ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Survey of Application Benchmark on Accelerated Hybrid Architecture

Nisha Bonde, Prof. Y. B. Gurav

Dept. of Computer Engineering, Padmabhooshan Vasantdada Patil Institute Of Technology, Computer Engineering Department Pune , India

ABSTRACT: Due to the rapid advancements and technology refreshes in computer architecture, it is becoming increasingly difficult to compare the real life performance of large computer systems based on specifications. Performance comparison of multicore processors and hardware architectures are done with the help of specifically designed benchmark tests. Computer systems are designed by manufacturers keeping in mind the full exploitation of the underlying hardware by the users, thereby resulting in high performance in the benchmark tests. HPC programmers are constantly looking for: (1) Performance – on all the systems they work on (including future ones too). (2) Portability – none like rewrites just because a new architecture comes along. (3) Maintainability – code should be clean, such that it can be changed without huge efforts, preferably by the scientists who are directly working with the applications. The traditional method used is dense linear algebra to evaluate the performance of multi-core processor based hardware accelerators.

KEYWORDS: Benchmark,High Performance Computing,Intel Xeon Phi coprocessor and Intel Xeon coprocessor.

I. INTRODUCTION

Most of the computer architectures make stronger demands than ever on writers of scientific code. The cost of moving data is increasing day by day as compared to the cost of computing [1]. Earlier, High Performance Computing essentially focused on a single box large scale system. But design of such mammoth systems will lead to development of architectures that are different and extremely proprietary from other contemporary architectures. There are a number of case studies, but are fairly small and conceptual, or they are very specific and hard to compare and reproduce. This paper highlights the different problems in the current designs as well as the benchmarks for comparing and evaluating tools for hybrid system design.

II. LITERATURE REVIEW

Benchmark is used to test a number of independent attributes of the performance of HPC systems. These provide an insight to hardware engineers for various architectural design considerations. Research in the area of benchmarking is increasing due to advancement of distributed memory computers as well as many cores packed onto a single die to address grand challenge problems. Sometimes, hybrid verification method is used in which the same problem is evaluated using different verification methods [2].

The list of top 500 high performing machines is generated twice a year to indicate latest high performance machines LINPACK benchmark is one of the methodologies used to measure the computer's floating point compute power. This benchmark is developed to determine how much amount of time is required by the system to solve the problem using LINPACK package. In addition to this, LINPACK parallel benchmark is developed for processing the data in parallel. Earlier LINPACK benchmarks were able to solve the problems with fixed size matrix, which were then later developed to solve problems with arbitrary size. As a result of this, the LINPACK benchmark is used to compute the performance of first 500 high performance computers [6].

As per the November 2015, top 500 supercomputers list, Tianhe 2 system is the fastest computer with peak of 54.9 Petaflop and a sustained of 33.8 Petaflop Linpack benchmark. The system has 31,20,000 hybrid cores, which

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

include Intel Xeon Phi and CPU. This depicts that for extreme scale computing, commodity processors coupled with computational accelerators are the way forward. [11].

Intel(R) Xeon Phi Coprocessor, Knights Corner (KNC), is a PCIe card and GDDR5 memory and offers very high memory bandwidth based on the many integrated Core (MIC) architecture [4].

Additionally, chip multiprocessors (CMP) supporting 4 to 32 hardware threads are available commercially and compute accelerators such as Graphics processing Units (GPUs) are also available which support 1000s of hardware threads [5].

Hybrid systems play a pivotal role in the current supercomputing paradigms and the systems are usually designed to achieve the highest possible performance in terms of the number of 64-bit floating-point operations per second (flops).

In current scenario, the effort and cost of moving data is much more than that of computing. Many recently installed clusters couple commodity processors to co-processors such as the Intel Xeon Phi Knights Corner' co-processor (KNC). Design of iterative solvers on extreme scale architecture, therefore becomes difficult. For example, Lattice Quantum Chromo dynamics (Lattice QCD) is the iterative solver of the Dirac equation. The Lattice QCD community was interested in porting their code onto the new powerful architecture. The initial implementation of Lattice QCD on KNC represents that it is possible to obtain high performance.

