management network and the service network are separate private (minimum 1Gb/s) networks connected to the BMC and ethernet (for example, embedded NIC) interfaces of all servers of the HPC system, respectively. As shown in Fig 3, for both management and service networks, the topology will consist of a layer of distribution switches in HA configuration and a layer of access switches having uplinks to two different distribution switches, all through 10Gb/s connections. The "jump server" will be used to connect the RDS distribution switches to the HPC-1 management and service distribution switches and therefore must have at least three dual 10Gb/s SFP NICs (for LACP), of which one dual 10Gb/s can be embedded (quadruple 10Gb/s SFP single NICs are prohibited).

The campus network is used to connect to the OIST network, which consists of all the labs, server rooms, and DC1 (first data center in Lab1). The service network is used for OS installation (PXE booting), OS management and for running services such as private DHCP and DNS, pacemaker, SLURM etc. The management network is for OOB and mainly used for monitoring and managing all servers and components BMCs or management port (including monitored power distribution unit), all of which shall be accessible to this network.

All HPC-1 spine and leaf switches must be compatible with and be able to run SONiC OS. Moreover, the HPC-1 spine A switches must support MLAG, EVPN/VXLAN multi-homing, and be SONiC EVPN/VXLAN compatible.

6-2. Interconnected network

All system nodes, including the high-performance parallel storage, must be interconnected with a minimum bandwidth of 800Gb/s available between nodes. The interconnected network will support all the data traffic required during HPC computation and must be RDMA capable. If the interconnect network is implemented using Ethernet switches, there should be no more than 3:1 oversubscription between nodes and have very high path diversity to maximize the bisection efficiency. If the interconnect network is implemented using Infiniband technology, the entire network must be symmetric, with no more than 50% blocking. A minimum of 32 free ports on the last layer of the switch tree must be available for additional devices to be connected. Fiber-based cables with high modal bandwidth will be preferred over copper cables for the interconnected network.

6-3. Jump server

A jump server will be used in each of the management and service networks. The jump server will be equipped with single- or dual-socket 64-bit x86 compatible processor(s) having a minimum of 32-core per socket, and at least 192 GiB of memory per node. Local storage will consist of at least 2TB SSD disks with RAID1 configuration. A server will have at least three dual 10Gb/s SFP NICs (for LACP), of which one dual 10Gb/s can be embedded (quadruple 10Gb/s SFP single NICs are prohibited). Each jump server will have 1Gb/s NICs connected to their respective distribution switch management port, and USB-serial connection to their respective distribution switch serial console port.

6-4. Cables, Cabling and Labeling work

All cables (including all SFP/QSFP transceivers on both ends) required to connect all the system components must be included in the proposal. This includes the cables and transceivers at both ends used to connect the proposed system to the RDS. The minimum grade of the cables must be OM4 for fiber and category 6A for UTP.

The cabling design must be planned and documented before installation work starts, and reviewed by OIST with the following requirements aligned too:

- Cabling must be neat and non-obstructive. Meaning cables should run down the side of racks rather than the middle and not restrict the install or removal of existing or additional infrastructure.
- The right length of cable should be used. Having some slack is ideal but if there is more than 50cm, then a more appropriate length cable should be utilized.
- All cabling between must be accessible and removable without having to stop operation of unrelated systems.
- OOB cabling should be within the same rack/enclosure when possible.
- All cables must be clearly labeled, which identifies what it is connected too (standard will be decided at kick-off meeting)
- Bundle cables together in groups of relevance. When bundling or securing cables, use Velcro-based ties. (every 30-60cm). Do not tighten the bundled cables.
- Cabling to OIST switches in RDS room must use the rack tray built between the rooms
- Interconnect network cabling must be as much as possible inside the rack/enclosure or over the rack/enclosure.

6-5. Connection to OIST switches in RDS room

For the connection between RDS and HPC-1 spine switches: the RDS Spine A is Dell Z9432F-ON running Enterprise SONiC 4.5.0a and the RDS Spine B is Arista 7060DX4-32 running SONiC 202411.949484-3435ddad9. The transceiver form factor is QSFP-DD (not OSFP) and the connection

standard (currently in use) is 400GBASE-SR4.2 with MPO-12 MMF OM4 cables. 3 x 400GE links per RDS/HPC-1 spine pair is required for a total of 12 x 400Gbps links between RDS and HPC-1 spines. The vendor will also provide 6 x QSFP-DD 400GBASE-SR4.2 transceivers for both RDS Spine A & B (Dell & Arista), 12 transceivers for HPC-1 Spines, and 12 x MPO-12 MMF cables, together with the cabling work.

