Integration and Development of the 500 TFLOPS Heterogeneous Cluster (Condor)

Mark Barnell, Qing Wu and Ryan Luley
Air Force Research Laboratory
Information Directorate
Rome, New York, USA

ABSTRACT
The Air Force Research Laboratory Information Directorate Advanced Computing Division (AFRL/RIT) High Performance Computing Affiliated Resource Center (HPC-ARC) is the host to a very large scale interactive computing cluster consisting of about 1800 nodes. Condor, the largest interactive Cell cluster in the world, consists of integrated heterogeneous processors of IBM Cell Broadband Engine (Cell BE) multicore CPUs, NVIDIA General Purpose Graphic Processing Units (GPGPUs) and Intel x86 server nodes in a 10Gb Ethernet Star Hub network and 20Gb/s Infiniband Mesh, with a combined capability of 500 trillion floating operations per second (TFLOPS). Applications developed and running on CONDOR include large-scale computational intelligence models, video synthetic aperture radar (SAR) back-projection, Space Situational Awareness (SSA), video target tracking, linear algebra and others. This presentation will discuss the design and integration of the system. It will also show progress on performance optimization efforts and lessons learned on algorithm scalability on a heterogeneous architecture.

INTRODUCTION
The Affiliated Resource Centers (ARCs) are Department of Defense (DoD) Laboratories and Test Centers that acquire and manage High Performance Computing (HPC) resources as a part of their local infrastructure, but share their HPC resources with the broader DoD HPC user community via the High Performance Computing Modernization Program (HPCMP) which coordinates allocation of their HPC resources. In order to provide tomorrow’s Air Force with massively parallel and scalable HPC applications, the software must be developed on large clusters. Unlike typical HPC clusters, all AFRL/RI clusters allow for interactive development and testing. In 2010, the AFRL Information Directorate won a two-million-dollar project, sponsored by the HPCMP, and built the Condor Cluster, which is DoD’s largest interactive super computer as of November 2011. The Condor Cluster consists of 84 Servers (2U Dual six-core Intel Westmere 5660, 24 or 48 GB RAM) each with 2 GPGPUs (NVIDIA C1060, C2050 or C2070s) [1]. The heterogeneous cluster has 22 PlayStation 3s (PS3s) connected to each of the 78 server nodes (1716 PS3s in total).

Figure 1. Condor Cluster: DoD’s largest interactive HPC.

The long-term goal of AFRL/RI’s high performance computing research is to provide the warfighters with Secure Embedded HPC (SEHPC) of the highest computing performance, under the Size-Weight-and-Power (SWaP) constraints. At the time when it was built, Condor was the largest, fastest and most energy-efficient interactive HPC in the Department of Defense.

The Condor HPC integrates the vast majority of the state-of-the-art HPC processing and networking architectures into one coherent functional system. This provides great R&D potentials and opportunities for the users so that they can explore and experiment with not only any single parallel computing architecture, but also any combinations of architectures, and evaluate their computing/communication performance and SWaP efficiencies under different programming and application scenarios. For processing architectures, the Intel Xeon server represents the multi-processor, super-scalar architecture; the NVIDIA Tesla GPGPU combines architectures of many-core, single-instruction-multiple-thread (SIMT, similar to SIMD), and streaming processing; the PlayStation 3 uses the IBM Cell BE processor, which
adopts the multi-processor, single-instruction-multiple-data (SIMD, or vector processing) architecture. These three processors represent most of the modern high-performance processor architectures and cover a wide range of trade-offs among performance, power, size and weight.

**DESIGN IMPLEMENTATION AND CONSTRAINTS**

The Condor application development focuses on two related ongoing programs, one applied research effort and one basic research effort. The applied research focuses on voluminous generation of synthetic aperture radar (SAR) images providing persistent surveillance of city-sized areas with 1Hz update rate yielding a previously unachievable “video SAR capability” previously unachievable. The basic research effort investigates massively parallel neuromorphic architectures that can exploit the video SAR outputs, or alternative high resolution video cameras, to deliver robust perception, anticipation, and focus of attention.