The KNC architecture is used which provides high performance without requiring new programming model and API, language or restrictive memory model. KNC has 61 cores on a single die, of which 60 are used for compute. Multi cycle instruction latency of each core is provided by four way hyper-threading support. These cores have minimal single threaded instructions which help to increase the area and power efficiency of a core. The vector units are 512 bits wide and are able to execute 8 wide double precision and 16 wide single precision single instructions multiple data instructions in a single clock. Following figure 1 shows the KNC architecture. Each core of KNC consists of simultaneous multi threading, vector processing unit, data and instruction caches and coherent caches. KNC supports multiple scalar and vector operations due to the rich instruction set present for it [1].

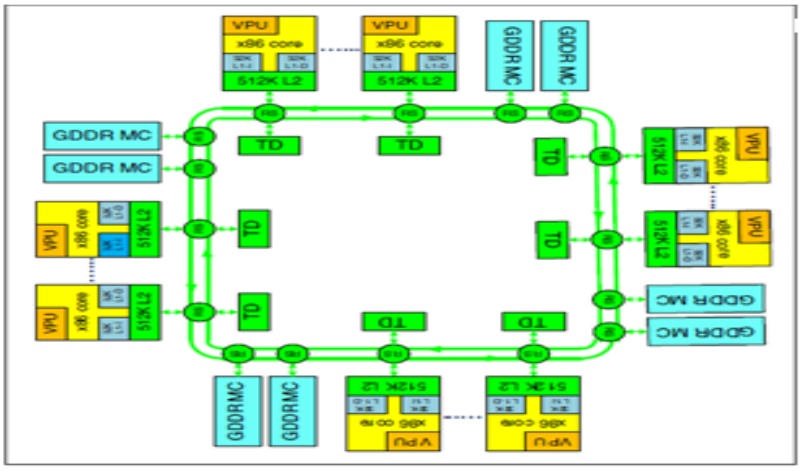

Figure 1: KNC Architecture

Another LINPACK implementation is on pure CPU Sandy Bridge EP based Architecture. This architecture supports 2 way hyper threading with multi core server architecture. It has 256 bit-wide SIMD unit that executes the AVX instruction set. Following figure 2 shows the Sandy Bridge EP Architecture [3].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

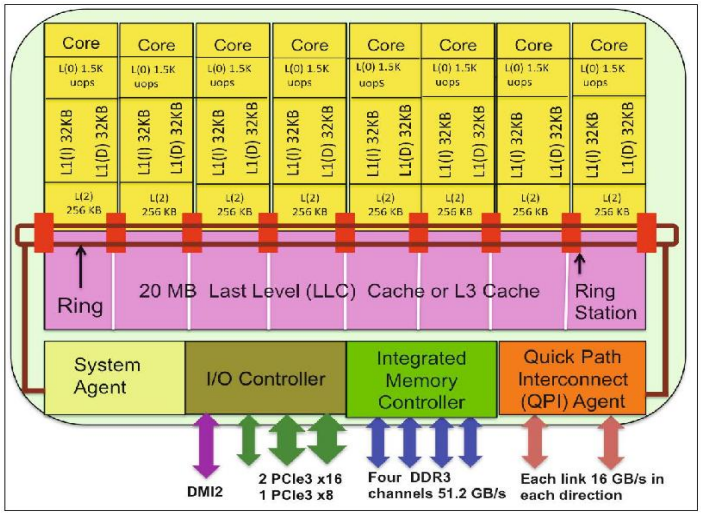

Figure 2: Sandy Bridge EP Architecture

A Sandy Bridge-based node has two Xeon E5-2670 processors, each with eight cores. Each processor is clocked at 2.6 GHz, with a peak performance of 166.4 Gflop/s. The total peak performance of the node is therefore 332.8 Gflop/s. Such nodes act as a base platform for accelerators of minimum 1 TF peak capacity.