For the connection between the RDS distribution switches and the jump servers: the RDS Distribution switches are Arista 7050SX3-24YC4C-S and the transceiver form factor is SFP28 (supports 25Gbps/10Gbps/1Gbps) and must be Arista-branded transceivers. The connection standard is 10GBASE-SR with Duplex LC MMF OM4 cables, with Arista 25GBASE-MR-SR which supports both 10Gbps and 25Gbps (multi-rate). Each jump server is connected to two distribution switches, for a total of 4 links.

Cable length between RDS and HPC-1 switches is estimated to be between 30 to 50 meters. The vendor should prepare the correct cable lengths and provide enough slack in case relocation within the HPC-1 room is needed.

7. High Performance Storage

The high-performance storage will consist of an ultra-fast parallel storage. All storage server components must have redundant high efficiency power supplies, and all hardware components must have the latest working firmware.

The proposed high-performance storage must provide the identical user experience, operational support effectiveness, monitoring infrastructure (email notifications, full access to SNMP and controller CLI, availability of source code for all Linux kernel components, etc.) and support service level as the high-performance storage in previous generations of HPC systems operated at OIST.

7-1. Ultra-fast parallel computing private storage

Should have at least 1PB of capacity using flash technology (NVMe, SSD, etc.) and equivalent RAID6 redundancy.

Read and write performance should be at least 900 GB/s and 700 GB/s, respectively, while providing at least 10 million 4k random read IOPS. The storage should also provide at least 20 billion inodes.

This storage should be accessible from the high-core-count, login, transfer, scheduler, and management nodes and only on the interconnect network. The storage must implement a distributed or parallel file system and feature user, group and project (or directory) quotas. The filesystem will be mounted on all the nodes through "/flash".

The system must demonstrate POSIX shared directory file creation rates above 80K and unique directory file creation rates above 550K.
To leverage local SSDs on compute nodes, the storage system should provide a persistent caching capability that transparently caches frequently accessed data on local storage. The cached data selection should be configurable by UID, project ID, and filename.
The proposed High Performance Storage system must provide seamless access and interoperability with the existing Research Data Storage system so that both systems can be used efficiently by users.
The system should provide an ultra-fast scanning capability that allows rapid scans of the overall file system for file system accounting and the implementation of purge policies. The ultra-fast scanning capability should allow scanning of 50M files at 100 second speed.

The filesystem should not require any licenses for operation, update or upgrade, and the source code of any necessary client software/kernel-module should be available for all required versions.

The proposed storage performance will also be subject to evaluation, and preference will be given to higher performance indicators (Read/write speed, file/directory creation rates, etc.).

7-2. Monitoring server

A monitoring server must be available together with a system allowing continuous monitoring of the storage activity (user and SLURM job-based IO and metadata activity monitoring, with logged metadata operations linked to user and to job IDs). Dashboard, graph customization and a remote API must be available through a WebUI (and optionally a CLI). The proposal should include the most suitable specifications for this server.

8. Cluster node composition

All nodes must have at least 100GB of storage for the OS and OS related data and a local /scratch partition with at least as much storage space as the node memory. For example, a 256GiB of memory node must have at least 275GB (1GB ≈ 0.931GiB) of storage space available in its /scratch partition.

All nodes should be connected to the four networks described in the network section. Connection to the network should be single port and not make use of LACP. Moreover, all nodes must be able to mount the research data storage in accordance with the Lustre nodemap rules.

Each node should have an optimal RAM (memory) configuration using memory modules with the highest possible frequency.

The OOB (out of band) management component of each server must provide the following secure capabilities (not provided by unsecure IPMI) to protect the server from inadvertent or malicious changes: RoT (Root of Trust) authentication must be available for BIOS and firmware during the server boot process, with the ability to lock the configuration and prevent firmware updates.

All nodes must have redundant high efficiency power supplies, and all the hardware components must have the latest working firmware.

The high core count nodes must have a compact configuration so that at least 2 nodes can fit in one rack unit (U), which can be RU or OU depending on whether 19-inch or 21-inch rack is considered. If enclosures are used, the enclosure node (N) density cannot be greater than 0.85 U/N.

