Blade-Based High-Performance Computing Cluster System

Tender Notice/Request for Proposal

by

Centre for Modeling and Simulation
University of Pune, Pune 411 007 India

Our Reference: CMS/0809/383 (2/1/2009)

Contents

1 Purpose of this Document

The Centre for Modeling and Simulation, University of Pune, plans to purchase a blade-based high-performance computing (HPC) cluster system for the needs of in-house scientific research in the area of computational/systems biology and biological sequence analysis. We are looking for a turn-key end-to-end solution complete with hardware, software, implementation, and support. The purpose of this Tender Notice (TN) / Request for Proposal (RFP) is to describe in detail (a) our requirements, and (b) our terms and conditions.

2 Terms and Conditions

2.1 Budgetary Provision for this Purchase

Rs. 30,00,000/-

2.2 Validity of Pricing

Minimum 120 days from the last date for tender/proposal submission.

2.3 Vendor Eligibility and Representation

1. Vendor Eligibility. A vendor much satisfy the following requirements to be eligible to submit proposals:

   (a) International original equipment manufacturers (OEM) with proven track-record in building and supporting HPC cluster platforms (both conventional and blade-based) for scientific research.

   (b) Continuous presence on `http://www.top500.org/` during 2003-08, with the maximum vendor share not less than 2.5% over the same time period.

   (c) Adequate documented experience during 2003-08 in setting up blade-based HPC clusters capable of at least 1TF sustained performance.

   (d) Adequate support infrastructure in India (preferably in the Pune region).

   (e) Adequate representation in India's scientific establishments.

2. Vendor Representation. An eligible vendor may designate, at their discretion and convenience,

   (a) one business partner to represent the vendor during the purchase process,

   (b) one implementation partner for implementing the HPC cluster system, and

   (c) one support partner for providing all post-implementation support.

   Designated partners must be provided with explicit authorization letters by the vendor. Implementation and support partners, in particular, must have adequate experience in implementing and supporting linux-based HPC cluster systems in scientific establishments in India. Irrespective of who is designated by the vendor as their representative, the ultimate responsibility for delivery, implementation, and support will be with the selected vendor.

3. Single-Point-of-Contact Support. We require a single point of contact with the vendor for the purchase process, implementation, and post-implementation and warranty support, irrespective of who represents the vendor.

2.4 Proposal Submission

All details on this can be found in Sec. 3.

2.5 Proposal Evaluation and Vendor Selection

1. All submitted proposals will be screened by a Technical Committee for their technical merit relative to the needs of proposed scientific research, computing power, power and cooling requirements, etc., and will be ranked accordingly. Technical evaluation of submitted proposals will be based on (a) benchmark results (see Sec. 5 for details), and (b) hardware configuration offered (see Sec. 4 for minimum requirements). Submission of benchmark results is a must for this purchase. Hardware configuration/s offered must satisfy the minimum requirements of Sec. 4. Proposals offering a technically superior hardware configuration with minimal cabling and greater computing power (as measured by the number of compute nodes) will be given preferential treatment. In all technical matters, the decision of the Technical Committee will be final.

2. Qualified technical proposals will be screened for commercials.
3. Final pricing negotiations and vendor selection will take place in a University of Pune Purchase Committee meeting. The date of this meeting will be communicated to qualified vendors.
4. University of Pune reserves the right to disqualify any or all proposals without giving any reasons. University of Pune is not bound to necessarily accept the lowest-priced proposal.

2.6 Delivery

A purchase order will be issued by the University to the vendor selected by the Purchase Committee. We expect delivery of the HPC cluster system in its entirety within 4 weeks after the date of this purchase order.

2.7 Implementation

We expect implementation of the HPC cluster system to be completed by the vendor within 2 weeks after delivery. End-goals of implementation are: (a) the deployment of the HPC cluster system complete with hardware, OS and clusterware, and user-specified software, and (b) a clear demonstration that the system is fully functional and usable for the end-user for scientific/computational research.

