The role of FPGAs in redefining the modern datacenter
Dell EMC Accelerating Understanding, Pisa Italy
27th September 2017

Craig Petrie
VP Business Development, Nallatech – a Molex company
craig.petrie@molex.com
Goal

Discuss the trends influencing the adoption of FPGAs as viable datacenter accelerators

Application case studies featuring FPGAs, including Machine Learning

Learn how DELL EMC, in collaboration with Nallatech – a Molex company, can address the requirements of customers wishing to use FPGAs
Molex FPGA Solutions

Microelectronics

Accelerated Computing
Nallatech – a Molex company

Server qualified accelerator cards featuring FPGAs, network I/O and an open architecture software/firmware framework

- Nallatech – a Molex company
- 25 years of FPGA heritage
- Energy-efficient heterogeneous computing
- Real-time, low latency network & I/O processing
- Intel (Altera) OpenCL partner
- Qualified FPGA product for multiple DELL EMC platforms
- Successfully deployed high volumes of FPGA accelerators with DELL EMC
- Application porting & optimization services
FPGAs have a product lifecycle of 15+ years
Audi Selects Intel SoC FPGAs
Autonomous Piloted Driving Capability
Deep Learning for self-driving cars

Source: MIT Micro Tutorial 2016
Intel acquires Altera
$16.7Bn

The datacenter is changing…
FPGAs address a wide range of markets

- **Automotive/Industrial**
  (Pedestrian Detection, Motion Estimation)

- **Computer & Storage**
  (HPC, Financial, Data Compression)

- **Military/Government**
  (Crypto, Image Detection)

- **Broadcast**
  (Video image processing)

- **Medical**
  (Diagnostic Image Processing, BioInformatics)

- **Networking**
  (DPI, SDN, NFV)

Two fundamental use cases:

1. Intelligent networking and integrated low latency I/O processing
2. Co-processing / CPU off-load for energy-efficient heterogeneous HPC
Machine Learning

» Extracting features from data in order to solve predictive problems

» Automatically recognize complex patterns

» Make intelligent (valuable) decisions

» Mature applications:
  » Speech & Language
    » Audio to Text (and vice-versa)
    » Translation
  » Image
    » Classification
    » Detection
Image Classification Accuracy

Top 5 Classification Error (%)

large error rate reduction due to Deep CNN

Source: [Russakovsky et al, IJCV 2016]
Example application: Medical, cancer detection

Source: [Jermyn et al., JBO 2016]
Opportunities


Cumulative Deep Learning Software Revenue by Region, World Markets: 2015-2024

Source: Tractica
30% of cloud will be FPGA by 2020
Novel use of FPGAs within the datacenter
Project “Catapult”: Accelerate Bing

Order of magnitude performance:

- ½ the number of servers
- <30% server TCO increase
- ~10% increase in power

Highly disruptive, market-changing
Spectrum of approaches to high performance

- CPUs
  - Single Cores
  - Multi-Cores
    - Coarse-Grained
      - CPUs and DSPs

- DSPs

- Multi-cores

- Arrays
  - Coarse-Grained
    - Massively Parallel Processor Arrays
  - Fine-Grained Massively Parallel Arrays

- FPGAs
Efficiency via Specialization

- **ASICs**
- **FPGAs**
- **GPUs**

Graph showing energy efficiency (MOPS/mW) vs. processor number (sorted by efficiency). The graph illustrates the efficiency of different types of hardware, with ASICs being the most efficient and FPGAs being more flexible. GPUs and microprocessors fall in between, showing significant improvements in efficiency with the use of dedicated hardware.
Deep Pipeline Parallelism = Acceleration

- Bit manipulation
- Integer arithmetic
- Single precision floating point (transcendental)
- Large local storage requirements
- Complex control flow
- Unusual function mix
- Predictably memory access patterns
It's not about clock speed

Performance is based on fine grain massive parallelism

More transistors = more performance!
Tool Flow Support

All Nallatech products support traditional HDL tool flows

- Intel-based product features support for OpenCL
- HDL cores for hardware-orientated customers

