Proceedings of the VII Bulgarian-Serbian Astronomical Conference (VII BSAC)
Chepelare, Bulgaria, June 1-4, 2010,
Editors: M. K. Tsvetkov, M. S. Dimitrijević, K. Tsvetkova, O. Kounchev, Ž. Mijajlović
Publ. Astron. Soc. "Rudjer Bošković" No 11, 2012, 147-15

HPC CLUSTER WITH GPGPU CAPABILITIES. PERFORMANCE AND FEATURES EVALUATION

E. ATANASSOV[1], M. DECHEV[2], G. PETROV[2], A. KARAIVANOVA[1], T. GUROV[1] AND M. DURCHOVA[1]

[1] *Institute of Information and Communication Technologies, Acad. G. Bonchev 25A,1113 Sofia, Bulgaria*
[2] *Institute of Astronomy and National Astronomical Observatory, 72, Tsarigradsko chaussee Blvd., 1784 Sofia, Bulgaria*

Abstract: The high performance computing clusters that are being deployed lately increasingly incorporate GPGPU processing cards, in order to achieve high energy and space efficiency. This leads to development of hybrid computing models that attempt to balance and optimize the performance of the CPU- and GPU-based hardware elements. After the expansion of the HPC cluster at the Institute of Information and Communication Technologies with HP SL390S G7 nodes equipped with NVIDIA M2090 cards, we performed careful evaluation and benchmarking of the new capabilities of the cluster. In this paper we present the software and hardware architecture of the cluster and the results of our benchmarking process.

1. INTRODUCTION

Last years high performance computing cluster was build at the Institute of Information and Communication Technologies, Bulgarian Academy of Sciences. It became a true center of the Bulgarian grid infrastructure and ensure a full membership into European virtual organizations. Its characteristics – infiniband connectivity (90 % efficiency) and 96TB storage volume, the High Performance Computing (HPC) Cluster at Institute of Information and Communication Technologies, Bulgarian Academy of Sciences (IICT-BAS) is more powerful than a big part of the European grid clusters. Recently it was extended with HP SL390S G7 nodes equipped with NVIDIA M2090 cards. Here we present an overview of the Bulgarian HPC infrastructure, the software and hardware architecture of the cluster and the results of our benchmarking process.

2. BULGARIAN HIGH PERFORMANCE COMPUTING INFRASTRUCTURE

The Bulgarian HPC infrastructure (Fig. 1) consists of the biggest HPC resource for research in Bulgaria – the supercomputer IBM BlueGene/P with 8192 cores,

two big HPC clusters with Intel GPUs and Infiniband interconnection at IICT-BAS and Institute of Organic Chemistry with Centre of Phytochemistry approximately 1400 cores total. In addition GPU-enabled servers equipped with state of the art Nvidia GPU cards are available for applications that can take advantage of them.

Figure 1. - The Bulgarian HPC infrastructure.

The main part of the HPC cluster at IICT-BAS (Fig. 2) consists of 3 chassis HP Cluster Platform Express 7000 with 36 blades BL 280c. The blades are equipped with dual Intel Xeon X5560 @ 2.8Ghz providing a total of 576 cores, 24 GB RAM per node. The cluster has 8 management servers HP DL 380 G6 with dual Intel X5560 @ 2.8 GHz and 32 GB RAM. They are connected with optical Fibre Channel connections with two SAN switches, ensuring redundant access to two Storage Area Network (SAN) storage systems - MSA2312fc and P2000 G3 FC. The total storage space is 96TB (Fig. 3). It is divided in two filesystems for users - /home and /scratch, for permanent and temporary storage, respectively. The management of these high-performance lustre filesystems is performed by 3 management nodes - one MGS/MDS and two OST nodes. The lustre filesystem is open source and popular choice for HPC computing clusters, due to its capability to perform parallel high-performance I/O.

The internal network connectivity of the cluster is provided by fully non-blocking DDR Infiniband at 20 Gbps line speed. The non-blocking feature ensures

Figure 2. - HPC Cluster at IICT-BAS.

