FPGA/GPU Cluster

The FPGA/GPU cluster is a cloud-based, remotely accessible compute infrastructure specifically designed to accelerate compute intensive applications, such as machine learning training and inference. Latest state of the art acceleration technologies including FPGAs, GP-GPUs and massively-parallel processing units, closely coupled with server processors constitute the backbone of this cluster. The software stack consists of a complete ecosystem of machine learning frameworks, libraries and runtime targeting heterogeneous computing accelerators.

Heterogeneous Compute Cluster
– 8 nodes

Node configuration:

- Dual 12 core 2.2-3.0 GHz CPU
- 192GB RAM
- 300GB local storage
- 100Gb EDR node interconnect
- 10GbE storage network
- **Xilinx Alveo U200 FPGA**
- **NVIDIA V100 GPU**

FPGA/GPU cluster configuration

Key Platform Benefits

- Secure remote access
- Machine learning frameworks: Tensorflow, Caffe and MXNet
- Support for deep learning training and inference
- Customizability: Select the right combination of accelerators for your application
- Reference designs using software stack for OpenCL, MPI heterogenous cluster computing
- Scalability: Create one node neural network graph and scale up by using more nodes
- Fast automated setup and configuration
- Technical support and training from CMC Microsystems

About CMC

Enabling innovation across Canada's National Design Network

CMC Microsystems delivers a nationwide, shared platform of tools and services to Canada's micro-nano innovators, helping to create the economy of the future.

25 multi-project wafer services available through nine foundries worldwide, offering industrial-scale manufacturing

40 university-based micro-nanotechnology (MNT) fabrication labs across Canada, helping researchers customize their designs

80 pieces of test equipment for loan in lab

560 CAD tools and modules

600 development systems

450 design flows, user guides and application notes

www.CMC.ca

Research Topics

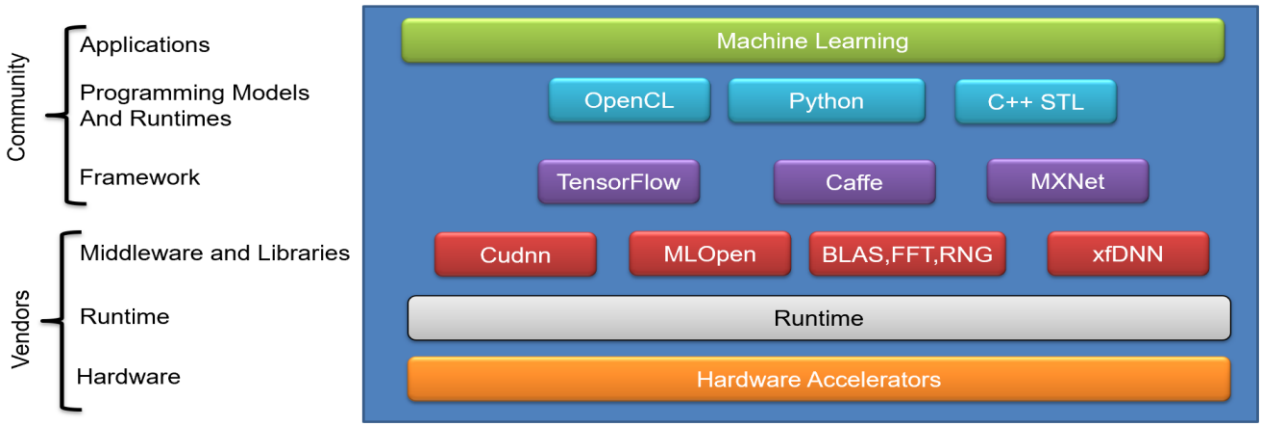
- Software IPs and applications targeting ML on heterogeneous computing systems (e.g. CNN, for object detection, speech recognition)
- Software stack including: Parallel programming models, Compilers, Middleware, Runtime, Drivers and OSes
- Case studies: ML, Big data analytics, data-intensive computing, cybersecurity.
- ASICs Prototyping: e.g., CMOS and other semiconductors, for implementing custom neural network accelerators.

FPGA/GPU Cluster Accelerators

Accelerator	Features	Host interface	Compute performance	Power
Xilinx Alveo U200	892,000 LUTs 64 GB Off-Chip mem. 77 GB/s bandwidth	Gen3x16	Unavailable	225W
NVIDIA V100 GPU	5,120 Cuda cores 640 Tensor cores 32GB /16GB HBM2 900GB/sec	Gen3x16	7 TFLOPS DP 14 TFLOPS SP 112 TFLOPS TP	250 W

FPGA/GPU Cluster Software stack

The FPGA/GPU cluster supports the three most commonly used deep learning frameworks, namely, TensorFlow, Caffe and MXNet. These frameworks provide a high-level abstraction layer for deep learning architecture specification, model training, tuning, testing and validation. Also included in the software stack are the various machine learning vendor specific libraries, that provide dedicated computing functions tuned for specific hardware architecture, delivering the best possible performance/power figure.



CMC Microsystems

Innovation Park at Queen's University
945 Princess St, Building 50
Kingston, ON, K7L 0E9
613.530.4666

Technical Contact

Dr. Yassine Hariri
Senior Engineer, Platform Design
613.530.4672
Hariri@cmc.ca