PROCEEDINGS OF SCIENCE

CICC JINR Cluster 2008 Performance Improvement

A. Ayriyan[*1†], **Gh. Adam**[1,2], **S. Adam**[1,2], **V. Korenkov**[1], **A. Lutsenko**[1], **and V. Mitsyn**[1]

[1] *Joint Institute for Nuclear Research, 6 Joliot Curie St., 141980 Dubna, Moscow reg., Russia*
[2] *Horia Hulubei National Institute for Physics and Nuclear Engineering (IFIN-HH), 407 Atomistilor, Magurele - Bucharest, 077125, Romania*
E-mail: ayriyan@jinr.ru, adamg@jinr.ru

The floating point computation runs at the Central Information and Computing Complex (CICC) cluster of JINR accommodate requests for traditional sequential applications, parallel computing applications, as well as Grid applications launched within various virtual organizations. During the second half of 2008, the CICC JINR comprised a heterogeneous computing array of multicore processors with a total of 560 cores grouped in seven modules. Performance measurements made in 2007 and system exploitation evidenced the existence of bottlenecks which were identified and alleviated. As a result of the implemented optimizations, the system works efficiently for all three abovementioned categories of jobs. Within the Russian Data Intensive Grid (RDIG) consortium, which comprises, besides the CICC JINR, 14 Russian computing centres, our cluster covered a sizeable part of the RDIG share to the LHC projects during 2007–2008. Performance measurements using the High Performance LINPACK Benchmark showed relative figures at the level of the best results reported in the June 2008 TOP500 edition of the most performing computers in the world.

XII Advanced Computing and Analysis Techniques in Physics Research
November 3-7 2008
Erice, Italy

[*] Speaker.

[†] Work done within JINR themes 05-6-1060-2005/2010 (A.A., Gh.A., S.A., A.L) and 05-6-1048-2003/2010 (V.K., V.M). A. Ayriyan acknowledges partial support from RFBR 08-01-00800 grant and from Hulubei-Meshcheryakov Programme; Gh. Adam and S. Adam acknowledge partial support from contract CEX05-D11-67. The authors are grateful to the referee for advice resulting in the improvement of the presentation.

PoS(ACAT08)054

1. Introduction

The computing facilities at the Central Information and Computer Complex (CICC) in the Laboratory of Information Technologies (LIT) of the Joint Institute for Nuclear Research (JINR) were substantially increased during 2007–2008 by acquisitions of multicore processor modules enabled with 2 Gb/core RAM and Gigabit Ethernet (GbE) interprocessor connections inside each module. In 2007, three rackmount modules (two from T-Platforms and one from Hewlett-Packard) each consisting of 40 dual-core 2.66 GHz Intel Xeon arrived (this set is called thereafter GbE-I). During 2008, three blade modules of 20 quad-core 2.66 GHz Intel Xeon E5430 processors each were acquired from SuperMicro (this set is called in what follows GbE-II). A fourth module from SuperMicro, consisting of 20 quad-core 3.0 GHz Intel Xeon X5450 processors can work in two OS defined regimes of interprocessor connection: either under GbE interconnect (case in which it is called GbE-III) or under InfiniBand (InfB) interconnect (case in which it is called IfB).

The CICC JINR is asked to solve, under very tight budget constraints, a large variety of user requests concerning floating point computation tasks. Traditional users need outputs to applications implementing sequential algorithms for the solution of either theoretical or data analysis models that are being developed within different projects. A number of applications concerning the solution of large scale problems ask for high-performance parallel computing. Last but not least, the implementation of the Grid environment at JINR in connection with its participation in the LHC and other very large scale projects needed the development of entirely new specific infrastructure. The main difficulty in the efficient simultaneous solution of all these tasks on the CICC JINR cluster was identified to stem from the occurrence of severe bottlenecks on the networking structure of the CICC.

The present paper discusses the developed in-house solution for this problem starting from performance measurements of the 560-core cluster as a whole and of its separate subsets, GbE-I, GbE-II, GbE-III, IfB. Results of performance measurements are given in section 2. System optimizations which secured improved work for all kinds of requests (traditional sequential computing, parallel computing, Grid computing) are discussed in section 3. Conclusions are drawn in section 4.

2. Performance measurement results

The characterization of computer performance is usually associated to parallel computing. The three classifications of the most performing computing systems, TOP500 [1], CIS TOP50 [2], and China TOP100 [3] are based on outputs obtained from the High Performance LINPACK (HPL) Benchmark [4]. Data reported in our investigations ([5, 6, 7] and present paper), are based on the use of the HPL benchmark as well (version 1.0a of January 20, 2004). We used an Intel C Compiler v10.1 and the Intel Math Kernel Library 10.0. Table 1 provides a summary of the obtained data.

