ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-WEPAB310

STUDY AND DESIGN OF A HIGH-PERFORMANCE COMPUTING INFRASTRUCTURE FOR IRANIAN LIGHT SOURCE FACILITY BASED ON THE ACCELERATOR PHYSICISTS AND ENGINEERS' APPLICATIONS REQUIREMENTS*

K. Mahmoudi†, A. Khaleghi[1], H. Haedar, Imam Khomeini International University, Qazvin, Iran
M. Akbari, Iranian Light Source Facility, Tehran, Iran
S. Mahmoudi, Iran University of Science and Technology, Tehran, Iran
[1]also at Iranian Light Source Facility, Tehran, Iran

Abstract

Synchrotron design and operation are one of the complex tasks which requires a lot of precise computation. As an example, we could mention the simulations done for calculating the impedance budget of the machine which requires a notable amount of computational power. In this paper we are going to review different HPC scenarios suitable for this matter then we will present our design of a suitable HPC based on the accelerator physicists and engineers' needs. Going through different HPC scenarios such as shared memory architectures, distributed memory architectures, cluster, grid and cloud computing we conclude implementation of a dedicated computing cluster can be desired for ILSF. Cluster computing provides the opportunity for easy and saleable scientific computation for ILSF also another advantage is that its resources can be used for running cloud or grid computing platforms as well.

INTRODUCTION

In the design phase of a synchrotron light source, various software is used to optimize and simulate the design of the accelerator lattice and its various components. The duration of these simulations and optimizations to achieve the desired accuracy in the results, depending on factors such as the dimensions of the simulated components, the specific geometry of the simulated components, the number of particles used in the simulation, the algorithms used in the optimization, etc. can be very long. As a bottleneck, it slows down the performance-dependent designs of these simulations and optimizations. Therefore, to speed up the design of accelerators and their components, various supercomputer systems are used, such as cluster computing, and grid computing.

In the commissioning and operation phase, a variety of software is developed and used for tasks such as analysis, optimization, improvement and troubleshooting of the accelerator and its components. The development of such software is ongoing and follows the new approaches in the accelerator community. The software used by physicists and accelerator subsystem specialists have different processing and hardware requirements depending on their development and functionality. In this paper with the review of various HPC scenarios and the software and hardware requirements and infrastructures used by other light sources [1-31] we have proposed a suitable HPC design for the ILSF as described in the next section.

PROPOSED HPC DESIGN

For the following reasons, a cluster system is recommended for use in optimizations, simulations, and analyses performed by various applications of accelerator physicists and accelerator specialists:

- Has a lower implementation cost regarding its performance.
- The price of hardware and software is reasonable and fairly low.
- The cost of support and maintenance is low.
- It is possible to develop and update the system at a relatively reasonable cost.
- The possibility of easy system upgrade according to the increasing need.
- If necessary, it is possible to use this system in a grid or cloud computing architecture.
- Most high-performance computing systems use this type of architecture for implementation.
- A number of synchrotrons including Diamond and SESAME light sources, also use cluster computing system for their calculations.

One type of cluster that has different applications in meteorological, seismic and science systems is the Beowulf cluster, which has a relatively good performance. This type of class has a simple architecture that consists of a server node and several computational nodes and a network for communication between nodes (Fig. 1). In the following, a brief explanation is given about each of the hardware and software components selected for the proposed system and the reasons for choosing each one.

Server and Compute Nodes Specifications

The processors used in the design of the proposed cluster are processors made by Intel. The server of this cluster has an 8-core Zeon processor that has good computing power. There is also 32 GB of main memory or RAM for this server. In addition to the 8-core Zeon processor, the server is also equipped with NVIDIA Tesla graphics cards.

Considering factors such as the time required and the type of optimizations and simulations performed in ILSF's specialized groups, the number of groups that need fast processing services and the estimated number of simultaneous simulations, 32 computational nodes are currently