The scalability and parallelism required to achieve sustained high computational throughputs demand low latency high bandwidth networking architectures. The Condor server nodes (custom built 2U X86 servers) were designed with both 20 Gb/s Infiniband and dual 10GbE network interface cards. This required the motherboard to support 48 PCI-E Gen2 (two Intel 5200 chipsets, 2x IOH-36D), allowing for four 16x Gen 2 slots. This supports maximum data throughput to all four PCI-E devices: two NVIDIA GPGPUs and the two network cards.

In a star-hub topology, 39 IBM BLADE RackSwitch G8000 Gigabit Ethernet spoke switches are connected to the PS3 compute nodes and aggregated to 12 RackSwitch G8100 10 Gigabit Ethernet switches. Dual 10 Gigabit Ethernet links are bonded for high-bandwidth switch-to-switch communications. The IBM BLADE RackSwitch G8100s are connected to the Condor server nodes. The IBM BLADE RackSwitch G8100’s CX4 transceivers ensure low transmission latency with an average of 60 to 70 microseconds even when going through three switches.

The condor server nodes can also communicate between each of the 78 nodes through an Infiniband mesh. This allows for very low latency and high bandwidth when applications only require the x86 processors and GPGPUs. While running benchmark tests and network OpenMPI applications, we routinely achieved a sustained 25-28 Gb/s performance across the entire network.

The design of the Condor HPC system had physical constraints and limitations. As shown in Figure 5, the actual footprint of the system, layout, power and cable trays were chosen carefully to allow for maximum cooling and minimum cable length.
ENERGY-EFFICIENT INTERACTIVE SYSTEM

Deployment and development of the Condor supercomputer was configured for two primary objectives: interactive (on-demand) and energy-efficient (green) computing. Interactive computing provides the users with direct access to the resources based on their schedule and scalability needs [2]. When the applications and software development activities use only a portion of Condor, the rest can be put in shutdown or put to sleep mode for significant energy savings. This has major impacts on the facility’s infrastructure and costs.

The current 100+ Condor users can login into one of six login servers and begin by reserving server nodes and PS3 clusters. Figure 6 shows the Condor status and reservation system as web-based user interface.

The PS3s are configured with Fedora 9 or Yellow-Dog Linux (YDL) and included with the bootloader and operating system is the wake-on-LAN option. This option allows all 1716 PS3s to be put in a power savings mode (sleep). A PS3’s typical idle power draw is 95 watts and 5 watts in sleep mode. The PS3s will consume 67 percent of the total 256 KWs when the entire Condor cluster is operational. The systems reservation mirrors the power draw is shown in Figure 7. The typical HPC system will run all of the nodes in idle mode, using up to 70% of the peak system power. Condor typically runs around 40% of peak during the work week, and 18% on the weekends. The estimated power cost saving is $219,964.00/yr and this achieves a reduction of 792 tons of carbon footprint on the environment [3].

PREPARING FOR THE FUTURE

Large scale computing systems provide the basis to investigate and implement solutions for C4ISR challenges. Fundamental for many of the Data-to-Decision problems is the ability to perceive, fuse, and exploit information within voluminous flows from increasingly capable and affordable sensors monitoring the air, space, and cyber domains. Signal and image processing, such as creating the video SAR capability, present significant computational loads near the sensor which then feed the even more challenging tasks of recognition, information fusion, tracking, and exploitation based upon this flood of imagery. HPC systems and the Condor cluster support
basic research into massively parallel neuromorphic models at scales approaching that of the human neocortex for robust visual perception and recognition (Figure 8).

We continue to expand our HPC portfolio and relationships with HPCMP and tailor our capabilities to solve significant Air Force challenges. Embedded HPC systems will be developed and integrated close to the sensor, enabling processing of high volume data with greatly improved information content. We are developing hybrid scalable computing framework for imagery information exploitation, real-time and autonomous sensing and deciding technologies on our Condor cluster. The scalable computing framework will be robust enough to run on tomorrows HPC architectures (Figure 9).

**CONCLUSION**

We have presented an interactive HPC supercomputer, Condor, which has been developed and designed to be energy-efficient and interactive with users. Condor provides the Air Force and the DoD community the ability to prototype, develop and evaluate large-scale massively parallel HPC applications.

**ACKNOWLEDGMENTS**

The contractor’s work is supported by the Air Force Research Laboratory, under contact FA8750-10-C-0216. Any Opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of AFRL or its contractors.

**REFERENCES**

