Unconventional Architectures with Reconfigurable Computing

Michaela Blott, Principal Engineer, Xilinx Research
Agenda

- Introductions
- Industry landscape
- Platform characterization & performance estimation
- Some unconventional examples
- Summary & future plans
Introductions – FPGAs, Xilinx, Xilinx research and the labs in Dublin

Industry landscape

Platform characterization & performance estimation

Some unconventional examples

Summary & future plans
What are FPGAs?

Customizable, Programmable Hardware Architectures

Great vehicle for implementation of unconventional architectures
Customized Hardware Architectures

Task: $X = (a + b) \times 256$;

1 cycle @ 400MHz

6 cycles @ 2GHz

FPGA promise very high throughput, latency, and power savings
Customizable Interfaces & Memory Architectures

Flexibility to interface to any other device and customize memory architectures
Introduction to Xilinx

- Fabless semiconductor company
- Founded in Silicon Valley in 1984
- Today: Approximately 3,500 employees and $2.25B revenue
- 20,000 Customers

1st FPGA in 1985: XC2064

- 64 FF
- 128 3-input LUTs
- 58 IOs
- 2um

Ultrascale +: VU13P

- 3.4M FF
- 1.7M 6-input LUTs
- 6.3Tbps IO
- 16nm
- + DSPs, ARM, 2.5D…

30 years
Xilinx is Diversified Across Multiple Markets

- Aerospace and Defense
- Industrial and Medical
- Test, Measure and Emulation
- Automotive and Transportation
- Wireless Communications
- Audio, Video Broadcast
- Consumer
- Wired Comms and Data Center

Mars Rover
MRI scanners
3D televisions
Machine Learning
ADAS
AR
Xilinx Research - Ireland

Applications & Architectures

Through application-driven technology development with customers, partners, and engineering & marketing
Introductions

Industry landscape

Platform characterization & performance estimation

Some unconventional examples

Summary & future plans
Moore’s Law & The Technology Pipeline

Scaling becomes increasingly esoteric
Transistor Cost Trend

Calculation of Cost Per Transistor by Node

Source: IBS

Cost per million gates

Economics become questionable
End of Dennard Scaling

Power dissipation becomes problematic

Source: Intel
Computing: Increasingly Heterogeneous and Integrated

- **Applications require**
  - Increasing compute (machine learning, data analytics)
  - Increasing storage capacity (photos, videos)
  - Lower power (OPEX = 2x CAPEX)

- **Heterogeneous compute is required to provide further performance scaling and reducing power consumption**

- **Accelerator integration transitions from**

  - Loosely coupled IO device, coherent accelerators (CAPI, QPI, CCIX) to on-chip integration with processors and memory

New decade of **application-driven architectures**

Diversification with increasingly heterogeneous devices
New Generation of Design Environments (FPGAs) make it easier

- **Legacy**
  - ISE, RTL-based design entry with IP library

- **Embedded CPU integration**
  - Microblaze, SDK, EDK

- **Raised abstraction for accelerators**
  - Vivado HLS
  - SDNet (DSL PX)
  - Block stitching and manual integration in platform in RTL

- **Simplified host integration & automated infrastructure creation**
  - SDSoc, SDNet, SDAccel

Monitoring & profiling infrastructure, Runtime OS, Dynamic and static workload partitioning, Cloud integration
For a given application, which architecture should I build?

For a given architecture, which applications are suitable?

=> Characterization & benchmarking
Agenda

- Introductions
- Industry landscape
- Platform characterization & performance estimation
- Some unconventional examples
- Summary & future plans
Agenda

- Introductions
- Industry landscape
- Platform characterization & performance estimation
- Some unconventional examples
- Summary & future plans
Peak performance as a function of operational intensity

- \( PT = \min \{ OI \times BW; P \} \)
- Takes into account maximum compute performance and memory bandwidth

Very crude but useful for performance estimates and platform comparison
Performance Estimation & Tracking

Allows performance estimates & tracking of optimizations

Achievable Performance
GOPS/sec (log)

Estimated peak performance

Average cost for a mix of operations
8b add
SP mult

Measurements

Current project: refining rooflines
Platform Characterization

- **FPGAs**
  - Highest performance for non-float (fixed point, characters, bit) with operational intensity > 16
  - Float/power
  - Lowest absolute power

- **GPU**
  - Absolute float performance for highly data parallel applications with little control flow

- **CPU**
  - Best balance, all-round average performance for all applications, in particular with large memory requirements
Introductions

Industry landscape

Platform characterization & performance estimation

Some unconventional examples

Summary & future plans
Applications under Investigation

Key value stores
Machine learning
Vision Processing
Genomics
Networking
Stencils
Fintech
Synthetic workloads

Unconventional memory architecture
Unconventional compute
Key-Value Stores

- Common middleware application to alleviate access bottlenecks on databases
  - Most popular and most recent database contents are cached in main memory of a tier of server platforms
  - Provides the abstraction of an associative memory

- $O_I = [3.65, 300]$

- Past: Scaling performance using custom dataflow architectures
  - Demonstrated 35x performance/power with dataflow architectures on FPGA
Dataflow architectures to scale performance

