Hardware for HBM-Centric Computing

Andrew McCormick, Alpha Data
Overview

• HBM Architecture
• FPGA Acceleration
• Applications
• Architectures
• Boards
HBM Architecture

FPGA Fabric
32 x 256bit@400MHz Independent AXI Ports

AXI Switching Hard IP

8 Memory Controllers per bank

4GB HBM 4GB HBM
FPGA Acceleration

• Improvements in Algorithm Implementation Efficiency
  (Better Performance per Watt and Performance per Gate)
  • Dataflow pipeline processing
  • Minimal Control Logic (no Instruction Set)
  • Reduced Arithmetic Complexity
  • Application Specific Memory Management/ Caching
• Improvements in Performance
  • Parallelization
• Acceleration will eventually become memory bound
  • Fortunately HBM gives O(10x) BW improvement for wide parallel designs
Applications

• Machine Learning
  • Most energy efficient if CNN Network Weights can be stored on-chip (Cache/URAM/BRAM)
  • But some Networks too large for on-chip storage (VGG)
    • Go multi-chip? Or Use external DRAM?
  • Batch processing CNNs can relatively efficiently stream both data and weights from off-chip DRAM
    • But this becomes memory bound as batch size (and result latency) reduces
  • HBM should give 5-10x performance boost
Applications

• Data Processing and Analytics
  • Search (parallel)
    • Memory bound, and trivially parallelizable
  • Sort
    • Parallel Merge Sort O(N), can parallelize up to 16MB using FPGA BRAM/URAM
    • HBM allows parallel merge sort up to 4GB O(N)
  • K-means Clustering
    • Fixed point implementation friendly
    • Easily Parallelizable
    • Memory boundary depends on K, data size and dimensionality
Applications

- **3D-FFT**
  - URAM good for Corner Turn Memory up to $128 \times 128 \times 128$ (double precision)
  - HBM needed for Corner Turn Memory for $8k^3$
- **Matrix Multiply ($N \times N$)**
  - On its own, not memory bound – good data reuse ($N$)
  - Matrix product core limited to $N \times N$
    - Divide and Conquer used for larger matrices
    - But requires memory bound Matrix Add
    - Shared access of many matrix operations to same HBM stored data can provide comprehensive linear algebra accelerator library.
Architectures

- **Basic HBM Accelerator**

- Very effective for working data sets between 50MB and 8GB
- Off-load to/from a host via CPU controller DMA
- Over-lapped processing required for larger data sets
- Limited performance >8GB
• OpenCAPI HBM Accelerator

- Beyond 8GB, OpenCAPI provides low latency high bandwidth access to host memory
- Comparable performance with onboard DDR4
- Memory coherent with CPUs without any driver intervention or overhead
- Action can work directly on host memory and HBM “cache”
• OpenCAPI HPC HBM Accelerator

- In HPC clusters, data set working memory can be extended “horizontally” rather than “vertically”
- Data set shared across multiple HBM FPGA nodes using RDMA for edge element swapping
- Can also extend host shared memory across nodes, via OpenCAPI and RDMA
• **OpenCAPI Networking HBM Accelerator**
  
  - Similar architecture for Networking
  - HBM used as working buffer for 100GE packets
  - OpenCAPI low latency host link can match this bandwidth
  - Actions can also process packets between interface and host for functions such as threat detection, filtering, accelerated responses, network data analytics, real time responses.
  - Significant HBM bandwidth still available for actions on the inflight traffic
Boards

- ADM-PCIE-9H7

- Xilinx HBM Ultrascale+ VU37P
  - 8GB HBM
  - 2852K logic cells
  - 9k DSPs
  - 47MB on-die memory
  - 28.1 INT8 TOP/s

- PCIe Gen3/4x16 Edge

- 225W FH ¾L DW PCIe Form Factor

- 4 x QSFP28 Front IO (4x100GE)

- 2 x OpenCAPI Rear IO (2x25GB/s)

- 8 x 28Gb/s capable Firefly
Boards

- ADM-PCIE-9H7
  - 8 x 28Gb/s capable Firefly
    - Connect to QSFP28 expansion cards
    - Connect direct to other 9H7 cards
    - Connect to in chassis storage
Boards (non-HBM)

- ADM-PCIE-9V3
- Xilinx Ultrascale+ VU3P
- OpenCAPI FPGA NIC card
- Gen3x16 /Gen4x8 Edge
  - CAPI 2.0
- Low Profile (<75W) PCIe
- 2xQSFP28 (100GE)
- 16/32GB on-board DDR4
- OpenCAPI 25G interface
  - Ideal OpenCAPI development platform
Boards (Power 8, CAPI 1.0)

- ADM-PCIE-8K5 and ADM-PCIE-KU3
- Xilinx Ultrascale KU115/KU060
- CAPI 1.0 Cards
- CAPI SNAP Networking cards
- 10G/40G capable
- 16GB DDR3/DDR4 SDRAM

https://www.alpha-data.com/ku3
https://www.alpha-data.com/8k5
Boards (coming soon)

- ADM-PCIE-9H3

  - Xilinx Ultrascale+ VU33P
  - OpenCAPI HBM FPGA NIC card
  - Gen3x16 / 2xGen4x8 Edge
  - Low Profile (<75W) PCIe
  - 1xQSFP-DD (2x100GE)
  - OpenCAPI 25G interface
Summary

• HBM Offers an Order of Magnitude Improvement in Memory Bandwidth for FPGA applications

• HBM and OpenCAPI are a great fit

• Alpha Data have off-the-shelf boards available for these 2 new technologies
Questions

• Any questions?

• Come see us at Booth #9

• Or at Super Computing ’18 Dallas, Booth #222