The First Workshop on the Intersections of Computer Architecture and Reconfigurable Logic (CARL 2010), Atlanta, Georgia - Sunday, December 5, 2010, Co-located with MICRO-43

The Workshop on the Intersections of Computer Architecture and Reconfigurable Logic (CARL) is a new forum for presenting FPGA and reconfigurable logic research relevant to a computer architecture audience. In recent years, there has been renewed interest in reconfigurable computing, driven by the need for greater computing performance and, at the same time, better power and energy efficiency. Reconfigurable computing is a key technology candidate to efficiently leverage exponential device scaling beyond current multicore processors.

This full-day workshop will be held on Sunday, December 5, 2010, co-located with MICRO-43 in Atlanta, Georgia. The meeting will include keynote presentations, research presentations and a brainstorming panel.

Program of Invited Presentations

Two categories of submissions were solicited for review, (1) new unpublished manuscripts and (2) audience-appropriate revisions of papers already published or under review outside of traditional computer architecture forums. (See Call for Papers below.) Each 4–6-page submission was assigned to 4 members of the program committee for review. At the end, the program committee invited 9 out of the 20 submitted papers for presentation at the CARL Workshop. Submissions selected for presentation at CARL are not published.

The workshop will be held on Sunday, December 5th in Room 1458, Klaus Advanced Computing Building, Georgia Tech.

- 8:45-10:00 Keynote
  - Welcome, Derek Chau, Joel Emer and James C. Hoe
  - Co-Designing a COTS Reconfigurable Exascale Computer, Steven J. Wallach (Convey Computer) (PDF)
  - 10:00-10:30 Coffee break
  - 10:30-12:00 Computing Abstractions (Kees Vissers, Xilinx)
  - Rethinking FPGA Computing with a Many-Core Approach, John Wawrzynak (UCB); Mingjie Lin (UCB); Ilia Lebedev (UCB); Shaoyi Chang (UCB); Daniel Burke (UCB) (PDF)
  - A Model for Programming Large-Scale Configurable Computing Applications, Carl Ebeling (University of Washington); Scott Hauck (University of Washington); Corey Olson (University of Washington); Marie Kim (University of Washington); Cooper Clausen (University of Washington); Baris Kogon (University of Washington) (PDF)
  - CoRAM: An In-Fabric Memory Abstraction for FPGA-based Computing, Eric Chung (Carnegie Mellon University); James Hoe (Carnegie Mellon University); Ken Mai (Carnegie Mellon University) (PDF)
- 12:00-1:30 Lunch
- 1:30-4:00 Languages and Environments (Gabriel Scheller, Intel)
Accelerating Deep Convolutional Neural Networks Using Specialized Hardware in the Datacenter

Kalin Ovtcharov, Olatunji Ruwase, Joo-Young Kim, Jeremy Fowers, Karin Strauss, Eric S. Chung
Top Row: Eric Peterson, Scott Hauck, Aaron Smith, Jan Gray, Adrian M. Caulfield, Phillip Yi Xiao, Michael Haselman, Doug Burger

Bottom Row: Joo-Young Kim, Stephen Heil, Derek Chiou, Sitaram Lanka, Andrew Putnam, Eric S. Chung


Huge thanks to our partners at ALTERA
Agenda

• Deep Learning on Catapult
• Academic Outreach Program
Deep Learning: The “Next Big Thing”?

- Significant advances in
  - Computer vision
  - Speech recognition
  - Natural language processing
  - Intelligent agents
  - Etc.

- State-of-the-art neural nets
  - Convolutional Neural Networks (CNNs)
  - Deep Neural Networks (DNNs)
  - ...
Goal: Deep Learning as a Cloud Service

• ML in the cloud
  • Leverage economies of scale in shared cloud infrastructure
  • Support training of new models
  • Deploy pre-trained models (e.g., classify images in OneDrive)
  • Scale training and deployment up to hundreds of thousands of machines

• Expose through cloud providers
  • Microsoft AzureML
  • Amazon ML-as-a-Service
  • Google Prediction API
Challenges

• Training very slow on conventional CPUs
  • Up to months
  • Yet, most cloud services built on commodity CPUs and components

• Deploying trained models also compute-intensive

• GPUs preferred by many practitioners but
  • Difficult to scale beyond 16-32 nodes
  • Limited in memory capacity (affecting model size and accuracy)
  • Too power-intensive for datacenters
  • Expensive to maintain
  • Have reliability issues
The Efficiency of Specialized Hardware

Source: Bob Broderson, Berkeley Wireless group
Datacenter Environment

• Software services change monthly
• Machines last 3 years, purchased on a rolling basis
• Machines repurposed \(~\frac{1}{2}\) way into lifecycle
• Little/no HW maintenance, no accessibility

• Homogeneity is highly desirable

The paradox: Specialization and homogeneity
Our Design Requirements

**Don’t Cost Too Much**
- <30% Cost of Current Servers

**Don’t Burn Too Much Power**
- <10% Power Draw (25W max, all from PCIe)

1. Specialize HW with an FPGA Fabric
2. Keep Servers Homogeneous

**Don’t Break Anything**
- Work in existing servers
- No Network Modifications
- Do not increase hardware failure rate
MICROSOFT SUPERCHARGES BING SEARCH WITH FPGAS

95% Query Latency vs. Throughput

2x Increase in Throughput
29% Latency Reduction

SW + FPGA
< 30% Cost
< 25 W Power
0 HW Failures

http://www.wired.com/2014/06/microsoft-fpga/
Catapult: An Elastic Reconfigurable Fabric for Datacenters
Catapult FPGA Accelerator Card