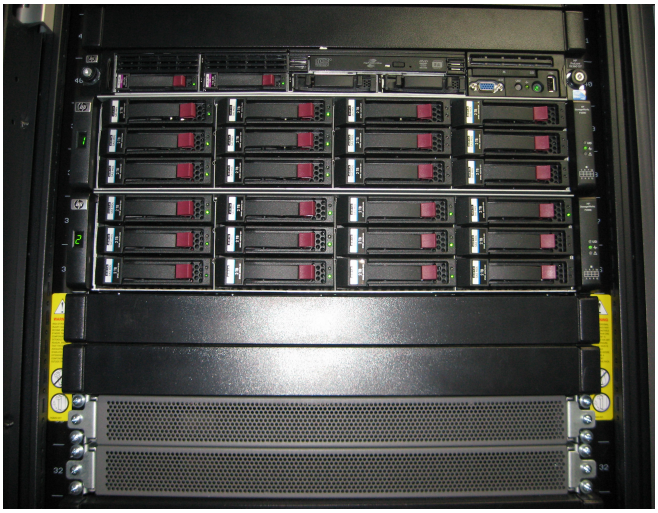

Figure 3. - Cluster Storage Area.

that applications do not run into network bottlenecks even when the cluster is fully used. The core switch of the Infiniband interconnection is a Voltaire Grid director 2004 Infiniband switch (Fig.4). This main advantage of Infiniband versus Ethernet consists in the low-latency of point-to-point communication, where the latency is measured at 2.5 μs between each two nodes.

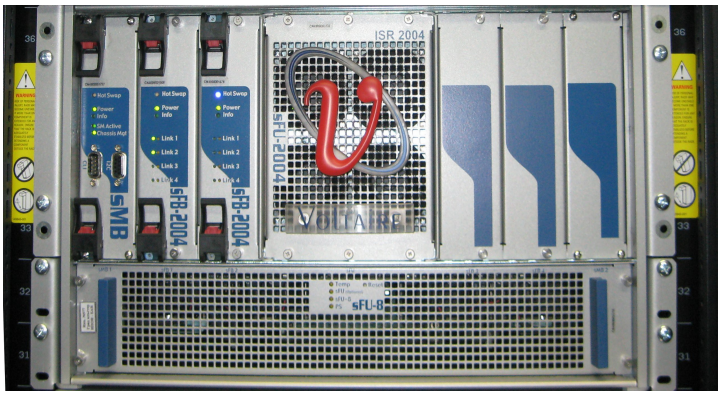

Figure 5. - Voltaire Grid director 2004 Infiniband switch.

As was mentioned above, two GPU-enabled servers were added recently. They are HP ProLiant SL390s G7 servers with Intel(R) Xeon(R) CPU E5649 @ 2.53GHz and 96 GB RAM, which support maximum 8 NVIDIA Testla cards. Currently they are equipped with M2090 cards. Every such GPU card has 512 graphic cores and has peak performance of 1331 Gigaflops in single precision calculations and 665 Gigaflops in double precision.

The cluster works under operating system Scientific Linux 5. The cluster can be accessed with Grid certificates, supporting several international and national Virtual Organizations.

3. TESTS AND BENCHMARKS

During the continuous certification process of the cluster, we performed a series of benchmarks.

a) High Performance Linpack

The performance of the supercomputers usually is done in billions floating points operations (double precision). The results of the first 500 systems are regularly published at www.top500.org.

For our test we used the software High-Performance Linpack (www.netlib.org/benchmark/hpl/). All test were done using all computing cores without hyperthrading since enabling the hyperthreading option does not improve

the result. In order to provide optimal parameters for the test we used the suggestions from http://hpl-calculator.sourceforge.net/.

Tests showed that the ratio between reached real performance and maximal theoretical performance is 3000/3225, i. e. more than 93%. This is an exceptional result and demonstrates the excellent parallel efficiency of the cluster, mainly due to the use of non-blocking Infiniband.

b) MPI Infiniband tests

One of the most popular tests for performance of the MPI communications are *osu_latency* and *osu_bw*. They measure the latency and bandwidth between servers in the cluster. O*su_latency* test results are presented in Table 1. They show that we can reach latency smaller than 2.5 μs. The results from the *osu_bw* tests are shown in Table 2. One can conclude that under 131072 byte size a maximal bandwidth of 1734 MB per second is reached, which is close to the theoretical maximum.