“Libraries”: incorporate optimised HDL cores into OpenCL frameworks
Application Development Paradigm

OpenCL expands the number of application developers.
Application Design Services

Nallatech offers 25 years of experience in helping customers achieve optimum performance with FPGA technology

- System Architecture
- Hardware / Software / Firmware design
- OpenCL optimization
- Full application ports from software to FPGA
- Specialist in data-centric computing
Real-Time Encryption/Compression

- Counter-based encryption/decryption
  - 256-bit key
- Advantage FPGA
  - Integer arithmetic
  - Coarse grain bit operations
  - Complex decision making

Results:

<table>
<thead>
<tr>
<th>Platform</th>
<th>Performance/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>E5503 Xeon Processor</td>
<td>0.0004 (single core)</td>
</tr>
<tr>
<td>AMD Radeon HD 7970</td>
<td>0.00165</td>
</tr>
<tr>
<td>PCIe385 A7 Accelerator</td>
<td>0.208</td>
</tr>
</tbody>
</table>

\[\times 126\]
Lattice Boltzmann – Complex Fluid System Simulation

» Pipelining implementation effectively implemented for ‘memory bound’ problem
» OpenCL programming abstracts complexities of the FPGA
» 2 weeks development effort
3D-Component Matching – Oil & Gas/Fracking

» Post processing of 3D textures for analyzing rock samples
» Sequential code: 99 years to process on a single core 😊
» Nvidia and Nallatech approached to write a kernel that offloads to a PCIe-attached accelerator to process the data in 30 minutes

Nvidia

» 438 K80 cards
» 4 racks
» 131kW (300W per card)

Nallatech

» 122 Intel Arria 10 FPGA cards
» ½ rack
» 4.86kW (40W per card)
Deep Learning with Convolutional Neural Networks

» Convolutional Neural Network (CNN)
  » Feed forward artificial neural network
  » Typically composed of multiple layers
  » Can tune weights to “learn”

» Several open source frameworks
  » Caffe (UC Berkeley)
  » Torch
  » Theano (University of Montreal)
Deep Learning

Image

“Volvo XC90”

Source: Lee et al., Comm. ACM 2011
FPGA acceleration focused on forward path

Benchmarking to date have concluded:

- Same performance as GPU, but using less power

Low bit precision required by CNN networks allows for significant optimisation, taking advantage of FPGA logic fabric

- The ImageNet CNN coefficient data can be reduced to 10 bits with less than 1% loss in accuracy

FPGA implementation is extremely low latency. Low power permits real-time CNN embedded in Edge platforms
CNN: ImageNet Acceleration

» FPGA acceleration of the ImageNet CNN versus a NVIDIA K40 using the same Caffe framework and setup

» The Nallatech 510T FPGA accelerator can categorise an image in average 0.748 milliseconds compared to 2 milliseconds for the K40 GPGPU
## AlexNet competitive analysis - Classification

<table>
<thead>
<tr>
<th>System</th>
<th>Throughput (img/s)</th>
<th>Power  (W)</th>
<th>Throughput/W (img/s/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arria 10 (Optimized: FP32, Full Size @350MHz)</td>
<td>800</td>
<td>~37W</td>
<td>21.6</td>
</tr>
<tr>
<td>Arria 10 (Optimized: FP16, Full Size @350MHz)</td>
<td>1200</td>
<td>~36W</td>
<td>33.3</td>
</tr>
<tr>
<td>2x Arria 10</td>
<td>2400</td>
<td>~75W</td>
<td>32</td>
</tr>
<tr>
<td>Caffe on NVIDIA Titan X</td>
<td>1000</td>
<td>~250W</td>
<td>4</td>
</tr>
<tr>
<td>cuDNN4 on NVIDIA Titan X</td>
<td>3216</td>
<td>~227W</td>
<td>14.2</td>
</tr>
</tbody>
</table>

Source: Intel – Machine Learning on FPGAs: Neural Networks
FPGAs for Real-Time ML/AI