The nodes shall be liquid cooled, with efficient non-leaking solutions. Higher preference will be given to solutions having the following specifications:

- DO NOT require any manipulation of the liquid cooling manifold or the piping connectors during the maintenance of a node
- DO NOT require manipulation or manual disconnection of PSU or power-cord during the maintenance of a node
- DO NOT feature fans or combinations of heatsink plus fan
- have the maximum number of components directly liquid cooled (e.g. > 85%).

Preference will be given to OCP-ORv3 rack, close-loop liquid-cooled enclosures, or equivalent solutions that optimize the liquid-cooling loop and the operation of the blades/nodes.

The proposed HPC system should also be able to monitor the total power consumption and provide monitored information, either through API, CLI, or SNMP.

Training on the operation, servicing and parts replacement of the node servers will be provided to OIST staff.

8-1. High core count nodes

Each high core count node will have a minimum of two 64-bit capable processors (dual-socket), x86 compatible, minimum 256-core per processor, maximum TDP of 600W, and at least 2TiB of memory (minimum 4GiB/core), with AVX-512 support. The motherboard of those nodes should be able to support a minimum of 8000 MT/s speed memory. A single SSD/NVMe can be used for local storage. It is possible to have the high core count nodes not directly connected to the campus network, but in

that case the vendor must propose “indirect connectivity” alternatives that can be implemented and used in operation at no cost.

8-2. Login nodes

Eight nodes with single- or dual-socket 64-bit x86 compatible processor having minimum 128-core per node, and 256 GiB of memory per node. Local storage will consist of SSD disks with RAID1 configuration.

8-3. Data transfer nodes

Four nodes with single- or dual-socket 64-bit x86 compatible processor having minimum 128-core per node, and 512 GiB of memory per node. Local storage can be a single SSD. The transfer nodes will be mainly used for:

- Fast data movement between the “ultra-fast parallel computing private storage” and the campus storage
- Fast data movement between OIST Research data storage (located in the RDS room) and the proposed HPC system

Each data transfer node will have a 100 Gb/s connection to the HPC-1 spine switches, a direct connection to the “interconnect network”, and NDR connection to Research Data Storage (through native Lustre FS). All cables necessary for connecting the data transfer nodes to the Research Data Storage must be provided in this proposal.

8-4. Management server

One server with single- or dual-socket 64-bit x86 compatible processors having a minimum of 32 core per socket, and 384 GiB of memory. Local storage will consist of SSD disks with RAID1 configuration.

In addition to the integrated BMC port connected to the management network, another 1Gb/s port is connected to the service network, and the server being connected to the interconnect network, the management server must also have an extra 1Gb/s Ethernet port available (for example, onboard 1G Ethernet port) to be connected to the management network, for accessing the OOB network and other system BMC from the OS of the management server.

8-5. Scheduling servers

Two servers with dual-socket 64-bit x86 compatible processors having a minimum of 32-core per socket with a frequency higher than 2.21 GHz, and 512 GiB of memory per node. Local storage will consist of at least 3TB SSD disks with RAID1 configuration.

9. Operating System and node configuration

All nodes and servers should be able to run Linux (Rocky Linux or equivalent open-source non-licensed based Linux operating system, ex: RL 9.8 or higher, Ubuntu 26.04 or higher). SLURM should be installed as the cluster scheduler on the scheduling nodes, with high availability configuration of SLURM controller daemons. The OS should also be able to support the compilation of third-party Lustre kernel modules for connection to the Research data storage.

OIST SCDA staff should be able to fully rebuild any compute node with newer versions of Linux and SLURM, and for this purpose, the vendor should provide OIST with all the required drivers, firmware, and tools and configuration files required for the rebuild.

9-1. BIOS configuration

For the compute nodes:

- Must be capable of PXE network boot and be configured to first try PXE and then hard disk at boot time.
- Should be configured to remain powered off in the event of power loss and return.
- Must have hyper-threading feature OFF.
- Must not have BIOS settings configured for overclocking.
- Other settings when applicable read as follows
    - Intel Turbo boost: Off
    - AMD performance boost: Off (if it increases the TDP), On (otherwise)
    - cTDP: Nominal

All BMC (baseboard management controller) of each server connected to the OOB (out of band) network must have administrator password setup to a value different from the default one (it can be the same password for all servers). The password will be provided to OIST during the acceptance phase.