* Work supported by Iranian Light Source Facility.
† kmahmoudi@edu.ikiu.ac.ir

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-WEPAB310

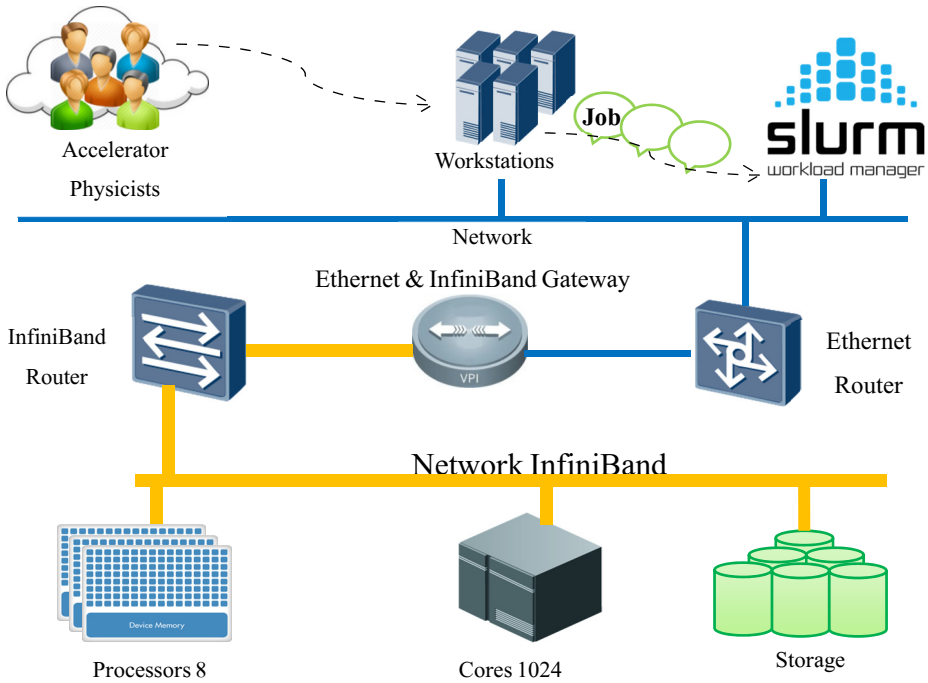

Figure 1: Simple scheme of high-performance computing system for Iranian Light Source Facility [32].

proposed for this computing system. Naturally, as the needs increase, the number of nodes can change. The current number of cores per node is considered to be 32, but due to technology upgrades and lower costs for processors with more cores, this number can be increased during actual implementation. In this case, assuming a constant budget, a compromise can be made between the number of nodes and the number of cores per node.

In cluster architecture, servers with one, two or more processors can be used as the main building blocks of the cluster. The most important criterion for selecting these building blocks is the balance between performance and cost. Multiprocessor servers often cost more than their single processor counterparts. If single-processor servers are used with a high-performance connection subsystem such as InfiniBand or Myrinet, the total cost of these connections will be higher than the cost of servers industry.

Network Technology

Different programs that run in parallel on high-performance computing clusters usually generate a lot of communication traffic on cluster connections, and the type of these connections can play an important role in the performance and bandwidth performance of the connections. Given that the computing system must be able to have acceptable computing power, the connection between the nodes of the computing system, which can be the bottleneck of the performance of this system, is of particular importance. Options such as fast Ethernet, Gigabit Ethernet, Myrinet, and InfiniBand are often used for network connections. The last two options are more expensive than the usual types of fast and Gigabit Ethernet connections. The type of connections required by the cluster can be selected based on its behaviour. For distributed and standalone applications that do not require large network connections, the use of high-speed Ethernet is a cost-effective option, although application with more data communication need require faster connections such as InfiniBand for better performance. Considering current needs and to maintain flexibility and scalability in responding to new needs, there should be a fast-internal connection between cluster nodes. One of the most widely used networking technologies is Gigabit Ethernet technology, which is used in most computing clusters, including those used in the Diamond Light Source. This technology has a high data transfer rate and low latency. On the other hand, having the ability to communicate through InfiniBand, provides more redundancy and the possibility of communication with less latency than Gigabit Ethernet. It is recommended to have this feature in applications where there is a lot of storage and processing load at the same time.

The Fat Tree topology can be used to network computational nodes. This topology, which is very common in the design of computational clusters, is configured as a tree with multiple roots. This topology can provide the best performance for a very large-scale computing cluster if configured as a dead-end free network. This topology usually uses the same bandwidth for all nodes and the network switches used in this network have the same number of ports [33].

Operating System

The proposed operating system for this cluster is Linux. The reason for choosing the Linux operating system is that it is an open source and is easily and freely available to users. For software that either does not have a Linux version, or for reasons such as lack of a license, it is not possible to use the Linux version, methods such as

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-WEPAB310

virtualization can be used to allocate a part of the computing cluster capacity to the operating system compatible with that software.

Programming Model

MPI is a standardized message-passing interface designed by a team of academic and industrial researchers to work on a wide range of parallel computers. Due to the widespread use of this parallel programming communication protocol in analysis software with parallel processing capability, providing the possibility of developing and executing such programs requires the implementation of MPI in the computational cluster system.