- Altera Stratix V D5
- 172,600 ALMs, 2,014 M20Ks, 1,590 DSPs
- PCIe Gen 3 x8
- 8GB DDR3-1333
- Powered by PCIe slot
- Torus Network
Microsoft Open Compute Server

- Two 8-core Xeon 2.1 GHz CPUs
- 64 GB DRAM
- 4 HDDs @ 2 TB, 2 SSDs @ 512 GB
- 10 Gb Ethernet
- No cable attachments to server

Air flow

200 LFM
68 °C Inlet
Scalable Reconfigurable Fabric

- 1 FPGA board per Server
- 48 Servers per ½ Rack
- 6x8 Torus Network among FPGAs
  - 20 Gb/s over SAS SFF-8088 cables

Data Center Server (1U, ½ width)
FPGA Accelerator for Bing Ranking

Document

FE: Feature Extraction

FFE: Free-Form Expressions

MLS: Machine Learning Scoring

Score

8-Stage Pipeline

FPGA 0

FPGA 1

FPGA 2

FPGA 3

FPGA 4

FPGA 5

FPGA 6

FPGA 7

Route to Head

Document Scoring Request

Return Score

Compute Score

Ranking Servers

Server

Server

Server

Server

Server

Server

Server
1,632 Server Pilot Deployed in a Production Datacenter
Scalable Deep Learning on Catapult

• Provide excellent **performance** and **accuracy** at fraction of cost of commodity CPUs

• Leverage abundant FPGA resources in MSFT’s datacenters for scaling up machine learning and model deployment

• Target high-valued kernels and expose to practitioners as composable SW libraries
Image Classification with Deep CNN

3-D Convolution and Max Pooling

Dense Layers

"Dog"

* Krizhevsky et al, NIPS’12
3-D Convolution

Input  Model Weights  Output
3-D Convolution

Input  Model Weights  Output
3-D Convolution

Input  Model Weights  Output
3-D Convolution

Input  Model Weights  Output
3-D Convolution

Input

Model Weights

Output
3-D Convolution

Input  Model Weights  Output
3-D Convolution

Input

Model Weights

Output
3-D Convolution

Input

Model Weights

Output
3-D Convolution

Input

Model Weights

Output
3-D Convolution

Input  Model Weights  Output
3-D Convolution

Input  Model Weights  Output
3-D Convolution

Input

Model Weights

Output
3-D Convolution and Max Pooling

- Input Feature Map
- Convolution Output
- Max Pooled Output (Optional)

*N, k, H, and p may vary across layers*

N = input height and width  
k = kernel height and width  
D = input depth  
H = # feature maps  
S = kernel stride
CNN Accelerator Building Block

- **Configurable**
  - Numerical precision (static)
  - Number of layers
  - Layer dimensions
  - Stride and pooling

- **Scalable**
  - Can compose multiple engines together over Catapult network

- **Efficient**
  - Minimize memory bandwidth via data re-distribution NoC
  - On-chip per-row broadcast
Scalable Deep Learning on Catapult
## CNN Classification Performance

<table>
<thead>
<tr>
<th></th>
<th>CIFAR-10</th>
<th>ImageNet 1K</th>
<th>ImageNet 22K</th>
<th>FPGA or GPU Power</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Server + Stratix V D5</strong></td>
<td>2318 images/s</td>
<td>134 images/s</td>
<td>91 images/sec</td>
<td>25W</td>
</tr>
<tr>
<td><strong>Server + Arria 10 GX1150</strong></td>
<td>-</td>
<td>~233 images/s (projected)</td>
<td>~158 images/sec (projected)</td>
<td>25W</td>
</tr>
<tr>
<td><strong>Best prior CNN on FPGA [FPGA’15]</strong></td>
<td>-</td>
<td>46 images/s</td>
<td>-</td>
<td>18W</td>
</tr>
<tr>
<td><strong>Caffe+cuDNN on Tesla K20</strong></td>
<td>-</td>
<td>376 images/s</td>
<td>-</td>
<td>225W</td>
</tr>
<tr>
<td><strong>Caffe+cuDNN on Tesla K40</strong></td>
<td>-</td>
<td>824 images/s</td>
<td>-</td>
<td>225W</td>
</tr>
</tbody>
</table>

See whitepaper @ [http://research.microsoft.com/apps/pubs/?id=240715](http://research.microsoft.com/apps/pubs/?id=240715)
DEMO
Related Work

• ASICs
  • [Holler’90], [Chen’14], [Cavigelli’15], etc.

• FPGAs
  • [LeCun’09], [Farabet’10], [Aysegul’13], [Gokhale’15], [Zhang’15], etc.

• GPUs/Appliances
  • Nvidia DIGITS, Ersatz, etc.

• Existing solutions not cloud-friendly
  • ASICs, GPUs, and appliances difficult to justify at scale in datacenter
  • ASICs lack flexibility
  • Existing FPGA designs target single FPGA
Conclusions

• Specialized HW for ML is promising for the cloud
  • Inter-networked FPGAs provide scalability, homogeneity, and flexibility
  • Offers compelling performance relative to conventional systems

• Future work
  • Multi-FPGA training pipeline
  • Prototyping on Arria 10
  • OpenCL

• Questions?
  • Eric Chung (erchung@microsoft.com)
Agenda

• Deep Learning on Catapult
• Academic Outreach Program
Academic Outreach

• Purpose
  • Create research eco-system around FPGAs in the data center that will include access to our enabling IP (drivers, shell), potentially research funding, and contests.

• Resources
  • Microsoft will provide FPGA boards, tools, and IP to academics
  • Access to full 48-server machines in a shared cloud

• Look for announcements @
  • http://research.microsoft.com/en-us/projects/catapult