10. HPC Facility Cooling

The proposal shall include cooling infrastructure required for the full operation of the proposed HPC system with piping and electrical solutions designed with expansion capability.

10-1. HPC facility cooling infrastructure

The proposal shall include the implementation of the HPC facility cooling and provide all the necessary components, for example chillers, cooling towers, CDUs, piping, additional pumps, and any other required or selected components to implement the HPC facility cooling for the whole system. The vendor will propose a solution for the HPC facility cooling. Liquid cooling proposals with efficient PUE and high temperature water usage will be given higher consideration in the evaluation. Power Usage Effectiveness (PUE) will be used for the evaluation, however whenever possible, the proposed cooling infrastructure must provide numbers on the following

- Water Usage Effectiveness (WUE in L/kWh), Carbon Usage Effectiveness (CUE in $kgCO_2e/kWh$), Energy Reuse Factor (ERF)
- Long term maintenance and utility costs
- HPC system performance
- Expandability (for example for HPC-2) and sustainability of the cooling infrastructure

Minimum specifications for the HPC cooling components, when included in the vendor proposed solution, such as CDU, Chillers, Cooling Towers, and Rear-doors, read as follows.

CDU

- In-Row CDU or in-rack (rack-mount) CDU
- Redundant design
    - in-rack: (all hot-swappable) fan 1+1, pump 1+1, power 1+1
    - in-row N+1 redundancy
- High availability (wide range of cooling capacity, pressure drop adjustable) and high cleanness (hygiene pipe with fine filtration)
- Sensors and controllers monitor the states of all components and all the measurables (temperature, flow, etc.) with the availability of SNMP, Modbus, or TCP protocol, with RJ45 connectivity available.

Cooling Tower

- Capacity: HPC system peak power usage + buffer
- Redundant design
- Sensors and controllers monitor the states of all components and all the measurables (temperature, flow, etc.) with the availability of SNMP, Modbus, or TCP protocol, with RJ45 connectivity available.

Chiller

- Capacity: HPC system peak power usage + buffer

- Redundant design: water pump 1+1,
- Sensors and controllers monitor the states of all components and all the measurables (temperature, flow, etc.) with the availability of SNMP, Modbus, or TCP protocol, with RJ45 connectivity available.

In case of the use of RDHx (rear door heat exchanger) in the solution, the proposal should include numerical results of thermal performance evaluation through CFD simulation (or analysis) and psychrometric chart analysis.

In case of use of any other cooling solution not included in the above, redundant design, monitoring capability according to requirements, and long-term warranty are required.

All the monitoring ports and BMC of the HPC cooling facility components must be connected to the management switch with their IP addresses configured.

Training on the operation, servicing and parts replacement of the HPC facility cooling components will be provided to OIST staff.

11. Physical Installation and acceptance

A spreadsheet will be used to track any issues or changes happening during physical installation and acceptance, and a daily report will be provided to OIST using this spreadsheet.

11-1. Physical installation

The vendor should bring all additional tools, power meters, parts, etc. required for installation and acceptance. Dedicated clean clothing must be used inside the OIST data center, and the vendor personnels must wear the required protection for the work. All servers and components will be unpacked in the preparation room, and no card boxes, plastic bags, bubble wraps, polystyrene foams, paper sheets, etc. shall be introduced inside the data center rooms.

A kick-off meeting will be held prior to the installation to discuss final technical aspects of the delivery, such as the final layout of the system components (storage, servers, switches), etc.

11-2. Acceptance

The acceptance will consist of the verification of the entire requirement and the proposed items agreed upon during the proposal evaluation. Moreover, the following checks must be cleared for the acceptance of the deliverables.

- OS and SLURM scheduler (slurmctld in HA configuration) installation.
- Operation verification of all the nodes and stress check using the following software (the system should remain stable over 3 days of continuous run):
    - OSU: http://www.nersc.gov/users/computational-systems/cori/nersc-8-procurement/trinity-nersc-8-rfp/nersc-8-trinity-benchmarks/omb-mpi-tests/
    - HPL: http://www.netlib.org/benchmark/hpl/
        - HPL stress tests should make use of the most advanced available SIMD instructions and optimized DGEMM implementations using MKL, BLIS, or OpenBLAS, or etc., when available.
- Storage performance evaluation
- The vendor will conduct power consumption measurements of the full system during the stress check.
- Show that measured SPECfp/SPECfp rate and SPECint/SPECint rate performance, for the BIOS setting in 9-1, on the compute nodes in 8-1 and 8-2, meet expected values (equal or better)