Sparse matrix vector multiplication has an equal importance in computational science and their structures are present in various computational activities. It is highly irregular compared to dense linear algebra. Alternate architectures are used such as GPU, which give desired performance for Sparse Matrix Vector Multiplication. These architectures are also called as throughput oriented architectures [7]. These throughput oriented processors give very high computational throughput. Most of the sparse matrix vector multiplication kernels are implemented using CUDA language. Methods of optimization of sparse vector matirx multiplication is different for different types of processors. Many core processsors avoids memory latency by using hardware multithreading. As a result, manycore processors demand a high degree of parallelism [7].

The excellent performance of accelerators especially vectorization along with limited space and power consumption has attracted the high performance computing community. Large scale applications can be executed using the parallel computation.

As per the top 500 supercomputers of November 2015, the ORNL Titan system is the second highest performing system, which is heterogeneous too. This system consists of 5, 60,640 processors with AMD 16 core systems as base servers along with GPU's as accelerators. In order to achieve a higher peak performance, CUDA processors are used in hybrid clusters. Highly parallel multithreaded multi processors with very high floating point performance and memory bandwidth is used in GPU architectures. The problems of data parallel computations can be solved by using GPUs and the parallel capabilities of GPUs are exposed by using CUDA language, which is a parallel programming model and software environment [9].

III. CONCLUSION AND FUTURE WORK

The paper has described various application benchmarks on Intel's recently released Intel(R) Xeon Phi co-processor (code-named Knights Corner), and Intel(R) Xeon(R)E5 2670 (Sandy Bridge EP) i.e. a hybrid configuration. The advanced offload DGEMM as well as the advanced look-ahead improves the performance and efficiency of hybrid LINPACK implementation. In future, efficient Linpack has to be designed on the upcoming Knights Landing many core processor architecture, which shall be placed on the main board with exclusive memory allocation. Another activity is development of scripts which can put unutilised cores into a deep sleep state to significantly reduce the total energy consumption.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

REFERENCES

1. Simon Heybrock, Bálint Joó, Dhiraj D. Kalamkar, Mikhail Smelyanskiy, Karthikeyan Vaidyanathan, Tilo Wettig Pradeep Dubey," Lattice QCD with Domain Decomposition on Intel Xeon Phi Co processors", ICHPC.
2. J. D. McCalpin. The STREAM Benchmark
3. Alexander Heinecke, Karthikeyan Vaidyanathan, Mikhail Smelyanskiy-Design and Implementation of the Linpack Benchmark for Single and Multi-Node Systems Based on Intel(R) R Xeon Phi TM Coprocessor 2013 IEEE 27th International Symposium on Parallel and Distributed Processing.
4. Intel Math Kernel Library (Intel(R) MKL) 11.0, 2012
5. F. Song and J. Dongarra. A scalable framework for heterogeneous gpu-based clusters. In Proceedings of the 24th ACM symposium on Parallelism in algorithms and architectures, SPAA '12, pages 91 to 100, New York, NY, USA, 2012. ACM.
6. www.top500.org. TOP500 list, June 2012 release. 2012.
7. N. Bell and M. Garland. Implementing sparse matrix-vector multiplication on throughput-oriented processors. In Proc. ACM/IEEE Conf. Supercomputing, SC'09, pp. 18. ACM, 2009.
8. T. Endo, A. Nukada, S. Matsuoka, and N. Maruyama. Linpack evaluation on a supercomputer with heterogeneous accelerators. In IPDPS, pages 1 to 8. IEEE, 2010.
9. Massimiliano Fatica, "Accelerating LINPACK with CUDA on Heterogeneous clusters", GPGPU '09 Washington DC, USA
10. Accelerators in HPC – Having the Cake and Eating It Too Written by Michel Müller, Typhoon computing.
11. www.top500.org