Job Scheduling

There are various queue, prioritization, and scheduling programs for processing jobs submitted to computing clusters, including SGE, Univa Grid Engine, HT-Condor, and Slurm. The Univa Grid Engine is commercial, but the other three options are open source. SLURM and HT-Condor have a more active community than SGE, and have been used in some accelerators with a large number of submitted processing tasks, such as CERN. On the other hand, compatibility or ease of adapting the job scheduler with the software used by accelerators is another point to consider when choosing it. HT-Condor is more commonly used for HTC applications and has fewer HPC-related capabilities than SLURM [34]. Therefore, at this stage of the design, the use of SLURM work scheduler is suggested, although if another specific scheduler is needed, using methods such as virtualization, it is possible to segment the computational cluster and use that special scheduler in the relevant section.

The type of hardware of the operating node and computational nodes, the type of network technology, the operating system and the type of programming model of the proposed cluster are summarized in Table 1.

CONCLUSION

The High-Performance Computing (HPC) systems or super computers usually can execute numerous operations in a fraction of time unit. They are capable to do more than 1012 operation per second using scalable hardware and reliable and high-capacity storage in national to multinational level.

Regarding technical and financial aspects, different architecture including shared memory, distributed memory, distributed memory having nodes with shared memory etc. are developed

To interconnect the computing resources and increase the computing capacity and performance, different approaches such as, cluster computing, grid computing and cloud computing are in front of us.

HPC solution of different synchrotron like ESRF, ELETTRA, Diamond and SESAME were studied and their design specifications regarding hardware, software and networking aspect were highlighted. These studies and the relevant literature review led to the proposed design and architecture which included 32 nodes, 1024 processing core and at least 8 high performance GPU. In this design different aspects of HPC such as processor, memory, storage network area, operating system, software and job scheduling system were addressed. Regarding the foreseen gradual increase in ILSF computing and storage needs and considering parameters such as maintainability, much effort was put on this design to be scalable and flexible.

Table 1: The Type of Hardware and Software Used in the Proposed Cluster

Item	Specification
Server type	HPE ProLiant G9 (Better or Similar one available at the implementation time)
Server CPU type	Intel Xeon (Better or Similar one available at the implementation time)
Server Memory	32 GB
Server GPU	NVIDIA TESLA K20 (Better or Similar one available at the implementation time)
Server Hard disk Capacity	2 TB
No. of Computing nodes	32
Computing node CPU type	CPU intel coreTM i7 (Better or Similar one available at the implementation time)
Per node memory	128 GB
Computing Node GPU	NVIDIA TESLA K20 (Better or Similar one available at the implementation time)
Hard disk capacity for each node	1TB
Network	10 Gigabit Ethernet+Infiniband
Ethernet Router	MikroTik Router RB1100AHx4
InfiniBand Router	SB7780 InfiniBand router
Storage	EMC Rackmount NAS Storage VNXB54DP25F
Operating System	Linux
Parallel Environment	MPI
Job Scheduler	Slurm

REFERENCES

[1] System Requirements for COMSOL Multiphysics, COMSOL Multiphysics Modeling Software, https://www.comsol.com/system-requirements

[2] What hardware do you recommend for COMSOL Multiphysics?, Comsol Multiphysics Modeling Software, https://www.comsol.com/support/knowledgebase/866/.

[3] N. Juntong and S. Krainara, "The New 118 MHz Normal Conducting RF Cavity for SIAM Photon Source at SLRI", in *Proc. 5th Int. Particle Accelerator Conf. (IPAC'14)*, Dresden, Germany, Jun. 2014, pp. 3896-3898. doi:10.18429/JACoW-IPAC2014-THPRI055

ISBN: 978-3-95450-214-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2021-WEPAB310

[4] ANSYS Platform Support by Application/ Product, 2019, https://www.ansys.com/content/dam/it-solutions/platform-support/previous-releases/platform-support-by-application-product-2019-r3.pdf

[5] ANSYS Hardware Information, SimuTech Group, https://www.simutechgroup.com/support/ansys-resources/ansys-hardware-support

[6] M. El Khaldi, J. Bonis, A. Camara, L. Garolfi, and A. Gonnin, "Electromagnetic, Thermal, and Structural Analysis of a THOMX RF Gun Using ANSYS", in *Proc. 7th Int. Particle Accelerator Conf. (IPAC'16)*, Busan, Korea, May 2016, pp. 3925-3927. doi:10.18429/JACoW-IPAC2016-THPOW002

[7] CST Studio Suite - Operating System Support, http://updates.cst.com/downloads/CST-OS-Support.pdf

[8] CST Studio Suite 2019 GPU Computing Guide, http://updates.cst.com/downloads/GPU_Computing_Guide_2019.pdf.

[9] CST Studio Suite Recommended Hardware, 3D Design and Engineering Software, https://www.3ds.com/support/hardware-and-software/simulia-system-information/cst-studio-suite/cst-studio-suite/.