Size (bytes)	Latency (μs)	Size (bytes)	Latency (μs)
0	2.45	2048	8.75
1	2.54	4096	10.73
2	2.46	8192	14.74
4	2.17	16384	20.84
8	2.14	32768	33.15
16	2.18	65536	53.76
32	2.24	131072	100.04
64	2.44	262144	178.19
128	4.10	524288	333.10
256	4.45	1048576	625.75
512	5.01	2097152	1249.96
1024	6.22	4194304	2462.30

Table 1. OSU MPI Latency Test v3.1.1

Size (bytes)	Bandwidth (MB/s)	Size (bytes)	Bandwidth (MB/s)
0	0	2048	936.44
1	1.13	4096	1211.40
2	2.32	8192	1528.57

4	4.68	16384	1471.90
8	8.82	32768	1572.26
16	18.51	65536	1680.17
32	36.68	131072	1727.01
64	67.71	262144	1727.80
128	124.40	524288	1731.80
256	228.00	1048576	1733.65
512	427.28	2097152	1734.14
1024	698.39	4194304	1734.69

Table 2. OSU MPI Bandwidth Test v3.1.1

c) Filesystem tests

One of the standard test for filesystem is *bonnie*++. It is included into the Operating System (OS). At that test cluster reached high results. As an example the block read speed achieved was 436MB/s.

d) For the GPU cards the device query tests show the following:

Device 0: "Tesla M2090"	
CUDA Driver Version / Runtime Version	4.0 / 4.0
CUDA Capability Major/Minor version number:	2.0
Total amount of global memory:	5375 MBytes
(16) Multiprocessors x (32) CUDA Cores/MP:	512 Cores
GPU Clock Speed:	1.30 GHz
Memory Clock rate:	1848.00 Mhz
Memory Bus Width:	384-bit
L2 Cache Size:	786432 bytes
Total amount of constant memory:	65536 bytes
Total amount of shared memory per block:	49152 bytes
Total number of registers available per block:	32768
Warp size:	32
Maximum number of threads per block:	1024
Texture alignment:	512 bytes

4. RESOURCE MANAGEMENT AND INFRASTRUCTURE MONITORING

The main control of the resource utilization is performed via the torque batch system and the maui scheduler. The computational jobs can be submitted both locally or through the Grid. A special accounting system is installed to allow quick overview of the resource utilization. Currently the cluster has around 50 local users, mainly from institutes of the Bulgarian Academy of Sciences. Most of the cluster resources are used for parallel jobs, since this is its main advantage versus regular clusters. A special accounting system is installed in order to track the utilization of the cluster as well as other HPC clusters in the region (Fig 5).

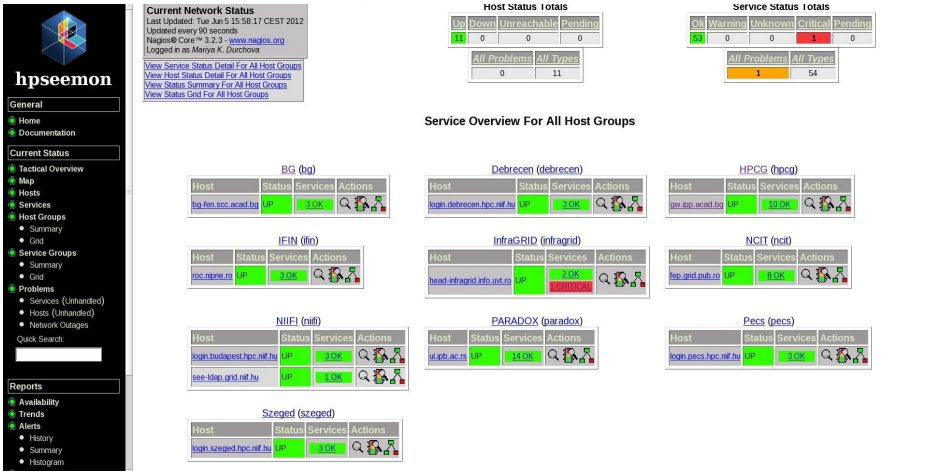

Figure 5. - Web-based cluster accounting system.

The cluster health status is monitored both though a local ganglia installation and using the EGI-Inspire nagios portal and the HP-SEE project regional nagios instance. Thus the administrators of the cluster receive alarms and notifications regarding major problems at the cluster.

5. CONCLUSIONS AND FUTURE WORK

The HPC cluster at IICT-BAS provides unique HPC resources, enabling development and use of advanced CPU-based and GPU-based parallel application from various areas of science. The use of the cluster enabled scientific research that otherwise would not have been possible. The balanced state-of-the-art hardware infrastructure can be further expanded by addition of more GPU-based

resources or more high-performance disk storage resources. In the future we plan to provide access also via open-source cloud middleware.

Acknowledgments

This work is supported by a grant of the Bulgarian National Science Foundation, Ministry of Education and Science, under number DVCP02/1 CoE Supper CA++ and DO-02-275.