IN11E: Architecture and Integration Testbed for Earth/Space Science Cyberinfrastructures

A Future Accelerated Cognitive Distributed Hybrid Testbed for Big Data Science Analytics

Milton Halem¹, John Edward Dorband¹, Navid Golpayegani², Smriti Prathapan¹, Yin Huang¹, Timothy Blattner³ and Kaushik Velusamy¹

(1)University of Maryland Baltimore County, Computer Science, Baltimore, MD, United States,
(2)NASA Goddard Space Flight Center, Greenbelt, MD, United States,
(3)National Institute of Standards and Technology Gaithersburg, Gaithersburg, MD, United States
What is Cognitive computing?

- **Cognition** is the process of acquiring knowledge and understanding through thought, experience, observations and use of other senses; It encompasses integrating reasoning, memory, language and tools for attaining knowledge from information.

- **Cognitive computing** is the simulation of the human learning processes by employing accelerated computerized component architectures that mimic the human brain. The components address problems that are characterized by ambiguity and uncertainty.

- **Cognitive systems** integrate diverse sources of massive information using machine learning algorithms to find patterns and then apply those patterns to respond to the needs.

- **Cognitive applications** such as machine learning to regress flux measurements from long records or to integrate a variety of sensor data, reanalysis model output along with propositional based rules to infer trends, issue warnings or recommendations to decision makers.
Basic general purpose compute cluster
The Bluewave Cluster

- Originally known at NASA Center for Climate Simulation (NCCS) as Discover Series Scalable Unit 8, was competitively awarded by NASA under the Congressional Stevenson- Wyndler Surplus System Act and came to CHMPR at UMBC in late 2013.

- The acquired NCCS system consisted of 6 racks with 512 nodes, of which 4 racks are employed by UMBC/CHMPR. Now called Bluewave, the system currently consists of 336 nodes which includes management nodes, storage nodes, and SLURM client compute nodes.

- Most nodes are running a Linux image based CentOS 6.5 system. Each node has 2 quad core Intel Nehalem processors (Xeon X5560) and 24GB RAM and 250GB of local storage.

- Bluewave includes a 32 node Hadoop cloud sub-cluster with each node containing 1TB of local disk storage. In addition, 4 nodes with 2 oct core Intel Sandy bridge and 4 Intel Many Integrated Core (Phi) co-processors are currently being integrated into the system.

- The cluster is used both for research by graduate students and faculty for integrating and testing high performance specialized prototyping computing components for Big Data Analytics.
A State-of-the-Art ‘Accelerated Cognitive Testbed’ (ACT) for CHMPR Big Data science

Acquired prototype of the next gen IBM 250 Petaflops summit processor, two IBM” Minsky” nodes each with dual Power 8+ processors with 10 cores at 3.0 GHz with 1TB ram and four Nvidia Tesla P100 GPUs (3584 cores) with 2 Nvidia's NVLinks from the power 8+ to each pair of GPUs have been integrated into Bluewave.

In addition, an IBM Flash system A 900 with 30TB of flashram expandable to 40 TB are connected to the Minsky nodes with a Coherent Accelerator Processor Interface (CAPI). The flashram A900 system supports both nodes and partitioning of flash ram is programmable.

An additional 8TB of SSD through an NVME card is available as local storage.
Acquired Two IBM Minsky Nodes

Each Node contains:
2 Upgraded Power 8 processors,
10 cores and 8 threads/core
4 Nvidia P100 GPUs with 2 Nvidia Links
1TB of DRAM
3.2 GHz Non Volatile Mem card
4 SSDs with 4 TBs each
Memory bandwidth 115 GB/s
Infiniband

In addition:
CAPI (Coherent Accelerator Processor Interface FPGA)
20 TB Flash Memory
Nvidia’s First GPU-to-GPU and GPU-to-CPU High-Speed Interconnect

• Accelerated computing with increasing numbers of highly dense GPU server nodes has become the de-facto standard for attaining HPC ratings.

• Interconnect bandwidth between GPUs is the most significant bottleneck to application performance.

• NVLink is the first high-speed interconnect for NVIDIA GPUs to solve the interconnect problem. Four NVLink connections per GPU can deliver a total 160 GB/s bidirectional bandwidth for the Tesla P100.

• This is over 5X the bandwidth of PCI Express Gen3.
**Figure 1:** The Tesla P100 significantly exceeds the compute performance of past GPU generations.

**Figure 2:** The Tesla P100 with HBM2 significantly exceeds memory bandwidth of past GPU generations.
The Network-Attached Storage (NAS)

- The NAS has 3 components.
- Store1 – 43 Terabytes
- Store2 - 21 Terabytes
- Store3 - 43 Terabytes

Each node in the Bluewave can access the above 3 stores for external persistent storage.
IBM Flashram Storage System – 30TB
IBM Flashram as Cognitive Memory

• Did not use traditional SSDs off the shelf, but created their own flash modules.
• Flashram is 3x-4x faster than SSDs since it is directly tied to motherboard similar to DRAM.
• Flash and SSDs are doing to hard drive arrays what hard drive arrays did to tape.
IBM claims “All-flash arrays are “cognitive in design and operation,” and include intelligent storage automation that enables data to move from flash to other storage media within an array. Moving data from tier-to-tier occurs by the storage management system which evaluates and learns data access usage patterns.”
Seagate: Holy Grail of Active Storage Research

- Acquired object oriented key value Seagate kinetic disk system consisting of 2 chassis each with 12 disks. Each disk has 4TB and its own ethernet cable, IP address and local processor. Total storage 96TB.

- Harnessing processor to allow an object storage platform to delegate intelligent functionality to the drives. Could run their own virus scanning, content discovery, and compression functions.

- Developed Lightweight Virtual File System (LVFS) which separates data retrieval from metadata information.

- Performed first ever satellite processing directly on Active Kinetic Disk using the MapReduce parallel programming model for gridding 2 years of global OCO-2 CO2 data.

- Interfacing active kinetic disks to IBM iDataPlex for Image Stitching, Deep belief Nets, Variable block size for out of core matrix multiplication.
REMOTE ACCESS TO THE D-WAVE 2X

• CHMPR has access to the D-Wave 2X at ARC at NASA. D-Wave 2X is the only commercially available quantum computer.

• We view the D-Wave not as a general purpose, but rather designed to be a hybrid quantum annealing coprocessor.

• D-Wave 2X consists of ~1152 qubits. 8 X 12 X 12 Qubit arrays and announced at SC16 the D-Wave 3X with ~ 2048 qubits to be delivered in 1st quarter 2017.

• Plan to couple Bluewave to D-Wave 3X over Internet 2 as an accelerator for optimization.
PGI OpenACC and NIST HTGS S/W

- Installing Portland Group Inc (PGI) OpenACC software for implementing C, C++ and Fortran 90 onto GPUs.

- Implementing a Hybrid Task Graph Scheduler developed at NIST to improve programmer productivity by implementing and optimizing parallel algorithms to fully utilize the multicore CPUs and multiple GPUs, while managing dependencies, locality of data and overlapping data motion with computation.
Summary

• Presented an existing accelerated cognitive distributed hybrid testbed ready for “Big Data Science Analytics”.

• We wish to publicly acknowledge the following organizations for enabling the provisioning of this prototype cognitive architecture a UMBC for future Big Data Science:
  NASA/GSFC, IBM, Seagate, D-Wave, NASA/AIST, NSF/IUCRC.
Thank You