10Gbps demonstrated with a 64b data path @ 156MHz using 3% of FPGA resources

80Gbps can be achieved by using a 512b @ 156MHz pipeline for example

Last Year: Scaling Capacity

- FPGAs enable custom memory architectures whereby storage media can be leveraged to their advantages

- Example:
  - SSDs combined with DDRx channels can be used to build high capacity & high performance key value stores
  - Concepts and early prototype to scale to 40TB & 80Gbps key value stores

Source: HotStorage 2015, Scaling out to a Single-Node 80Gbps Memcached Server with 40Terabytes of Memory
Object distribution on the basis of size

<table>
<thead>
<tr>
<th>Value Size (B)</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>768</th>
<th>1K</th>
<th>4K</th>
<th>8K</th>
<th>32K</th>
<th>1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>0.55</td>
<td>0.075</td>
<td>0.275</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>Twitter</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.85</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Wiki</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.29</td>
<td>0.008</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Flickr</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.9</td>
<td>0.05</td>
<td>0.03</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Advantages:
- Larger objects require larger storage
- Larger granular access to flash suits page-size access granularity of flash

Concerns:
- Large access latency on flash
- Variations in access bandwidth and latency between DRAM and flash

Source:
Dataflow architectures can accommodate high latency accesses without sacrificing throughput.

- In dataflow architectures: no limit to number of outstanding requests
- Flash can be serviced at maximum speed
Custom memory controllers

with out of order processing

Hybrid Memory Controller

---

© Copyright 2015 Xilinx
Today:
Networked Object Storage Board with MPSoC

50Gbps key value store with 2TB, 25W

Unconventional memory architecture to achieve high capacity while maintaining performance
Machine Learning:
Top-5 accuracy image classification

*Image-Net Large-Scale Visual Recognition Challenge (ILSVRC*)

CNNs deliver super-human accuracy

Human @95%

* http://image-net.org/challenges/LSVRC/
**http://www.slideshare.net/NVIDIA/nvidia-ces-2016-press-conference, pg 10
Compute and Memory Requirements

CNNs are highly compute and highly memory intensive. GPUs deliver highest performance for AlexNet with 4000+ fps.

<table>
<thead>
<tr>
<th>CNN for ImageNet datasets</th>
<th>Memory (SP) [MB]</th>
<th>Operations [MOPS]</th>
<th>Operational Intensity [OPS:B]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet – convolutions only</td>
<td>9.3</td>
<td>1332</td>
<td>143</td>
</tr>
<tr>
<td>AlexNet – complete</td>
<td>244</td>
<td>1456</td>
<td>5.97</td>
</tr>
<tr>
<td>VGG-16</td>
<td>552</td>
<td>30823</td>
<td>55.84</td>
</tr>
<tr>
<td>GoogleNet</td>
<td>27.2</td>
<td>1502</td>
<td>55.24</td>
</tr>
</tbody>
</table>
Emerging: Low-Precision Networks

- Reducing precision is shown today to work to 6b
  - 50x reduction in model size (no external memory needed) [1]
- Reducing to the extreme: binary neural networks (BNNs)

“The performance gap between the floating-point and {ternary fixed-point: -1, 0, +1} networks almost vanishes {with a big/complex enough network}.” [2]

## Binary and Almost Binary Networks

### Accuracy (published & reproduced results)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>FP32</th>
<th>BNN</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>99%</td>
<td>99%</td>
<td>[1]</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>92%</td>
<td>90%</td>
<td>[1]</td>
</tr>
<tr>
<td>ImageNet (GoogleNet arch)</td>
<td>90% top-5</td>
<td>86% top-5</td>
<td>[2] binary weights</td>
</tr>
<tr>
<td>ImageNet (DoReFaNet)</td>
<td>56% top-1</td>
<td>50% top-1</td>
<td>[4] 2-bit activations</td>
</tr>
</tbody>
</table>


Roofline model for KU115 (ADM-PCIE-8K5) & CNNs

Xilinx FPGA Rooflines

- AlexNet (complete)
- VGG-16
- LeNet-5
- KU115-SP
- GoogleNet
- SqueezeNet/FireCaffe
- KU115-16bint
- KU115-8bint
- KU115-1b

26 GByte/sec

BNN (avoid mem bw)

BNN – 2.5 LUTS/OP

2 Tops peak, 16b
Lab setup using the MNIST dataset, Zynq chip

First prototype in Xilinx labs Dublin* shows
In hardware: 12Mfps for MNIST, 2usec latency, ~7.4Watt

*Yaman Umuroglu (NTNU, Xilinx); Nicholas Fraser (University of Sydney, Xilinx); Giulio Gambardella (Xilinx Research); Michaela Blott (Xilinx Research)
Introductions

Industry landscape

Platform characterization

Some unconventional examples

Summary & future plans
Summary & Future Plans

➢ Trend towards unconventional architectures
  – A diversification of increasingly heterogeneous system

➢ Characterization leveraging Berkeley roofline models
  – Visualizes application suitability for different accelerators and performance estimation

➢ Some unconventional examples

➢ In collaboration with leading customers, partners and universities

➢ Future:
  – Facilitate ease of use for reconfigurable computing
  – Bring clarity & understanding on applications
Thank You
mblott@xilinx.com

Any questions?