Cloud Data Center Acceleration

2015
Agenda

- Computer & Storage Trends
- Server and Storage System
  - Memory and Homogenous Architecture
  - Direct Attachment
- Memory Trends
- Acceleration Introduction
- FPGA Adoption Examples
Server (Computer) & Storage Trends

Cloud computing, virtualization, convergence
- Server & storage consolidation & virtualization
- Convergence to PCIe backplane and low latency 25GbE
- Distributed storage and cache for cloud computing
- Convergence is enabling lower power and higher density

Lots of interest in Storage, Storage Class Memory
- Capacity Expansion
- DRAM to flash, and flash cache
- Intermediate Storage
- Disaggregation of Storage and Data
- Rapid change & new cloud architectures
- “Verticalization”, disaggregation, dense computing for cloud servers
- Acceleration option with FPGA per node, or pool heterogeneous accelerators
Memory & IO bottlenecks limit utilization
- Typical server workloads run at ~20% processor utilization
  - Virtualization driving application consolidation
  - But memory and IO are limiting factors to better utilization
- “Big Data” configurations are also bottlenecked
  - Especially search and analytics workloads

The Processor mostly waits for RAM
- Flash / Disk are 100,000 … 1,000,000 clocks away from cpu
- RAM is ~100 clocks away unless you have locality (cache).
- If you want 1CPI (clock per instruction) you have to have the data in cache (program cache is “easy”)
- This requires cache conscious data-structures and algorithms sequential (or predictable) access patterns
- In Memory DB is going to be common (SPARK Architecture)
O/S Bypass: DMA, RDMA, zero copy, cpu cache direct

- Avoid memory copies
- NICs, clusters, accelerators
- DMA, RDMA
  - Mellanox RoCE, Infiniband
- Intel PCIe steering hints
  - Into cpu cache
- Heterogeneous System Architecture (HSA)
  - For accelerators
- Direct Access to cpu cache
  - QPI, CAPI
  - Low latency
  - Simplified programming model
  - Huge benefit for flash cache
Computing Bottlenecks

Memory bottleneck
- Need faster & larger DRAM
  - CPU core growth > memory b/w
  - CPU has limited # of pins
  - DRAM process geometry limits
- Emerging:
  - Stacked DRAM in package
  - Optics from CPU package
  - Optics controller for clusters

Cluster networking
- Over optics

Main Storage Data response time
- Impact Big Data Processing
Emerging: Data Stream Mining, Real Time Analytics

Data stream examples
- Computer network traffic
- Data feeds
- Sensor data

Benefits
- Real time analytics
  - Predict class or value of new instances
  - e.g. security threats with machine learning
- Filtering data to store

Topology
- Single or Multiple FPGA accelerators
Enter New Memory solutions (A new Dawn Awaits)

- Did the 14nm NAND delay drive these solutions to become the next gen?
- Or did the need for more flexible memory and storage applications drive this transition?
- New Memories are complementary to existing solutions
  - How to Adopt
  - Where do they go
- How do they fit in tomorrow's Server/storage Architectures
3D XPoint vs. NAND

- 1000X faster write
- Much better endurance
- 5X to 7X Faster SSD’s
- Cost & Price in between DRAM and flash
- Altera FPGA controller options
Rapid Change in the Cloud Data Center

- **Rapid change & new cloud architectures**
  - “Verticalization”, disaggregation, dense computing for cloud servers
  - Intel offering 35 custom Xeons for Grantley
  - Software Defined Data Center
    - Pool resources (compute, network, storage), automate provisioning, monitoring
  - Intel MCM & Microsoft Bing FPGA announcements
  - Intel Standard to Custom Roadmap showing 35 Grantley SKU’s:
Accelerator Spectrum

Application Spectrum

Computer Vision  Data Analytics  Language Proc.  Visual Analytics  Search Ranking

Data Streaming  Computational  Medical Diagnosis  Image Pattern Recognition  Best match Engines

Algorithms

Database  Graph  Machine Learning  Numeric Computing
Efficient Data Centric Computing Topologies

Server with Unstructured Search Topology – e.g. Hadoop + Map/Reduce
Switch or/and Large Aggregating Processor
e.g. map result collection and Reduce Function

Small processors close to storage e.g. Map Function
P1
Memory
Memory
Flash Drive(s)
Flash Drive(s)
Pn
Network Attach