2.8 Testing and Certification

The warranty on the HPC cluster system will begin on the date the HPC cluster is demonstrated by the vendor to the Centre's technical team to be fully operational and working satisfactorily. This date will be decided as follows: Upon completion of implementation of the entire cluster system (hardware+software) by the vendor, the Centre's technical team will test it for not more than one week at full computational load. If no problems of any kind show up during this test period, the system will be certified by the Centre's technical team as "fully functional and working satisfactorily". Warranty on the cluster system will begin on the day of this certification.

If any problems show up, they will need to be corrected by the vendor, and the Centre's technical team will again subject the cluster system through the mandatory testing period. This test cycle will be repeated as many times as required until the cluster system is demonstrated to be fully functional to the Centre's technical team's satisfaction.

2.9 Warranty and Support

1. Warranty. Your proposal must provide, in the least,
   (a) 3-year on-site comprehensive warranty with next-business-day response/support for all hardware.
   (b) 3-year on-site next-business-day support for everything related to the operating system, clusterware, and software setup.
2. Single-Point-of-Contact Support. Irrespective of who represents the vendor (see Sec. 2.3), we need one single point of contact with the vendor for all and post-implementation/warranty support.

2.10 Payment

Upon certification of the fully-implemented HPC cluster system by the Centre's technical team as "fully-operational and working satisfactorily" (see Sec. 2.8 for details), University of Pune will make full payment within 4 weeks by a mutually agreeable method.

3 Proposal Submission

3.1 General Instructions

1. Technical and Commercial Proposals. Technical and commercial proposals should be addressed to Director, Centre for Modeling and Simulation, University of Pune, and submitted in separate sealed envelopes. Envelopes containing proposals should clearly indicate
   (a) vendor name,
   (b) type of proposal: technical or commercial, and
   (c) our reference number for this Tender Notice/RFP (see below).

   Detailed instructions on the format of the technical and commercial proposals can be found in Sec. 3.2.

2. Our Reference Number. Please quote our reference number "CMS/0809/383 (2/1/2009)" on all correspondence.
3. Clarification Enquiries. Any clarification enquiries may be directed to the office of the

   Centre for Modeling and Simulation, University of Pune, Pune 411 007 India
   Phone: (20).2569.0842, (20).2560.1448. Email: office@cms.unipune.ernet.in. Web: http://cms.unipune.ernet.in/
4. **Last Date for Tender/Proposal Submission:**
   7 business days from the _date_of_publication_ of a short tender notice in newspaper/s.

3.2 Response Format

3.2.1 Technical Proposal

1. Vendor Information. Vendor profile, together with sufficiently detailed notes on
   (a) Expertise+experience in building HPC cluster systems (both conventional and blade-based) with linux-based software setup.
   (b) Support infrastructure in India and in the Pune region.
   (c) Presence in India's scientific establishments: Provide sufficiently detailed information including name of the establishment, purpose of the HPC cluster system supplied, nature/configuration of the HPC cluster system, year of purchase, contact person information if available.
   (d) Any prior presence on the University of Pune campus.
2. Vendor Representation. Sufficiently detailed information and profile of the implementation and support partners, if applicable (see Sec. 2.3), together with contact details, and focus on their experience in implementing and supporting linux-based HPC cluster systems in scientific establishments in India.
3. Complete Technical Specification of the Offered HPC Cluster System. Technical specifications of the offered system must satisfy the minimum requirements of Sec. 4. **A vendor may offer more than one alternative/option for the hardware configuration.** For each hardware configuration alternative offered, we need the following information:
   (a) Complete spec-sheets, brochures, and URLs to information pages on the vendor's website, etc., for each major cluster component (chassis, compute nodes, master node, external storage, any peripherals, software setup, etc.).
   (b) Peak power rating of the complete system, plus power and cooling requirements.
   (c) A clear summary of what is offered over and above the minimum requirements.
4. Results of the Benchmark Tests. Include a CD with requested results of benchmark tests (Sec. 5) together with relevant information about the hardware configuration + software set-up of the system used for running the benchmarks.

