FPGA based HPC Acceleration

Houston Tang
Computer & Storage BU
PSG, Intel
htang@altera.com
Agenda

- Challenges we are facing

- Algorithm Acceleration
  - Neural Algorithms, Financial, Bio-Science, Oil & Gas, Recognition, Imaging, Search, …

- Network Acceleration
  - Virtualization, Encryption, Compression, Deduplication, DPI
  - CPU Offload, Low Latency

- Data Access Acceleration
  - Data Analytics, Filtering, Compression, Deduplication
  - O/S & Driver Bypass, NVMe

- Summary
Challenges we are facing

- Variety of applications are becoming bottlenecked by scalable performance requirements
  - E.g. Big Data Analytics, object recognition, search engines, etc…
- Overloading CPUs capabilities
  - Frequencies are capped
  - Processors keep adding more cores
  - Need to coordinate all the cores and manage data
  - TCP/IP, UDP, OVS increasing CPU utilization while scaling up to 50G, 100G Ethernet
- GPUs require re-optimization between generations
- Power consumption of CPUs and GPUs limits system size
- Maintaining coherency throughout scalable system
Agenda

Algorithm Acceleration
- Neural Algorithms, Financial, BioScience, Oil & Gas, Recognition, Imaging, Search, ...

Network Acceleration
- Virtualization, Encryption, Compression, Deduplication, DPI
- CPU Offload, Low Latency, Compliance

Data Access Acceleration
- Data Analytics, Filtering, Compression, Dedupe
- O/S & Driver Bypass, NVMe

Summary
Case Study 1: CNN Algorithms

Convolution Neural Network
- A branch of machine learning targeting for image classification

Altera CNN IP
- Based on Alexnet network, focusing on scoring
- Developed on Arria 10 hardware and OpenCL compiler
- Demonstrated on Zhenzhen IDF this April
- Achieved 500+ throughput with around 35W power
- Demo introduction video available soon
Case Study 2: JPEG decoder for image storage

JPEG decoding in OpenCL
- Ported and optimized in 2 weeks

Decompresses images on the fly from storage system
- FPGA architecture allows several accelerators to run concurrently
- Utilized ~36% of Stratix V
- Maxes out PCIe performance

<table>
<thead>
<tr>
<th>JPEG file</th>
<th>Frames / s</th>
<th>File size (MB)</th>
<th>Bandwidth (MB / s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1025.jpeg</td>
<td>151</td>
<td>1.459</td>
<td>2739</td>
</tr>
<tr>
<td>1006.jpeg</td>
<td>1310</td>
<td>0.685</td>
<td>1600</td>
</tr>
<tr>
<td>1023.jpeg</td>
<td>2015</td>
<td>0.326</td>
<td>1856</td>
</tr>
<tr>
<td>1036.jpeg</td>
<td>3667</td>
<td>0.033</td>
<td>896</td>
</tr>
<tr>
<td>1044.jpeg</td>
<td>1347</td>
<td>0.651</td>
<td>1645</td>
</tr>
</tbody>
</table>
Case Study 3: OpenCL Example-GZIP Compression
Great Quality of Results, TTM, and Ease of Use

Alterna summer intern ported and optimized GZIP algorithm in a little more than a month

Industry leading companies FPGA engineer coded Verilog in 3 months

Much lower design effort and design time

OpenCL Was

10% Slower
12% more resources
3x faster development time
Case Study 4: Document Filtering Benchmark

- Unstructured data analytics
  - Bloom Filter
- Advantage FPGA
  - 1.5X versus Intel 6 core
  - Similar performance to GPU
  - 5X+ Performance to Power

Results:

<table>
<thead>
<tr>
<th>Platform</th>
<th>Power (W)</th>
<th>Performance (MTs)</th>
<th>MTs/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>W3690 Xeon Processor</td>
<td>130</td>
<td>2070</td>
<td>15.92</td>
</tr>
<tr>
<td>nVidia Tesla C2075</td>
<td>215</td>
<td>3240</td>
<td>15.07</td>
</tr>
<tr>
<td>PCIe385 Stratix 5-A7 Accelerator</td>
<td>25</td>
<td>3602</td>
<td>144.08</td>
</tr>
</tbody>
</table>

Case Study 5: Multi-Asset Barrier Option Pricing

- Monte-Carlo simulation
  - No closed form solution possible
  - High quality random number generator (RNG) required
  - Billions of simulations required
- Used GPU vendors example code
- Advantage FPGA
  - Complex Control Flow
- Optimizations
  - Channels, loop pipelining

<table>
<thead>
<tr>
<th>Platform</th>
<th>Power (W)</th>
<th>Performance (Bsims/s)</th>
<th>Efficiency (Msims/s/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W3690 Xeon Processor</td>
<td>130</td>
<td>.032</td>
<td>0.0025</td>
</tr>
<tr>
<td>Largest 28nm GPU</td>
<td>212</td>
<td>4.5 w/MT RNG</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.1 w/XORWOW</td>
<td>48</td>
</tr>
<tr>
<td>Altera Stratix D8 (w/Mersenne Twister RNG)</td>
<td>45</td>
<td>12.0 w/MT RNG</td>
<td>266</td>
</tr>
</tbody>
</table>
Case Study 6: H.265 Server Co-processing for stream video

