Advanced Architecture Test Beds

Suzanne M. Kelly, James A. Ang, and James H. Laros III
Sandia National Laboratories*

smkelly@sandia.gov, jaang@sandia.gov, jhlaros@sandia.gov

ABSTRACT

As part of NNSA's ASC program, Sandia National Laboratories is addressing a critical need for experimental architecture test beds to support path-finding explorations of alternative programming models, architecture-aware algorithms, low-energy runtime and system software, and advanced memory subsystem development. The new advanced architecture systems are being used to develop Mantevo proxy applications, enable application performance analysis with Mantevo proxy applications, support heterogeneous computing and programming model R&D projects, and for Structural Simulation Toolkit (SST)[1] HPC architectural simulation validation efforts. This paper describes the available platforms, their environment, current activities, and future plans.

Keywords

Exascale, Co-design, Computer Architecture, Test Beds, Programming Models, HPC Architectural Simulation, System Software, HPC.

1.0 Problem Description

The transition from single-core to multi-core processor technology, and the advent of heterogeneous compute node architectures and accelerators, coupled with the continually increasing demand for more computing cycles, is necessitating we explore revolutionary changes to our HPC foundation. We are seriously questioning how much further an MPI-everywhere programming model can scale. Power usage curves of the largest systems are reaching practical limits and are untenable when projected to Exascale requirements. Current algorithms and solvers must be reconsidered in light of many-core and accelerator technology. Even applications themselves may need to be reworked to some degree to take advantage of the huge node and platform level parallelism that we anticipate. None of these issues will suddenly appear on an Exascale-class system. Instead, they will become more apparent as we continue on the path to Exascale. Now is the time to explore what and by how much do things need to change.

A logical first step is to study the aforementioned challenges, as well as others, on the key building block of an HPC system—the node. Sandia has obtained several state of the art test beds, ranging in size from one node to a few racks of nodes. At the present time, it is more important to explore a diverse set of architectural alternatives than to push large scale. These alternatives offer various methods designed to increase node level computational ability, all presented as increased node-level parallelism of some type. The systems represent various hardware architectures and are described in Section 4.0.

2.0 Current and Expected Use Cases

Prior to procuring these systems, we identified the likely scenarios for how they would be used. We envisioned them as enablers for co-design so that hardware and software communities could share the resources and communicate their findings on a common platform. They are not production resources, but intended for "test pilot" users that understand the experimental nature of these test beds.

The major focus areas were expected to be:

- Alternative programming models
- Architecture-aware algorithms
- Low-energy runtime and system software
- Advanced memory sub-system development

The Mantevo mini-applications[2] are also co-design enablers as these proxies for ASC applications can be more easily used by non-application developers, such as computer

* Sandia National Laboratories is a multiprogram laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the United States Department of Energy's Nuclear Security Administration under Contract DE-AC04-94AL85000.

[1] http://sst.sandia.gov/

[2] https://software.sandia.gov/mantevo/index.html

architects and system software developers. They play an important role on the test beds. This is particularly true when they do not have to adhere to the same access control restrictions that are levied on the applications they emulate.

HPC architectural simulation also plays a key role in co-design. Validation efforts of Sandia's SST can be done using the advanced architecture processors and memory systems on the test beds. Once again, Mantevo applications are useful in this process. The same mini application can more easily be run on both SST and a test bed.

The first efforts on the test beds have therefore focused on porting the Mantevo applications to the platforms. And where possible, the results are compared to the full application. Subsequently, alternative programming models (e.g. MPI plus OpenMP) are being implemented in the mini applications to assess their impact.

We are beginning to ramp up activities related to power research and system software. Since some of the systems have solid state storage devices attached to each node, we also plan to explore opportunities for their use. In particular, a study of a memory hierarchy is likely. However, it may also be studied as an I/O device.

Some of the test beds will be undergoing hardware upgrades in the near future. We are planning a centralized file system. Currently, every system has independent storage devices.

3.0 Network Environment

The lack of a common file system resulted from the selection of the network environment for a number of the test beds. In order to serve as large of a community as possible, a number of the systems reside in a newly created external collaborative network that is Internet-facing and is also available to non-US citizens. Because of its newness, the supporting infrastructure and resources are limited. We hope to rectify this over time. Export controlled applications are not allowed on systems in this collaborative network.

4.0 Description of Current Systems

We now describe the seven systems being managed as part of our advanced architecture test bed project.