As part of the acceptance, the following system details must be provided to OIST (as Excel sheets):

- Serial numbers (or service TAGs) and part numbers of all system hardware and components
- List of all firmware versions for each hardware part (including firmware release date).
- All configuration file and settings used for the installation of the OS and SLURM and for the operation verification
- MAC addresses of all NIC devices together with their node/server names (before the delivery, OIST will provide the name of the system that will be used for the nodes/servers naming)
- All configuration parameters used during the switches (Ethernet, IB, and Management) installation

- Operation manual for the storage that include emergency safe shutdown procedure

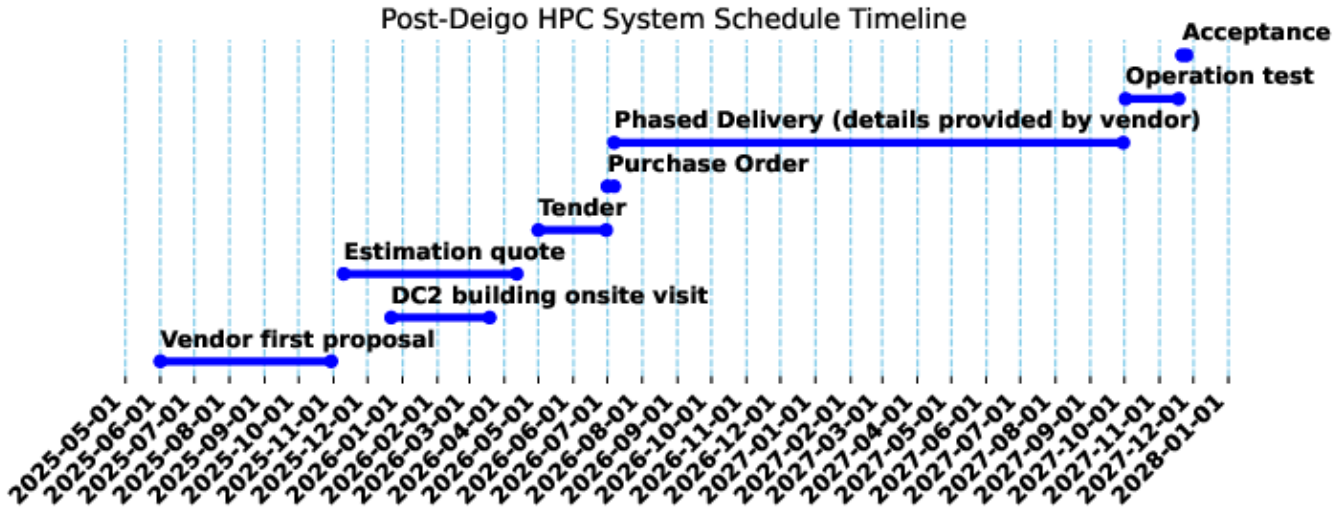
The vendor will provide a report for each verification. In the event of failure to clear a verification, the vendor should provide a remedy at no cost.

11-3. Delivery and schedule

Tentatively, the system must be delivered and accepted by the end of 2027.
The place is OIST data center DC2, at the address below:

Okinawa Institute of Science and Technology Graduate School
1919-1, Tancha, Onna-son, Kumigami-gun, Okinawa 904-0412, Japan

The project schedule timeline is provided below (dates are indicative). ***The detailed schedule timeline for delivery (from "purchase order" to "Acceptance") will be provided in the vendor proposal and finalized after the bidding, during the contract establishment.***



Action	Start	End
Vendor first proposal	2025-6-1	2025-10-30
DC2 building onsite visit	2025-12-22	2026-3-19
Estimation quote	2025-11-10	2026-4-12
Proposal competition (Budget amount will be announced at opening)	2026-5-22	2026-6-29
Purchase Order	2026-7-1	2026-7-7
Phased Delivery (details provided by vendor)	2026-7-7	2027-9-30
Operation test	2027-10-2	2027-11-18
Acceptance	2027-11-21	2027-11-30

12. Evaluation Sheet

Detailed information on OIST building facility, data center and actual computing system will be available during the public announcement period of the tender (Q&A).

12-1. Proposal evaluation details

The vendor proposal will be evaluated accordingly to the following criteria and the required document check list. Any "required document" that fails the check will lead to rejection of the proposal. Accepted proposals will compete based on the criteria below.