[10] M. Borland and T. Berenc, *User's Manual for elegant*, Advanced Photon Source, Lemont, IL, USA, Jul. 2021. https://ops.aps.anl.gov/manuals/elegant_latest/elegant.html

[11] M. Borland, *Getting Started with SDDS*, Argonne National Laboratory, Lemont, IL, USA. https://ops.aps.anl.gov/manuals/GettingStartedWithSDDS/GettingStartedWithSDDS.pdf

[12] M. Borland, "Multi-objective optimization of storage ring dynamic acceptance and lifetime", presented at The Accelerator and Detector Research and Development Program Principal Investigators' 2011 Meeting, Annapolis, MD, USA, Aug. 2011, unpublished.

[13] X. N. Gavaldà, "Multi-Objective Genetic based Algorithms and Experimental Beam Lifetime Studies for the Synchrotron SOLEIL Storage Ring", Ph.D. thesis, Accelerator Physics, Université Paris-Saclay, Gif-sur-Yvette, France, 2016.

[14] R. Armstrong *et al*., "Toward a common component architecture for high-performance scientific computing", in *Proc. 8th Int. Symposium on High Performance Distributed Computing*, Redondo Beach, CA, USA, Aug. 1999, pp. 115-124. doi:10.1109/HPDC.1999.805289

[15] Arista, HPC Deployment Scenarios, https://www.arista.com/assets/data/pdf/Whitepapers/HPC_Deployment_Scenarios.pdf

[16] D. Culler, J. P. Singh, and A. Gupta, *Parallel Computer Architecture: A Hardware/Software Approach*, San Francisco, CA, USA: Morgan Kaufman, 1998.

[17] The free dictionary, "symmetric multiprocessing", https://encyclopedia2.thefreedictionary.com/symmetric+multiprocessing

[18] S. Mittal, "A Survey of Techniques for Architecting and Managing Asymmetric Multicore Processors", *ACM Computing Surveys*, vol. 48, p. 45, 2016. doi:10.1145/2856125

[19] D. A. Patterson and J. L. Hennessy, *Computer Architecture: A Quantitative Approach*, Burlington, MA, USA: Morgan Kaufmann, 2006.

[20] WikiPedia, Asymmetric Multiprocessing, https://en.wikipedia.org/wiki/Asymmetric_multiprocessing

[21] WikiPedia, Mémoire distribuée, https://fr.wikipedia.org/wiki/M%C3%A9moire_distribu%C3%A9e

[22] D. Graham-Smith, "Weekend Project: Build your own supercomputer", PC & Tech Authority, Jun. 2017. http://www.pcauthority.com.au/Feature/306972,weekend-project-build-your-own-supercomputer.aspx

[23] Mellanox, Interconnect your future, https://www.mellanox.com/related-docs/solutions/hpc/TOP500-JUNE-2019.pdf

[24] NICE Current Status, ESRF, http://www.esrf.eu/Infrastructure/Computing/NICE/Implementation

[25] Infrastructure, Elettra, https://www.elettra.trieste.it/lightsources/labs-and-services/scientific-computing/what-is-our-job.html

[26] U. K. Pedersen, N. Rees, M. Basham, and F. J. K. Ferner, "Handling high data rate detectors at Diamond Light Source", *Journal of Physics: Conference Series*, vol. 425, no. 6, p. 062008, 2013. doi:10.1088/1742-6596/425/6/062008

[27] M. Heron, "Computing and Networking at Diamond Light Source", 2016, https://technodocbox.com/123584608-Computer_Networking/Computing-and-networking-at-diamond-light-source-mark-heron-head-of-control-systems.html

[28] IMAN1: Jordan's national supercomputer center, http://www.iman1.jo/iman1/.

[29] M. Nasiri, "Difference between SAN and NAS", https://storage.tosinso.com/fa/articles/7256/.

[30] T. Muggendobler, "Storage: DAS, SAN, NAS", unpublished.

[31] C. T. Yang, "Cloud Storage and FreeNAS", unpublished.

[32] Y. Cheng, "Status of IHEP Site", presented at HEPiX Fall 2016 Workshop, Berkeley, CA, USA, Oct. 2016, unpublished.

[33] Designing a HPC cluster with Mellanox InfiniBand. Mellanox Interconnection Community, https://community.mellanox.com/docs/DOC-2392

[34] C. Hollowell, J. Barnett, C. Caramarcu, W. Strecker-Kellogg, A. Wong, and A. Zaytsev, "Mixing HTC and HPC Workloads with HTCondor and Slurm", *Journal of Physics: Conference Series*, vol. 898, p. 082014, 2017. doi:10.1088/1742-6596/898/8/082014