Microsoft: Project Brainwave
“Persistent” Neural Nets

Observations
State-of-art FPGAs have $O(10K)$
distributed Block RAMs $O(10MB)$
⇒ Tens of TB/sec of memory BW

Large-scale cloud services and
DNN models run persistently

Solution: persist all model
parameters in FPGA on-chip
memory during service lifetime

When single request arrives, all chip resources (on-
chip memories and compute units) are used to
process a single query (no batching required)

Source: Microsoft Brainwave Hot Chips presentation 2017
Narrow Precision Inference on FPGAs

FPGA Performance vs. Data Type

- Stratix V D5 @ 225MHz
- Stratix 10 280 @ 500MHz

Tera-Operations/sec

Source: Microsoft Brainwave Hot Chips presentation 2017
DELL EMC can supply Intel-based Nallatech FPGA platforms directly to customers under existing purchasing agreements.
One stop shop

Nallatech FPGA cards are available **pre-integrated** in DELL EMC server platforms

- **Development Systems**
  Customers place a single order and receive a complete system ready to start programming with OpenCL:
  - Nallatech FPGA Card
  - Quartus and OpenCL SDK
  - Performance-spec server for FPGA compiling

- **Production Platforms**
  Nallatech and DELL EMC ensure system qualification for production:
  - Thermal qualification & management
  - Power qualification & management
  - Integration with server’s BIOS and system management tools
Form Factor

Standard products will conform to two standard PCI Express physical sizes supported in most server and blade platforms

Half Height, Half Length
» Default size for **NIC** cards - ubiquitous
» Typically **25W** power consumption, single-width
» 8-lane PCI Express Gen 3.0 support (10GB/s)

Full Height, ¾ Length
» Default size for **GPU/Xeon Phi** cards
» Up to **300W** power consumption, double-width
» 16-lane PCI Express Gen 3.0 support (20GB/s)
**Altera Arria 10 Products**

**385A SoC**

Embedded ARM processors

- **NIC** Form Factor - Low Profile, Half Length PCIe form factor easily fits in most server platforms
- **Altera Arria 10 SoC** (dual ARM Cortex-A9 processors)
- 8-lane PCIe Gen 3.0
- (2) QSFP Cages supporting dual 1G, 10G & 40G line rates
- (2) banks of SDRAM @ 2133MT/s, up to 32GB total @ 34GB/s
- Board Support Packages (BSP) for Altera OpenCL SDK
- Typical Power Consumption: <25 watts
Intel Arria 10

510T
Extreme Compute Acceleration

- GPU/Xeon Phi form factor - Dual slot standard configuration
- **DUAL Intel Arria 10 FPGAs** 10A1150GX speed grade 2
- Active and Passive Cooling Options
- 16-lane PCIe Gen 3.0
- **(8)** Banks of DDR4 (4 banks per FPGA): ~3240 GBytes/s memory bandwidth
- Board Support Packages (BSP) for Intel OpenCL SDK
- Typical Power Consumption: up to 150 watts (50% of Tesla-class GPU card)
Maximize Acceleration – Increase Efficiency – Save Time

The Nallatech FPGA Accelerated Compute Node allows you to drive the most demanding HPC, data visualization and rendering workloads with a flexible, extremely dense 1U rack server optimized for accelerators.
The next generation is here...
Summary

» FPGAs are highly disruptive for datacenter applications
» FPGA-enabled platforms can deliver significant performance improvements using less power
» Intel’s acquisition of Altera legitimises FPGAs and has kickstarted the ecosystem
» Intel’s OpenCL SDK allows software (CPU/GPU) customers to program FPGAs
» Next generation applications, such as Machine Learning require the use of heterogeneous architectures featuring FPGAs

» FPGAs are now a key technology for accelerating the datacenter
» In collaboration with Nallatech – a Molex company, DELL EMC has a leadership position in the market
Thank You!

Craig Petrie
VP Business Development
craig.petrie@molex.com