3.2.2 Commercial Proposal

1. Vendor Representation. Relevant information and profile of the business partner, if applicable (see Sec. 2.3), together with contact details.
2. Complete Pricing Details of the Offered HPC Cluster System. For each of the technical alternatives proposed:
   (a) Pricing for each major HPC cluster system component (master node, compute nodes, infiniband switch, chassis, external storage, software and implementation).
   (b) Per-node pricing of the offered compute node.

   All hardware prices are to be quoted in USD (CIF Mumbai): Customs clearance will be taken care of by the University of Pune. University of Pune is exempt from octroi duties levied by the Pune Municipal Corporation, and the Centre's office will provide an octroi exemption certificate if and when necessary. Any other applicable taxes should be mentioned clearly.
3. Warranty Details. We assume that all terms and conditions from our side (see Sec. 2) are accepted by the vendor. Any additional features offered over and above the minimum required warranty and support terms (Sec. 2.9) should be clearly mentioned.
4. Information about Single-Point-of-Contact for Warranty and Support. Complete contact information for the single-point-of-contact for warranty support (Sec. 2.9).

4 Minimum Requirements for the HPC Cluster System

We require a HPC cluster system consisting of

1. One master node,
2. At least 8 blade servers to act as compute nodes,
3. An Infiniband switch to interconnect the nodes,
4. A blade chassis capable of housing at least 10 blade servers (to accommodate future expansions),
5. An external DAS storage (to be physically connected to the master node but accessible to all nodes) with at least 1.3TB of usable storage at RAID 5,
6. Software components (operating system, cluster management and monitoring software, compilers, etc.) necessary for the operation of the complete system as a scientific computing platform, plus implementation and deployment of the complete system.

Minimum requirements on each of these components are specified in Sec. 4.1–4.6.

4.1 The Master Node

Configuration	The master node should have the same base-level configuration as for a compute node (see Sec. 4.2), except for the form factor. In addition, it should be able to connect to the external storage via SAS connectivity satisfying the following requirement/s:
SAS Connectivity to the External Storage	Requisite card/s and cable/s must ensure at least 3GBPS SAS connectivity between the external storage and the master node. SAS connectivity between external storage and the master node must be such that it is able to channel I/O to-and-fro between the external storage and the compute nodes without additional load on the master node's processors. Features ensuring redundancy of I/O paths will be preferred. In case the master node offered is a blade server, the blade chassis must include redundant SAS connectivity modules of compatible rating.
Form Factor	Either a blade server or a conventional rack-mount server external to the blade chassis.
Quantity	1

4.2 Compute Node Blades

Sockets	2 processor sockets per node, both populated with the processor specified below.
Processor	Intel Xeon E5450 (3.0GHZ, 12MB L2 cache, 1333MHZ FSB, 80W).
Chipset	Intel 5000P chipset.
Memory	16GB PC2-5300 667MHZ DDR2 SDRAM RDIMMs/DIMMs with advanced ECC, to be installed in matched pairs. Scalability upto 32GB plus adequate number of empty DIMM slots for upgrades.
Internal Storage and Controllers	2×146GB 3G SAS 10000RPM slim-form-factor/2.5" HDDs. One integrated two-port SAS controller / PCI-Express 64-bit / 133MHZ. Internal integrated hardware RAID controller with 128MB battery-backed write cache. Hot-swappability is not essential, but will be considered favourably.
Networking	Dual gigabit 10/100/1000 ethernet controllers with features including wake-on-LAN, full duplex, TCP/IP offload engine, load balancing or teaming, etc. Redundancy features (e.g., dual connections to ethernet switch/s) will be preferred.
Interconnect	Infiniband 4X DDR HCA. ConnectX or equivalent will be preferred.
Diagnostics	LED lights indicating failing components and on-board diagnostics.
Security Features	Power-on password, administrator password, unattended boot, selectable boot, etc.
OS Support	Blade servers should be able to support industry-leading operating systems such as standard Linux distributions (commercial and non-commercial).
Management Features	Each blade should have the capability of being managed individually and remotely through CLI-, GUI- and IPMI-based communication, irrespective of the operating system.
Connectivity to External Storage	Each compute node should be able to communicate with the external storage (see Sec. 4.5) via the master node (Sec. 4.1). We do not expect each compute node to be connected to the external storage separately.
Form Factor	We require the compute nodes to be blade servers.
Quantity	At least 8.