Criteria	**Weight**
Main performance criterion of the vendor x is based on the proposed total number of cores Nx (from 8-1) in the system, the proposed power usage effectiveness PUEx of the system, and the proposed number of cores of the other competing vendors a, b, ..., according to the formula below $$80 \cdot \frac{N_x}{PUE_x \cdot \max(N_a,\ N_b, \dots, N_x,\ \dots)}$$	80
Power usage effectiveness contribution using the formula below $$5 \cdot e^{1-PUE_x}$$	5
Direct liquid cooling coverage factor $$(\%\ \text{of server/chassis components cooled by DLC}) \cdot 0.05$$	5
FLOPS per Watt Px contribution of the vendor x $$10 \cdot \frac{P_x}{\max(P_a,\ P_b,\ \dots,\ P_x,\ \dots)}$$ where, $$P_x = \frac{Peak\ FLOPS\ from\ HPL}{Peak\ Power\ usage}$$	10

Required document	**Check**
Does the proposal is comprehensive and include all the requirements stated in the specification, especially the ones in section 4?	Y/N
Have the integrators, hardware manufacturers, and vendors demonstrated evidence of edibility in deploying an HPC system at this scale?	Y / N
Has the vendor provided all the required quotes?	Y / N
Does the storage system meet or exceed all the requirements in 7?	Y / N
Does the storage monitoring solution meet the requirements in 7-2?	Y / N

Are the SPEC performance values provided by the vendor consistent?	Y / N
Does the acceptance testing procedure have all the details about the test to be performed?	Y / N
Does the proposed solution meet all the specific requirements of the specification?	Y / N
Does the proposed solution meet the OIST power requirement?	Y / N
Does the proposed full cooling solution meet the requirements in 10-1?	Y / N
Does the maintenance and support service for the storage meet OIST requirements (refer to last paragraph of section 5)?	Y / N
Do the integrator, hardware manufactures and vendors have the capabilities to provide full hardware support for the HPC cluster and system support for the storage?	Y / N
Is the provided delivery schedule, deliverables, and acceptance test plan consistent and comprehensive?	Y / N

13. DC2 Building information

Details of the DC2 building facility are given here. Documentation and drawings of the building and facilities are provided in the following supplementary documents.

【DC2】完成図 PDF 意匠構造.pdf

【DC2】機械設備完成図.pdf

【DC2】電気設備完成図.pdf

When available, additional details (CAD files, etc.) can be provided as supplementary data upon demand. Moreover, onsite visits to the OIST DC2 facility are also possible upon demand.

DC2 is an individual building dedicated to OIST scientific computing systems and storage with restricted access to only specific designated staff, and the activities monitored through CCTV cameras.

13-1. DC2 facility power specifications

The estimated peak power consumption of the proposed system should be provided and should not be higher than the 1,100kW maximum capacity available for this system.

The storage systems and essential network components (spine and distribution switches, jump server(s), scheduler and management nodes) shall be connected to the UPS source (separate from the main source) provided by OIST DC2 facility, in the listed priority order, for up to a total no higher than 50kW.

Because the power available from the existing HPC-1 distribution panel is limited to about 500kW, to use the maximum power available, the vendor will have to (but is not limited to):
a) either provide and change the breakers at the distribution panel and the busduct with higher current rated breakers, together with changing the cable to the busduct with a thicker one,
b) or provide and add distribution panels in the HPC-1, and connect them in place of the HPC-2 to the busduct
c) A third option would consist of moving the existing distribution panels from HPC-2 to HPC-1. However, this work will have to be ordered from the vendor who originally installed the distribution panels. Further discussion, not in the scope of this specification, will be needed with OIST building facility management (BFM).

Electrical power for the HPC system will be available through u-v-w-N-E connections to breakers (A-P-2, A-P-3 and UPS1-2) from three power panels (2 for commercial lines, one for UPS) located inside the HPC-1 room.

Electrical power for the outside cooling facility will be taken from the HPC-1 room by branching from the shunt switch circuits of the existing distribution panels (A-1, A-2, and UPS) to a new distribution panel. The new distribution panel (can be more than one if necessary) and branching will be provided by the vendor.

The vendor must provide meters for all the distribution panels used or added in the proposal having either SNMP, Modbus, or TCP protocol monitoring capability, connected to the management network. The PUE should be directly measurable from the meters on the distribution panels.