Application: Data Analytics / Data Search / Video Server

Server with Balanced FLOPs/Byte/s and FLOPs/Byte Depth

X TFlop Processor
X TB/s
Network / Storage Attach
X TBytes Memory

Application: Large Dataset HPC with Compute intensive function that do not scale well – e.g. FEA

Server With Multi Node Pipeline
Network / Storage Attach

Application: Deep Pipeline DSP, e.g. Video Analytics

Server with 3D Torus Configurations

Application: Classic HPC, e.g. QCD, CFD, Weather Modeling

130 GB Memory 130 GB Memory 130 GB Memory

Network Attach / Storage Attach

Network Attach / Storage Attach

130 GB Memory

Network Attach / Storage Attach

Network Attach / Storage Attach

© 2015 Altera Corporation—Public
Microsoft SmartNIC with FPGA for Azure
(8-25-15 Hot Chips Presentation)

- Scaling up to 40 Gbs and beyond
  - Requires significant computation for packet processing
- Use FPGAs for reconfigurable functions
  - Already used in Bing
  - SW Configurable
- Program with Generic Flow Tables (GFT)
  - SDN i/f to hardware
- SmartNIC also does Crypto, QoS, storage acceleration, and more...

http://tinyurl.com/p4sghaq
FPGA AlexNet Classification Demo
(Intel IDF, August 2015)

**CNN AlexNet Classification**
- 2X+ Performance/W vs cpu (Arria 10)
- 5X+ performance Arria 10 → Startix 10
  - 3X DSP blocks, 2X clock speed

**Microsoft Projection**
- 880 images/s for A10GX115
- 2X Perf./W versus GPU

**Altera OpenCL AlexNet Example**
- 600+ images/s for A10GX115 by year end

<table>
<thead>
<tr>
<th>CNN Classification Platform</th>
<th>Power (W)</th>
<th>Performance (Image/s)</th>
<th>Efficiency (Images/sec/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E52699 Dual Xeon Processor (18 cores per Xeon)</td>
<td>321</td>
<td>1320</td>
<td>4.11</td>
</tr>
<tr>
<td>PCIe w/ dual Arria 10 1150</td>
<td>130*</td>
<td>1200</td>
<td>9.27</td>
</tr>
</tbody>
</table>

Note *: CPU low power state of 65W included.
Why Expansion Memory?

- Enable memory-intensive computation
  - Change the way we look at data

- Increase users’ productivity
  - Boost scientific output
  - Broaden participation

- machine learning
- graph-based informatics
- high-productivity languages
- Load Balancing
- data exploration
- statistics
- Big Data
- ...
Advanced memory controller market
Memory innovation will change how computing is done

- Emerging market for “Advanced Memory Controllers”.
  - These devices interface to the processor by directly attaching to their existing memory interface bus.
  - Memory Types will require New Controller implementations

- Memory offload Applications
  - Filtering, Acceleration, Capacity, Sub-Systems

- FPGA can translate between existing memory interface electricals and a plethora of backend devices, interfaces, or protocols to enable a wide variety of applications.
  - Initial examples of this include:
    - Bridging between DDR4 and other memory technologies such as NAND Flash, MRAM, or Memristor.
    - Memory depth expansion to enable up to 8X the memory density available per memory controller.
    - Enable new memory adoption quickly
    - Enable acceleration of data processing for analytics applications
    - Enable offload of data management functions such as compression or encryption.
Application: DDR4 DIMM Replacement
- Memory Bridging and/or In-line Acceleration

Key Memory Attributes
- Capacity
- Sub System mixed Memory
- Optimized Solution for App
- Database Acceleration