The measurements have been done by the end of August 2008 during an interruption of the Grid connection caused by maintenance and upgrade works. The performance measurement of the CICC cluster as a whole was possible during two runs only (when the 2982 GFlops figure reported in Table 1 was obtained). A two-dual-core-processor dye got then defective within the GbE-I subset and the measurements were continued with a 236-core configuration of this subset.

	Totals	**GbE-I**	**GbE-II**	**GbE-III**	**IfB**
Structure	7 modules Heterogeneous	3 rackmounts Homogeneous	3 blades Homogeneous	1 blade Homogeneous	
Year	2007-2008	2007	2008	2008	
CPU/cores	200/560	120/240	60/240	20/80	
Intel Processor		Xeon 5150 (dual-core)	Xeon E5430 (quad-core)	Xeon X5450 (quad-core)	
Frequency, MHz		2660	2660	3000	
RAM, Gb	1120	480	480	160	
Network	GbE	GbE	GbE	GbE	InfB
Performance Linpack/Peak, GFlop/s	2982/6067.2	1325/2511.04 *)	1414/2553.6	598.4/960	757.5/960
Efficiency (Linpack/Peak performances)	0.491	0.528	0.554	0.623	0.797
Coefficient matrix order	274400	220000	220000	120000	120000
Basic aim of computing	Distributed Parallel	Distributed	Distributed	Parallel Distributed	

*) During measurements a two-dual-core-processor-chip has fallen down. These figures and those in Fig. 1 for GbE-I correspond to a 236-core configuration.

Table 1: Summary of CICC JINR performance characterization

The performance reported in Table 1 placed CICC JINR at the 24-th place in the ninth edition (September 2008) [8] of the TOP50 rating list available at the date of this Conference.

3. CICC optimization

Performance measurements done in 2007 for the 240 dual-core processor cluster GbE-I [5, 6] and comparison with the Myrinet SIMFAP cluster implemented in Bucharest-Romania [5] as well as with TOP500 data provided hints on the need of further optimization of the cluster performance. The connection of each GbE-I module to the main Backbone Ethernet switch was modified to a four-port GbE trunk, such that the effective rate inbetween modules remained near to 1 Gbps. The present GbE-I performance of 1325 GFlops obtained under a 236-core configuration is about twenty percent higher than the previously obtained 1124 GFlops for the 240-core configuration [5, 6]. As it concerns the efficiency ratio, this was increased from 44 percent to nearly 53 percent. The almost 9 percent increase of the efficiency ratio is fully attributable to the latency reduction originating in the improvement of the inter-rack communication.

Since the processor clock frequencies within modules GbE-I and GbE-II are the same, the higher performance value obtained for the latter cluster comes from the lower inter-blade latency

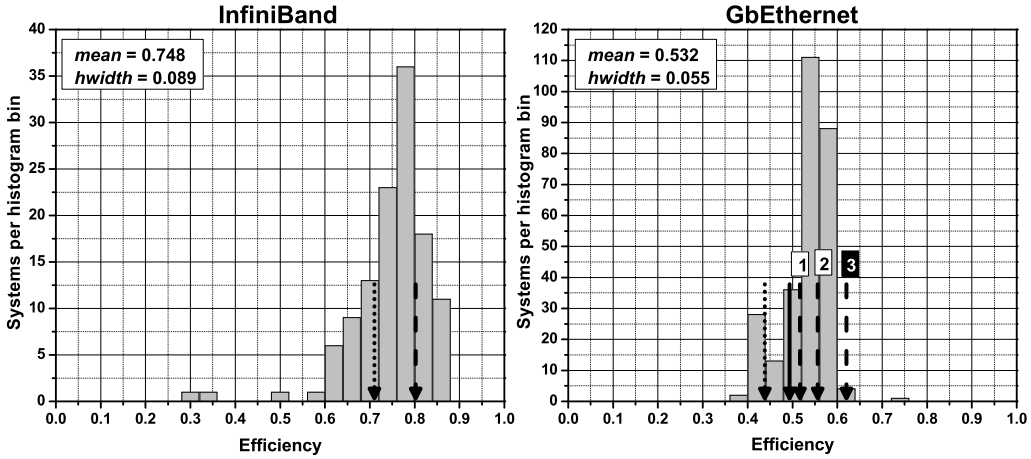

Figure 1: Performance increase of CICC JINR during 2008 as a result of optimization implementations and comparison with the histogram representations of the June 2008 issue of the TOP500 data for the InfB and GbE interconnects. *InfiniBand*: Dotted arrow points to the previously reported data in [7]. Interrupted-line arrow points to the newly measured performance. *Gigabit Ethernet*: Dotted arrow points to the previously reported data in [5, 6]. Solid line arrow points to present 560-core heterogeneous structure. Interrupted-line arrows point respectively to: 1 – GbE-I; 2 – GbE-II; 3 – GbE-III.

within GbE-II as compared to the latency among racks within GbE-I.