13-2. DC2 layout

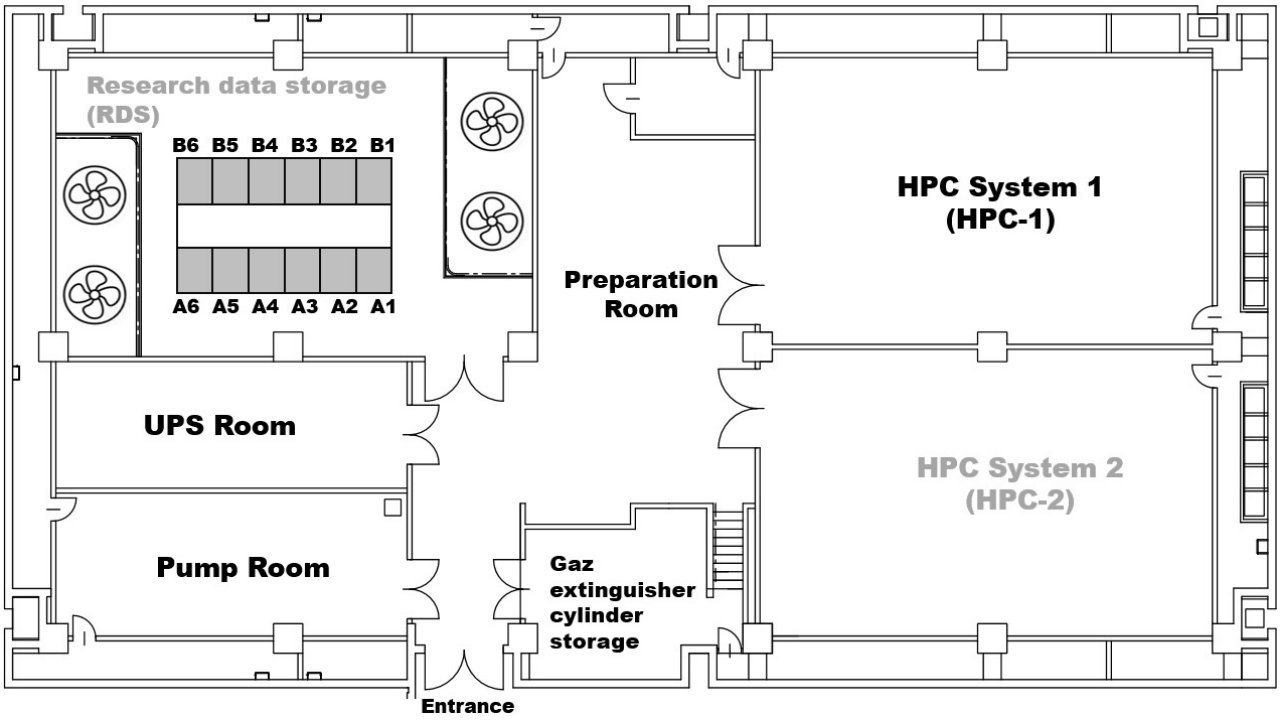


Fig. 4 Top view of the data center (DC2) on the first floor.

The HPC system 1 (HPC-1) space in OIST DC2 is available for cluster installation. All the computing components (network, storage, nodes, servers) and eventually part of the cooling system (ex: CDUs) should be located in the HPC-1 room. HPC facility cooling will be located outside, and access for pipes will be available from the building wall side of the HPC-1 room facing the outdoor space for HPC facility cooling.