On-Chip Cache
DIMM Module

FPGA
- Ctrl/Accel Logic
- Memory Filter/Search
- On-Chip Cache

ADV MEM CTRL

DDR4 Slave

DDR4 Slot 0

DDR4 Slot 1
Acceleration Solutions

Data making money the new Way
### Acceleration Memory Applications

<table>
<thead>
<tr>
<th>Accelerator Application</th>
<th>Memory Function</th>
<th>Memory Type</th>
<th>Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Analytics</td>
<td>Temporary Storage</td>
<td>DDR3/4</td>
<td>Storage Class, HBM, HMC</td>
</tr>
<tr>
<td>Computer Vision/OCR</td>
<td>Buffer</td>
<td>DDR3/4</td>
<td>Storage Class</td>
</tr>
<tr>
<td>Image Pattern Recognition</td>
<td>Storage, Buffer</td>
<td>SSD, DDR</td>
<td>Storage Class, HBM, HMC</td>
</tr>
<tr>
<td>Search Ranking</td>
<td>storage, Working</td>
<td>DDR3</td>
<td>Storage Class</td>
</tr>
<tr>
<td>Visual Analytics</td>
<td>Buffer</td>
<td>DDR3</td>
<td>Storage Class</td>
</tr>
<tr>
<td>Medical Imaging</td>
<td>Storage, Buffer</td>
<td>SSD, DDR3/4</td>
<td>Storage Class, DDR4,</td>
</tr>
</tbody>
</table>

- As FLOPs increase Memory Bandwidth will need to scale
- As Data increases capacity will also increase to sustain computation
Accelerator Board Block Diagram

- 4GB DDR4 SODIMM
- 4GB DDR4 SODIMM
- 4GB DDR4 Soldered On
- 4GB DDR4 Soldered On
- FPGA Arria 10 1150 GX
- 2GB HMC Hybrid Memory Cube (Optional)
- PCI Express Bridge

Connections:
- x72
- x16
- x8
- PCIe x16

Other Components:
- Micro-USB
- Flash FPGA Config

© 2015 Altera Corporation—Public
Dual Arria 10 High Memory Bandwidth FPGA Accelerator

- GPU Form Factor Card with 2x Arria 10 10A1150GX FPGAs
  - Dual Slot Standard Configuration, Single Slot width possible, if user design fits within ~100W power footprint

- 410 GBytes/s Peak Aggregate Memory Bandwidth
  - 85GB/s Peak DDR4 Memory Bandwidth per FPGA
  - 60GB/s Write + 60GB/s Read Peak HMC Bandwidth per FPGA

- 132 GBytes Memory Depth or 260GBytes with Soft Memory Controllers
  - 4GByes of HMC memory shared between FPGAs

- 60 GBytes/s, 7.5GBytes/s/Ch/Dir, board to board pipelining bandwidth
  - (4) Communication channels running at 15Gb/s or (4) 40GbE Network IO channels

NOTE: Performance numbers are absolute maximum capability & peak data rates.
Dual Stratix 10 3D Torus Scalable FPGA Accelerator

- GPU Form Factor Card with 2x Stratix 10 FPGAs
  - Support Majority of Stratix 10 Family – Both large and small devices from 2 to 10 TFlops
- 204 GBytes/s Peak Aggregate Memory Bandwidth
  - 102GB/s Peak DDR4 Memory Bandwidth per FPGA
- 256 GBytes Memory Depth
- 336 GBytes/s, 14GBytes/s/Channel/Direction, board to board Scaling

NOTE: Performance numbers are absolute maximum capability & peak data rates
Minimise Multiple accesses to External Memory

Traditional CPU/GPU Implementation

Iterate many times

Access entire volume data storage in system memory

Result read from p & q buffers after many thousands of iterations
Minimise Multiple accesses to External Memory

Traditional CPU/GPU Implementation

Iterate many times

Function A

Function B

Function C

Function D

Function E

Access entire volume data storage in system memory

Result read from p & q buffers after many thousands of iterations
Minimise Multiple accesses to External Memory

Iterate many times

Global memory

Function A
Function B
Function C
Function D
Function E

Access entire volume data storage in system memory

Result read from p & q buffers after many thousands of iterations
Minimise Multiple accesses to External Memory

FPGA Implementation

Iterate many times

Access entire volume data storage in system memory

Result read from p & q buffers after many thousands of iterations
Try to Minimise Multiple accesses to External Memory

FPGA Implementation

Iterate many times

Access entire volume data storage in system memory

Global memory

Global memory

Function A

Function B

Function C

Function D

Function E

Delay Line External Memory

Delay Line

Deeper than blockram when large algorithm data alignment is required to further extend the deep pipeline

Result read from p & q buffers after many thousands of iterations
Summary

- FPGA utilizes less external memory bandwidth for Reverse Time Migration, CNN and other common acceleration algorithms.
- The growth in data and TFLOPs for Acceleration will require more BW and in a orderly fashion. New Memories will require higher bandwidth and controller changes.
- Memory and System solution to increase compute efficiency are changing architectures, networks and the type of memory.