The 471.3 GFlops per-blade performance of GbE-II is 27 percent lower than the 598.4 GFlops figure found for GbE-III. Only half of this misfit may be attributed to the clock frequency differences. The other half comes from the inter-blade interconnect latency.

The IfB module measurements reported in [7] were done under 1 Gb/core RAM. The final acquisition of this module was made with 2 Gb/core RAM. The corresponding efficiency increase of eight and a half percent fully comes from the RAM/core increase.

The efficiency comparison of GbE-I with the November 2007 issue of the TOP500 data [6, 7] has shown consistency, even if below the average values, with the best results reported worldwide. The comparison of the efficiency of the present data with the June 2008 TOP500 data is done in Fig.1. While the 2008 TOP500 data showed sensible improvement as compared with the November 2007 TOP500 data, the present efficiency obtained for the CICC-JINR cluster has evolved faster. All the GbE homogeneous modules stay now inside the main peak, with the GbE-II efficiency being higher than the TOP500 peak and the GbE-III one being near to the highest maximum end. Once again, this is consistent with the above conclusion that the latency decrease of the inter-rack communication is a substantial source of improving the speed of the high performance computing. Finally, the present IfB efficiency stays beyond the maximum of the TOP500 distribution. This provides a supplementary confirmation of the above made statement that the 2 Gb/core RAM is a parameter the (low) cost of which is fully justified by the increase of the speed of the high performance computing.

4. Conclusions

The present study reported performance investigations devoted to better understanding and more efficient use of the computing facilities installed at LIT-JINR. It provided straightforward confirmation of the usefulness of such investigations for the solution of computing tasks both at LIT JINR and at computing centres in JINR Member States.

We have found that low cost improvements of the hardware connections resulted in substantial gains concerning the latency decrease and the consistent data handling to/from the associated mass storage peripherals. The increase of the RAM/core also resulted in sensible performance gains too.

With these improvements, the efficiency of the high performance computing done on each of the three homogeneous modules of the LIT-JINR cluster stays at the upper end of the reported statistics in the June 2008 edition of the TOP500 [1] list. Therefore, the home made hardware optimizations and free software installation brought significant speedup for all the kinds of the incoming computing tasks (traditional sequential, Grid distributed, or parallel).

The CICC JINR is the main provider of resources for intensive floating point computations asked by a large variety of projects done either in JINR or with JINR participation. Under the implementation of the "first come, first served" principle in the dynamic allocation of the available resources according to the users' requests, CICC JINR has been able to cope in 2008 with all the tasks under tough budgetary constraints.

References

[1] *http://www.top500.org/*

[2] *http://www.supercomputers.ru/*

[3] *http://www.samss.org.cn/2007-China-HPC-top100-20071110-eng.htm*

[4] A. Petitet, R.C. Whaley, J. Dongarra, A. Cleary, *HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers*, *http://www.netlib.org/benchmark/hpl/*, 2004.

[5] Gh. Adam, S. Adam, A. Ayriyan, E. Dushanov, E. Hayryan, V. Korenkov, A. Lutsenko, V. Mitsyn, T. Sapozhnikova, A. Sapozhnikov, O. Streltsova, F. Buzatu, M. Dulea, I. Vasile, A. Sima, C. Visan, J. Busa, I. Pokorny, *Performance assessment of the SIMFAP parallel cluster at IFIN-HH Bucharest*, *Romanian Journ. Phys.*, **53** (2008) 665-677.

[6] A. Ayriyan, Gh. Adam, S. Adam, E. Dushanov, V. Korenkov, A. Lutsenko, V. Mitsyn, O. Streltsova, *Performance assessment of JINR CICC supercomputer*, *Proceedings of XII Scientific Conference of JINR Young Scientists and Specialists*, JINR Dubna, ISBN 978-5-9751-0045-0 (2008), pp. 71-74.

[7] Gh. Adam, S. Adam, A. Ayriyan, V. Korenkov, V. Mitsyn, M. Dulea, I. Vasile, *Consistent performance assessment of multicore computer systems*, *Romanian Journ. Phys.*, **53** (2008) 985-992.

[8] *http://supercomputers.ru/?page=archive&rating=9*