4.3 Infiniband Switch

Speed Rating	4X DDR. Should be compatible with HCA on master (Sec. 4.1) and compute nodes (Sec. 4.2), an provide fully non-blocking switching/interconnectivity between master and compute nodes.
Number of Ports	24, if infiniband switch offered is external to the chassis. If internal to the chassis (preferred), the number of ports on the switch should be adequate to interconnect a fully-populated chassis + one external master node.
Form Factor	May be internal to the blade chassis (preferred) or external. If external, we require a Voltaire.
Quantity	1

4.4 Blade Chassis/Enclosure

General Description	Upto 10U blade enclosure/chassis capable of (a) housing at least 10 hot-pluggable blade servers, and (b) of providing common resources for blade servers such as power, system management, cabling, ethernet management and expansion, external storage, switching and connectivity, and I/O (ports for USB, keyboard, video, mouse, optical drive, etc.). Chassis should provide adequate redundancy features.
Internal Resource Connectivity	Within-chassis connectivity of shared resources may be through a redundant mid-plane $(1 + 1)$ or a passive back-plane.
Ethernet Switch Modules	Redundant internal gigabit ethernet switch modules having at least 5 up-link ports.
SAS Connectivity Modules	In case the master node offered is a blade server that will be housed in the chassis, the chassis should be equipped with redundant internal 3GBPS SAS connectivity modules to enable connection with external storage (Sec. 4.5).
Infiniband Switch Module	See Sec. 4.3.
Management Modules	Chassis to be equipped and configured with hot-pluggable management modules (preferably in redundant configuration) in order to provide IP-based management of the blade servers and other vital components, switching, health monitoring, inventory management, and remote console to each blade.
KVM Support	Blade chassis should be equipped with support for keyboard, video, and mouse along with management modules. Local ports preferred and redundancy features welcome.
CD/USB	Chassis may be equipped with at least 1 USB port and a DVD-ROM drive (internal or external). Both should be accessible from and usable by individual blades inside the chassis.
Cooling	100% redundant cooling to be provided inside the chassis. Cooling system should be capable of dissipating, **efficiently and without processor throttling**, all heat generated by a fully-populated chassis with highest-configuration blade servers running at highest possible power rating.
Power Modules	$N + N$ redundancy to be provided on hot-pluggable power supplies powering the chassis, with each power supply unit of the highest capacity available with the vendor. Redundancy in power paths powering individual blades will be considered favourably.
Chassis Management and Software	Chassis should provide support for (a) remote console management, (b) powering on/off individual blades, (c) monitoring of power status, operating system events, temperature, disks, blowers/fans, power modules, and system diagnostics through the chassis management software. Chassis management controller/software (CMC/S) should be from the OEM itself, and software licenses for a fully-populated blade enclosure should be included in the chassis price. The CMC/S should provide standard features such as: (a) role-based (admin, user, operator, etc.) security that allows effective delegation of management responsibilities by giving systems administrators granular control over which users can perform which management operations on which devices, etc.; (b) proactive notification of actual or impending component failure alerts; (c) automatic event handling that allows notification of failures via e-mail; (d) performance monitoring and analysis features such as detection and analysis of hardware bottlenecks; (e) user-friendly GUI/console-based configuration and deployment of OS and software applications on individual blades. Desirable features include: (a) comprehensive system data collection and ability to produce detailed inventory reports for managed devices; (b) proactive identification of out-of-date BIOS, drivers, and server management agents; (c) remote update of system software/firmware components; (d) scheduling periodic server configuration snapshots; etc.
System Panel	LED panel to provide adequate information about power-on, location, over-heating and other system error conditions, etc.
Platform Support	The same chassis should be capable of simultaneously housing blade servers with (a) Intel Xeon and AMD Opteron processors, and (b) dual- and quad-core processor varieties.
Clustering Support	The chassis should support configuration of high-availability cluster with no single point of failure on components like switches, I/O connectors, power modules, etc.
Form Factor	Rack-mountable on a standard 19" rack.