4.1 *Arthur – An Intel® Many Integrated Core (MIC) test bed with Knights Ferry co-processors*

The machine with hostname *arthur.sandia.gov* is a 42 node cluster that was integrated by Appro International. Each node has:

- two six-core (Westmere-EP) Xeon® 5600 processors running at 3.46 GHz
- 24 GB Double Data Rate (DDR)3-1600 MHz memory
- two 30-core Knights Ferry software development cards running at 1.05 GHz. Each card has 2 GB Graphics (G)DDR5-1800MHz memory
- one 80GB Intel® Solid State Disk (SSD) Serial Advanced Technology Attachment (SATA) 3Gb/s MLC NAND flash drive

The interconnection network is a Mellanox Infiniscale IV Quad Data Rate (QDR) Infiniband. There is a separate Ethernet system management network.

Figure 1 – Arthur Rack: One of Seven

The Westmere processors will be upgraded to the E5 (Sandy Bridge) family of processors in the second quarter of 2012. We plan to replace the Knights Ferry cards with pre-production Knights Corner co-processors in the third quarter of 2012.

4.2 *Ferry and Ferry2 – Single node Intel® Knights Ferry systems*

We received early single node Intel® systems for evaluation of node specific capabilities. They reside in two different network environments, one of which supports processing of export controlled applications and data.

4.3 *Teller--An AMD Llano Fusion™ Cluster*

Teller.sandia.gov, integrated by Penguin Computing, is a 104 node cluster where each node is configured as follows:.

- one single socket Llano Fusion™ Accelerated Processing Unit (APU) which integrates:
- one quad-core AMD K10 X86 running at 2.9 GHz
- a 400-core Radeon HD 6550D at 600 MHz
- one 256 GB Micro C400 SSD SATA 6 Gb/s MLC NAND Flash drive
- 16GB DDR3-1600 MHz memory (four have 8GB DDR3-1866 MHz)

The Interconnection Network is QLogic Quad Small Form Factor Pluggable (QSFP) QDR Infiniband. Again, there is a separate Ethernet system management network. The large capacity of the flash drives on *Teller* is of particular interest for I/O explorations. Discussions are underway with AMD and Penguin regarding the potential to upgrade the Llano APUs with Trinity APUs

Figure 2: Teller Rack: One of Six

4.4 *Curie – A Cray XK6 test bed*

Curie.sandia.gov has 52 compute nodes. This is not an advanced system per se, as it already exists as a product and has been exercised at high scale on Oak Ridge Leadership Facility's *Titan* system. Nonetheless, it offers useful architecture features, such as the ability of its management network to collect power usage metrics. Each compute node has

- One AMD Opteron™ Interlagos sixteen-core 6272 processor running at 2.1 GHz
- 32 GB DDR3-1600 memory
- One NVIDIA® Fermi accelerator with 6.0 GB GDDR5 memory

The interconnection network is Cray's Gemini custom network. Curie also has an out of band Ethernet system management network and embedded management controllers. Discussions are underway with NVIDIA® and Cray regarding potential upgrades to replace the Fermi GP-GPUs with Kepler GP-GPUs.

4.5 *Watt – A Cray CX-1 Heterogeneous Cluster*

Watt.sandia.gov is a ten-node cluster comprised of a four blade (node) Cray CX-1 and six 1U Microway Graphics Processor Unit (GPU) enabled nodes. Each Cray CX-1 blade (node) consists of:

- two Intel Nehalem X5550 quad-core 2.67 GHz processors
- 24GB DDR3-1333 MHz memory
- one NVIDIA® Tesla C1060 GPU with 4 GB GDDR5 memory

Each of the six 1U Microway nodes is configured with:

- two Intel Nehalem E5520 quad-core 2.27 GHz processors
- 12 GB DDR3-1333 MHz memory
- two NVIDIA® GPUs, either Tesla C1060 with 4 GB GDDR5 memory or Fermi C2070 with 6 GB GDDR5 memory

All blades and nodes are connected via an Infiniband Double Data Rate (DDR) 12 port switch embedded in the Cray CX-1. All nodes are also connected via the Ethernet management switch, also embedded in the CX-1. Bright Cluster manager is used to support this cluster and is also currently being evaluated.

4.6 *Wingus – Convey HC-1ex*

This single node system resides in Sandia's internal restricted network that allows processing of export controlled information. The board has:

- One Intel® Nehalem Quad-core X86 and 2.13GHZ
- Four Xilinx Vertex6 LX760 Field programmable Gate Array (FPGA) Co-processor
- Eight Xilinx FPGA for programmable memory controllers that support 16 channels of Convey-designed Scatter-Gather DIMMS

The Convey test bed is available for proxy application development and testing, but our primary efforts have been focused on multi-threaded runtime system software development. Convey's advanced memory subsystem is useful for quantifying the impact of a node architecture that places priority on breaking down the memory wall.