1 channel 4Kp60

Cost: ~$9600
Power: 4x135W = 540W

Cost: ~$4800 + 2 FPGAs
Power: 2x135W + 40W = 310W

Harmonic Chooses Altera Solution for H.265 4Kp60 Video Encoding

- Lead Customer: Harmonic #1 in Broadcast Encoding
- “We need this...a room full of servers is tough to show...”
Case Study 7: AES Encryption
(40Gb Ethernet or Data Access)

- Encryption/decryption
  - 256bit key
  - Counter (CTR) method

- Advantage FPGA
  - Integer arithmetic
  - Coarse grain bit operations

<table>
<thead>
<tr>
<th>Platform</th>
<th>Power (W)</th>
<th>Performance (GB/s)</th>
<th>Efficiency (MB/s/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E5503 Xeon Processor (single core)</td>
<td>est 80</td>
<td>0.01</td>
<td>0.125</td>
</tr>
<tr>
<td>AMD Radeon HD 7970</td>
<td>est 100</td>
<td>0.33</td>
<td>3.3</td>
</tr>
<tr>
<td>PCIe385 A7 Accelerator</td>
<td>25</td>
<td>5.20</td>
<td>208</td>
</tr>
</tbody>
</table>

Counter (CTR) mode encryption
Agenda

Algorithm Acceleration
- Neural Algorithms, Financial, BioScience, Oil & Gas, Recognition, Imaging, Search, ...

Network Acceleration
- Virtualization, Encryption, Compression, Deduplication, DPI
- CPU Offload, Low Latency, Compliance

Data Access Acceleration
- Data Analytics, Filtering, Compression, Dedupe
- O/S & Driver Bypass, NVMe

Summary
Network Rapid Customization Value Propositions

- **Lower Latency**: critical in High Frequency Trading (HFT) or program trading, to get the latest price before the competition. 100 nsec. matters.

- **CPU Offload**: by filtering data or offloading algorithms onto the fpga, use less cpu’s, power, and space and save on IT costs.

- **Compliance**: enforce rule checking or monitor trades going to the exchange.
Network Rapid Customization Too!

- Mellanox
  - Virtualization
  - Encrypt/decrypt
  - Dedupe, compression

- NASDAQ
  - CPU Offload

- SolarFlare
  - Low latency 10G
  - CPU Offload
“SmartNIC”

- Utilize FPGA fast transceiver
- OVS offload
- TCP/UDP offload
- DPI
- RoCE
- Microsoft SmartNic
Agenda

Algorithm Acceleration
- Neural Algorithms, Financial, BioScience, Oil & Gas, Recognition, Imaging, Search, ...

Network Acceleration
- Virtualization, Encryption, Compression, Deduplication, DPI
- CPU Offload, Low Latency, Compliance

Data Access Acceleration
- Data Analytics, Filtering, Compression, Dedupe
- O/S & Driver Bypass, NVMe

Summary
A tradeoff between capacity, speed and cost
- Exploit the principle of locality.

CPU attach goals
- Reduce OS awareness of I/O bus, fast large storage

Memory coherent attach
- OS bypass and larger working memory space

FPGA memory controllers with accelerators
- FPGA supports DMI, QPI, PCIe, NVMe, DDR Target, etc…
Key Value Store w/Algo-Logic Systems (Altera Partner)

- Compelling KVS Results
  - 150M searches/s.
  - Sub-microsecond
  - Better energy efficiency

- Comprehensive KVS White Paper
  - Comparison to:
    - Cpu with sockets
    - Cpu with DPDK

- Real-time Hands-on
  - KVS Tutorial at IEEE HotI 2014
SSD Controller Reference Design for Data Center
High IO Rate and Large Flash Cache Environment

- 2X to 7X Write Endurance
  - Better ECC w/BCH or LDPC
- Better “tail latency”
- 100K IOPS per channel
- NVMe interface
- Support new technologies
- Support new vendors
- SSD’s & Arrays
- Features
  - Entry: Limited features
  - Mid: w/compression or encryption
  - High: filtering, etc.

A10 Reference Design w/Mobiveil & NVMDurance
Thank You
Case Study 6: Oil & Gas – 4K³ RTM TTI Overview

Memory Bound with GPU
- <10% FLOP Efficiency, 100% Memory BW
- GPU Memory limits volumes to <1K³
- Poor Scaling beyond 1 GPU
  - PCIe Bottleneck

Balanced FLOP/Byte with FPGA
- Sliding Window
  - Less external memory accesses required
  - Efficient SIMD re-use of local data
  - Can supports dense stencils
- Deep Pipelines
  - Minimizes External Memory Accesses
- >50% FLOP Efficiency
- Very Power Efficient
- Deep DDR4 Memory on Accelerator
  - Natively supports volumes >1K³
- Linear Scaling to 4K³ volumes and beyond
  - Direct FPGA to FPGA accelerator Comms
Oil & Gas Prototyping Board With Dual A10 and HMC

510T

- GPU Form Factor Card with (2) Arria 10 10A1150GX FPGAs
  - Dual Slot Standard Configuration
  - Single Slot width possible, if user design fits within ~100W power footprint
- PCIe Gen3 x 16
- 290 GBytes/s Peak Aggregate Memory Bandwidth
  - 85GB/s Peak DDR4 Memory Bandwidth per FPGA (4 Banks per FPGA)
  - 30GB/s Write + 30GB/s Read Peak HMC Bandwidth per FPGA