4.5 External Storage

Type	DAS storage with SAS disks and at least 3GBPS SAS connectivity with the master node.
OS and Clustering Support	Storage array shall support industry-leading OS platforms including such as Linux and Windows. Storage system shall support all above operating systems in clustering.
Capacity and Scalability	Storage array shall be offered with appropriate number of 146GB 15000RPM SAS drives to ensure at least 1.3TB of usable storage at RAID 5 (typical/standard linux filesystems such at `ext3` assumed). Storage should be scalable to at least 48 drives.
Global Hot Spare	At least one global hot-spare drive shall be provided and configured for the array.
Processing Power	Controllers shall be based on latest SAS technology to ensure no bottleneck for I/O.
Architecture and Processing Power	Storage array shall support dual, redundant, hot-pluggable, active-active array controllers to avoid issues related to software-based RAID. Storage array shall have switched architecture for disk drive connectivity. Desirable and preferable: active-active controllers to support dynamic controller failover and failback.
No Single Point of Failure	Storage array shall be configured in a no-single-point-of-failure configuration; this includes components like array controller card, cache memory, fan, power supply, etc.
Disk Drive Support	Storage array should be able to support dual-ported 15000RPM 146/300/450GB hot-pluggable enterprise SAS hard drives together with SATA drives in the same device shelf.
Cache	Storage array should have at least of 1GB cache per controller.
RAID Support	Storage array shall support RAID levels 1, 1+0, and 5 in the least. On deployment, we expect the storage to be configured at RAID level 5.
Data Protection	In case of power failure, storage array shall be able to hold data in the cache for at least 72 hours of time. For optimal data protection the units software logic shall not reside on HDDs reserving them only for user data.
Host and Back-End Ports	Storage shall have at least 2 host ports for connectivity to servers.
Ports Bandwidth	Storage shall provide 3GBPS SAS connectivity to the host/master node.
Performance	Storage subsystem shall support more than 90,000IOPS and 700MBPS sequential throughput.
Software	Software+licences necessary for storage array configuration, operation, multi-path/load-balancing, and performance analysis should be included. Any performance management features will be considered favourably.
Form Factor	Rack-mountable on a standard 19" rack.

4.6 Software and Implementation

General Requirements	The proposed HPC cluster system should be deployed with (a) an open-source Linux-based operating system with adequate device driver support; tools for cluster installation and management that support node-group and repository manager for deploying updates, patches, etc., or for quickly re-imaging new nodes with no interruption in uptime; tools for monitoring cluster health, resource usage; and a job scheduler; (b) compilers, MPI, and code development tools; (c) installation/integration of user-specific scientific applications (see below); (d) integration of all software components so as to make the complete HPC cluster system fully functional and usable (e.g., integration of the scheduler with MPI, any license managers, etc.)
Operating System	`Rocks` 5.1 or latest stable release.
Workload and Cluster Management	Latest stable release of a reputed workload and cluster management software suite. Scheduling and cluster management software should support policy-based workload management, graphical cluster administration interface, monitoring and reporting tools, etc. Open-source software preferred.
Compilers and MPI	1 academic single-user license of the Intel Cluster Toolkit Compiler Edition 3.2 for Linux (`http://www.intel.com/cd/software/products/asmo-na/eng/375500.htm`) or latest release.
Scientific Applications	Installation, integration, and any performance tuning of (a) standard numerical libraries (`BLAS, LAPACK, ATLAS, FFTW`), (b) `MEME, ClustalW-MPI` and `VASP` (source codes will be provided by us), (c) non-default `Rocks` rolls (e.g., `Bio`) and other standard open-source scientific tools/applications in consultation with the Centre's technical team.
HPC Cluster Implementation	End-goals of implementation are (a) the deployment of the HPC cluster system complete with hardware, OS, clusterware, and user-specified software, so that it is functional and usable for the end-user for scientific/computational research, and (b) a clear demonstration of the same.