Discussions are underway with Convey on their next generation node design.

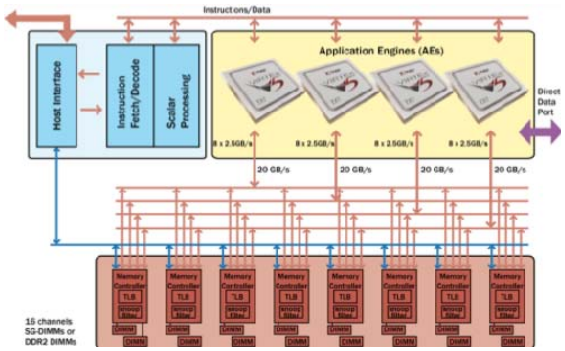

Figure 3: Schematic of a Convey GV HC-1EX board

4.7 *Tilera TILE-Gx36*

The system has four Gx8036™ nodes that are connected by 4x10 gigabit Ethernet low latency network interfaces. Each node consists of:

- 36 1.2 GHz cores
- 16 GB DDR3-1333 MHz memory with 9 MB coherent L3 cache
- 64 bit MIPS-derived Very Long Instruction Word (VLIW) instruction set
- 256 KB Level 2 cache per core
- 32 KB Level 1 instruction cache per core
- 32 KB Level 1 data cache per core

It has a user-accessible iMesh™ network on a chip.

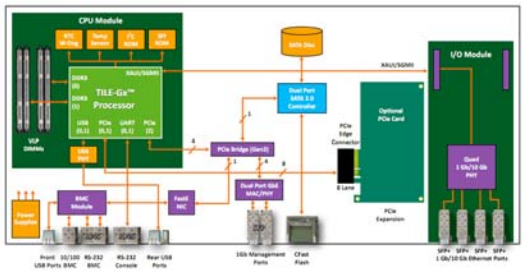

Figure 4: Tilera GX block

5.0 Conclusions and Future Plans

The six test beds described in this paper represent the fairly wide variation of architectural offerings that have potential to impact next generation platforms. While we cannot, at this time, anticipate exactly what an Exascale platform will look like, we do have confidence that a huge amount of parallelism will have to be leveraged to achieve an ExaFLOP within practical power constraints. We can also speculate that node-level parallelism will be a very important factor and will possibly present the greatest challenge, and hopefully opportunity, to our existing applications.

Intel MIC, AMD Fusion and discrete GPU platforms like the Cray XK6 (NVIDIA) all require high levels of parallelism to exploit the available computational capability. The MIC architecture leverages many (10s of) low frequency general-purpose x86 cores and presents a more traditional cache coherent memory space. The Cray XK6 loosely integrates few high frequency general-purpose x86 processors with a discrete NVIDIA GPU, which contains a very large number (100s) of simple cores optimized for Single Instruction Multiple Data (SIMD) parallelism. The AMD Fusion architecture presents a similar heterogeneous environment where the general-purpose x86 cores and the simple SIMD cores are more tightly integrated on the same chip. While each architecture exposes a large amount of node level parallel processing capability, each must be exploited in different ways and each, at this point, has strengths and weaknesses. It is critical that each approach is investigated and providing a test platform for researching new programming paradigms will be critical in achieving our Exascale goals.

Investigating node-level parallelism, while critical, is not the only factor. Investigating advanced networking and memory interfaces and techniques is equally important. The Tilera and Convey architectures are allowing us to investigate some of these critical platform considerations.

Finally, power is well recognized as a critical factor in achieving our future goals. The Cray architecture has been leveraged in the past to investigate how we can improve application energy efficiency and the XK6 test bed will be used to further this research by investigating energy use on GPUs. We are additionally designing a power monitoring capability with Penguin computing that will enable fine-grained component level measurement of not only our test bed platforms but future commodity clusters.

We are executing a number of technology refresh options on our current test bed platforms, which will allow us to evaluate emerging capabilities in advance of general availability. These advanced architecture test beds support quantitative performance measurements, experience with

future programming paradigms, and advanced system software development. We will continue to seek opportunities for analysis and testing with advanced architectures that help us characterize the path to Exascale.

Acknowledgement

We are grateful for our industry collaborators: Intel®, AMD, Appro, Penguin, Convey, Tilera and Micron.