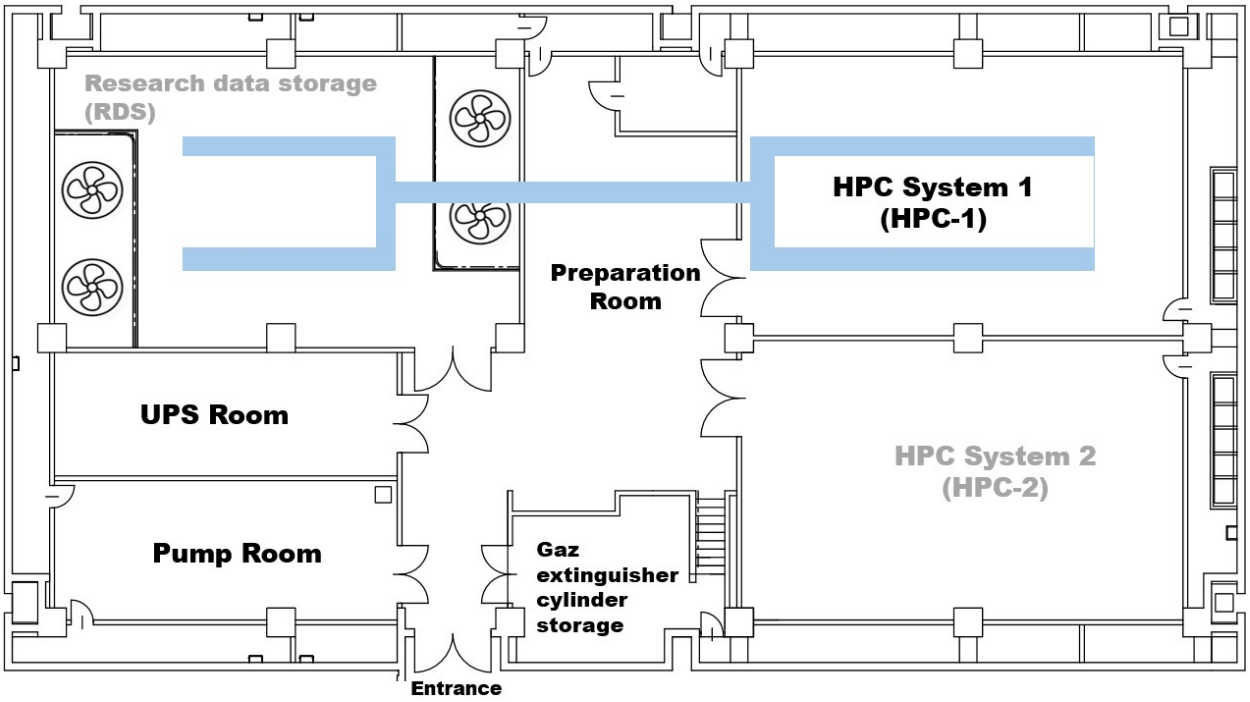


Fig. 5 Rack tray in blue for cabling between RDS and HPC-1 rooms and within HPC-1 room if needed

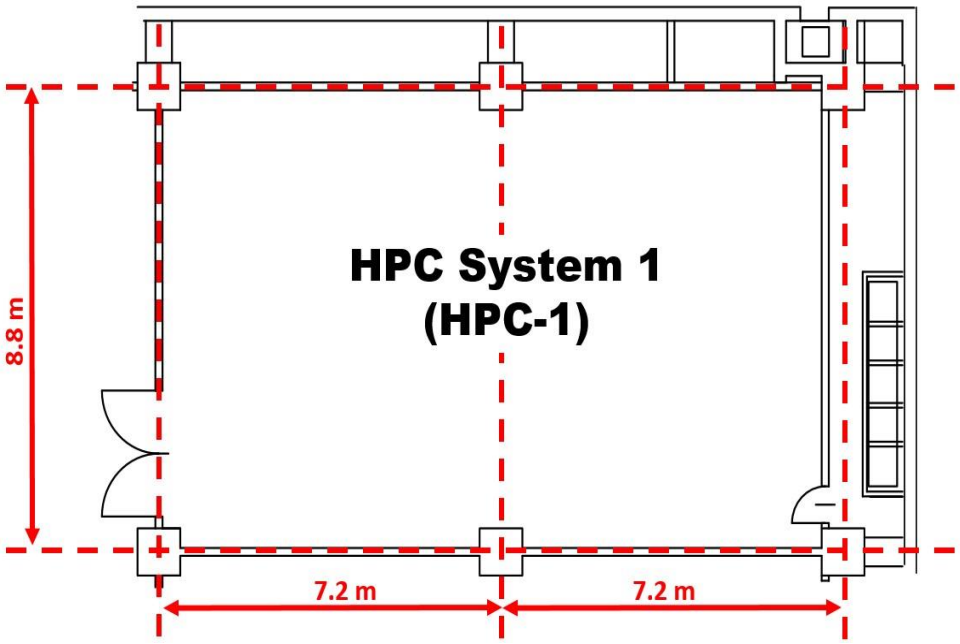


Fig. 6 HPC-1 room dimensions, there will be at least 5m of space usable between the floor and the ceiling

Fig. 7 Relative location of the room HPC-1 and the outdoor space (about 230 m^2 available) for the HPC facility cooling