5 Benchmark Tests for Technical Evaluation

Technical evaluation of a proposed HPC cluster solution will be based on the results of a number of benchmark tests described below in Sec. 5.1–5.3. Our benchmark tests have been designed in consideration of the computational resource usage profile (i.e, heavy number-crunching by and large, with moderate to heavy I/O in spurts) of the actual tools that will be used for the scientific research using the proposed HPC cluster.

The HPC cluster system used for benchmarking must satisfy all other minimum requirements (Sec. 4) (except for peripherals such as the external storage, etc.). In case a vendor does not have an Intel-Xeon-E5450-based system available for benchmarking, these tests may be run on a system with a related processor (e.g., Intel Xeon E5472). Results obtained on processors other than E5450 may be scaled appropriately for fair comparison.

Specifically, we need the following data for each of the benchmark tests described in Sec. 5.1–5.3:

1. The following results on 8, 16, 32, 64, 96 cores:
    - Run time, as provided by the Unix `time` command or equivalent.
    - Output generated by each of the runs (`stdout`, `stderr`, plus any output files). Please make sure that output files do not get overwritten across runs over different number of cores.
2. Complete and detailed configuration information (hardware configuration + software setup) of the HPC cluster that was used to run these benchmark tests.
3. Output (`stdout`+`stderr`) of the compilation sequence for the tools below.

In the command lines below,

- you may need to add the prefix `nohup time` and/or the suffix `2>&1` for output redirection, depending on your cluster setup; and
- `<cores>` stands for the number of cores used for the run.

5.1 MEME

MEME 3.5.4 (`http://meme.nbcr.net/downloads/meme_3.5.4.tar.gz`) is an open-source motif search tool. Input files `input-meme-seq-2.txt` and `input-meme-bg-2.txt` are available upon request (see Sec. 3.1 for Clarification Enquiries). The `meme` command line to be used is:

```
meme input-meme-seq-2.txt -dna -bfile input-meme-bg-2.txt -nmotifs 2 -mod anr -revcomp -nsites 372 -maxsize 1500000 -text -p <cores>
```

5.2 ClustalW-MPI

ClustalW-MPI (`http://www.bii.a-star.edu.sg/docs/software/clustalw-mpi-0.13.tar.gz`) is an MPI implementation of the open-source multiple sequence alignment tool ClustalW. Input sequence data file `input-clustalw.txt` are available upon request (see Sec. 3.1 for Clarification Enquiries). The `clustalw-mpi` command line is:

```
clustalw-mpi -infile=input-clustalw.txt -outputtree=dist
```

5.3 The NPB Benchmark Suite

Detailed information on NPB is available at `http://www.nas.nasa.gov/Resources/Software/npb.html`. Use version 3.3 (non-grid, non-MZ variety). We need results (as described above) for all the NPB benchmarks (i.e., levels 1, 2, and 3).

5.4 Optional Benchmarks

A vendor may, at their discretion, also provide results of other relevant international benchmark suites such as the SPEC MPI2007 (`http://www.spec.org/mpi2007/`) in support of their proposal. These additional benchmark results are optional.