SPECIFICATIONS

PNNL Battelle, Pacific Northwest Division, Management & Operating Contractor of the U.S. Department of Energy's Pacific Northwest National Laboratory (PNNL), is requesting proposals to provide the following components for a new computer cluster.

The request is for configurations of both a 32 and 64 compute node system.

The cluster's primary use will be to exploit novel networking architectures able of sustaining high throughput rates for small messages. Current technology can send around 140 million messages per second. We require the system to be delivered with a network technology that can send at least 3 times that number of messages per second between any two compute nodes.

Specifications for each compute node:

1) Dual socket Intel E5-2623 v3 CPUs

2) 128GB of memory (8 x 16GB DDR4 2133 MHz RDIMM)

3) Mellanox FDR ConnectX III Infiniband connection either built-in on the motherboard or provided as an HCA. If not built-in on the motherboard, the HCA will be attached to a PCI Express 3.0 8x slot running at full 8x speed.

4) Network capable of transferring in excess of 520 million messages per second between any two nodes.

5) At least 1 TB of local disk storage on each compute node. Each local disk(s) shall be able to sustain 80MB/s of I/O for 1MB or larger block size reads and writes as measured by the FIO benchmark tool. Preferred configuration is a single disk drive on each node. Drive must be front mounted and replaceable without removing node from rack.

6) Capable of running Redhat Enterprise Linux 6.6 or later. The cluster will be configured with Rocks 6.2 during acceptance testing.

7) At least one gigabit Ethernet interface which is set to PXE boot

8) IPMI or other management network interface with the ability to remotely power cycle the node and remotely establish a console connection.

9) BIOS and BMC (management interface) upgrades to the compute nodes must be able to be performed from either the host Linux operating system or by unattended PXE boot.

10) Power cables to connect to in-rack PDUs should not exceed 2 feet in length

11) Cat6 Ethernet cables shall be used to connect to rack level Ethernet switches

FDR Infiniband network specifications:

1) In the case of a 32 node cluster, a single 36 port FDR Infiniband switch. In the case of a 64 node cluster, 2 x 36 port switches.

2) All cables required to connect the compute nodes and the 1 head node to the Infiniband switch(s).

3) In the case of a 64 node cluster, 2 FDR Infiniband cables to connect the 36 port switches to each other

Additional Network requirements:

1) Rack level gigabit Ethernet switch(s) shall have at least 1 10 Gigabit Ethernet port capable of connecting into existing 10 Gigabit Ethernet fabric and have basic Ethernet Layer 3 management capabilities. . If IPMI network is separated from regular Ethernet network, IPMI network switches must have at least 1 Gigabit uplink port.

2) The cluster shall be equipped with 1 ancillary node to handle login and cluster management tasks. The nodes shall meet the same specifications as the compute nodes with the exception of the novel network interface which is not required.

Physical space requirements:

All components shall draw air from the front of the cabinet and exhaust out the back.

Maximum power draw for any one rack unit shall not exceed 48kVA.

Cooling in the room will be provided by standard under the floor CRAC cooling.

Vendor will supply rack(s) and Power Distribution Units (PDUs) with the rack(s).

If cluster exceeds 1 42U rack, proper cabling must be supplied to connect compute nodes between racks via existing overhead cable trays. The distance from the top of the rack to the bottom of the copper cable tray is 6 inches and the distance from the top of the rack to the bottom of the fiber tray is 11 inches. See Figure 1

Warranty requirements:

All items shall include a 3 year warranty service with a cross shipment of warranty parts.

Figure 3 Cable Trays

I. Requirements Checklist

Fill in the checklist in its entirety. Answer “yes” or “no” to each requirement. Provide supporting detail in every appropriate space. Incomplete checklists will cause a proposal to be rejected as unresponsive.

	Requirement	Requirement Satisfied?	Explain how your proposal meets the requirement

1	**Systems consists nodes containing dual E5-2623 v3 with 128GB of registered ECC 2133Mhz memory per node**		
2	**Each node has BIOS settings as appropriate with PXE boot first**		
3	**Each node local disk storage of at least 1 Terabyte**		
4	**Each node has Mellanox ConnectX III FDR on-board or via HCA in PCI Express 3.0 8x slot running at full 8x speed**		
5	**Network able to send 8 Byte packets between nodes at 4 Gigabytes/second**		
5	**Each node has IPMI Interface with DHCP assignable address**		

6	**Warranty service with next day part replacement**		
7	**Appropriate Ethernet and IPMI network switches and cables**		
8	**Appropriate cabling to connect compute nodes to Infiniband switch**		

II. Acceptance Testing Criteria

Acceptance testing will be performed by Battelle as follows:

1. Physical verification of equipment ordered.
    a. Verify quantities & specifications match invoice.
    b. Verify equipment is of high quality: Build tolerances reasonable, hot swappable parts slide in and out easily, etc.
2. Bring system online.
    a. Verify CPU speed
    b. Verify RAM total and speed
    c. Verify BIOS settings
    d. Verify PXE enabled Ethernet interface
    e. Verify IPMI interface exists, is enabled, and receives IP address by DHCP

3. Performance & reliability testing.
    a. Install system and run "CPU burn" for 8 hours
    b. Boot nodes with memtest86 via PXE and run for 8 hours
4. Ethernet/IPMI testing.
    a. Verify ability to PXE boot nodes and perform BIOS upgrades via PXE
5. FDR network
    a. Verify expected bandwidth and latency between computed nodes via OSU benchmarks
6. Novel networking

a. Verify 520 million messages per second can be transferred between any two compute nodes.