Inference Architectures @Xilinx

Graham Schelle, PhD
Principal Engineer
Xilinx Research Labs
Twitch Chooses Xilinx to Enable its Broadcast-quality Livestream of eSports

Children’s Hospital of Philadelphia And Edico Genome Achieve Fastest-Ever Analysis Of 1,000 Genomes
Agenda

- Xilinx Adaptive Architectures
- Inference Architectures
- Open Source
Traditionally, FPGAs for massively data-parallel applications
Traditionally, FPGAs for massively data-parallel applications

In 2011, Zynq introduced (ZU+ in 2015) ARM CPUs added for embedded applications
In 2018, Alveo introduced Accelerator cards for data center workloads

Coming in 2019, Versal Platform Adaptive compute acceleration platform (ACAP)
Inference Architectures
Increasing, Evolving Workloads

- New acceleration needs & algorithms
- ML “infused” in many applications
- Adaptable HW a key benefit

Andrej Karpathy on Twitter
Inference Architectures – Evolving Frameworks

Increasing, Evolving Workloads
- New acceleration needs & algorithms
- ML “infused” in many applications
- Adaptable HW a key benefit

Move to Lower Precision
- ML inference moving to INT8 & lower
- Better Perf/W with similar accuracy
- Xilinx devices natively support variable precision

Compressed Networks
- Higher performance with reduced compute / memory needs
- Pruning & load balancing to match network requirements

Andrej Karpathy on Twitter
Increasing, Evolving Workloads

- New acceleration needs & algorithms
- ML “infused” in many applications
- Adaptable HW a key benefit

Move to Lower Precision

- ML inference moving to INT8 & lower
- Better Perf/W with similar accuracy
- Xilinx devices natively support variable precision

Compressed Networks

- Higher performance with reduced compute / memory needs
- Pruning & load balancing to match network requirements
Inference Architectures – Evolving Workloads

1. Inference is hard
2. Huge variation in compute and memory requirements
3. Models typically don’t fit into cache

Increasing, Evolving Workloads

- New acceleration needs & algorithms
- ML “infused” in many applications
- Adaptable HW a key benefit

Move to Lower Precision

- ML inference moving to INT8 & lower
- Better Perf/W with similar accuracy
- Xilinx devices natively support variable precision

Compressed Networks

- Higher performance with reduced compute / memory needs
- Pruning & load balancing to match network requirements
Inference Architectures – Precision vs Power

FPGA:

Target Device ZU7EV ● Ambient temperature: 25 °C ● 12.5% of toggle rate ● 0.5 of Static Probability ● Power reported for PL accelerated block only

ASIC:

Source: Bill Dally (Stanford), Cadence Embedded Neural Network Summit, February 1, 2017

Rybalkin, V., Pappalardo, A., Ghaffar, M.M., Gambardella, G., Wehn, N. and Blott, M. "FINN-L: Library Extensions and Design Trade-off Analysis for Variable Precision LSTM Networks on FPGAs."
Xilinx Cloud Inference - ML Suite Overlays with xDNN

- Built in Programmable Logic
- High Utilization, Thput or Latency Variants
- CPU offload for new layer exploration

© Copyright 2018 Xilinx
“Learning both Weights and Connections for Efficient Neural Networks”, NeurIPS 2015

“EIE: Efficient Inference Engine on Compressed Deep Neural Network”, ISCA 2016

“ESE: Efficient Speech Recognition Engine with Compressed LSTM on FPGA”, FPGA 2017
Xilinx Edge Inference - DeePhi

“Learning both Weights and Connections for Efficient Neural Networks”, NeurIPS 2015

“EIE: Efficient Inference Engine on Compressed Deep Neural Network”, ISCA 2016

“ESE: Efficient Speech Recognition Engine with Compressed LSTM on FPGA”, FPGA 2017

“Xilinx Announces the Acquisition of DeePhi Tech
Deal to Accelerate Data Center and Intelligent Edge Applications
Jul 17, 2018

© Copyright 2018 Xilinx
Integrate the AWS Cloud with Responsive Xilinx Machine Learning at the Edge

Richard Elberger
Partner Solutions Architect
AWS

Wesley Skeffington
Principal Architect – Industrial & Medical
Xilinx
Xilinx and Open Source
Xilinx and Open Source

PYNQ

Quantized Neural Networks

Xilinx Runtime for PCIe Attached FPGAs

...More on www.github.com/Xilinx
Xilinx and Open Source

PYNQ

Quantized Neural Networks

Xilinx Runtime for PCIe Attached FPGAs

...More on www.github.com/Xilinx
Python is increasingly the Language of Choice

Top Programming Languages, IEEE Spectrum, July’18

To date


https://stackoverflow.blog/2017/09/06/incredible-growth-python/
Python is increasingly the Language of Choice

Top Programming Languages, IEEE Spectrum, July’18

Python is listed as an embedded language for the first time


To date

https://stackoverflow.blog/2017/09/06/incredible-growth-python/
Python is increasingly the Language of Choice

Top Programming Languages,
IEEE Spectrum, July’18

Python is the fastest growing language: driven by data science, AI, ML and academia


Python is listed as an embedded language for the first time

https://stackoverflow.blog/2017/09/06/incredible-growth-python/

To date
Python Productivity for Zynq

- Jupyter web server
- IPython kernel
- Ubuntu-based Linux
- ARM A9 / A53
- Overlays/designs
- ZU+ Fabric
Python Productivity for Zynq

Jupyter notebooks, browser-based interface

- Jupyter web server
- IPython kernel
- Ubuntu-based Linux
- ARM A9 / A53
  - Overlays/designs
  - ZU+ Fabric
Python Productivity for Zynq

Jupyter notebooks, browser-based interface

PYNQ enables JupyterLab on Zynq and ZU+

<table>
<thead>
<tr>
<th>Jupyter web server</th>
<th>Overlays/designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPython kernel</td>
<td></td>
</tr>
<tr>
<td>Ubuntu-based Linux</td>
<td>ZU+ Fabric</td>
</tr>
<tr>
<td>ARM A9 / A53</td>
<td></td>
</tr>
</tbody>
</table>
Python Productivity for Zynq

Jupyter notebooks, browser-based interface

PYNQ enables JupyterLab on Zynq and ZU+

FPGA designs delivered as Python packages

<table>
<thead>
<tr>
<th>Jupyter web server</th>
<th>Overlays/designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPython kernel</td>
<td></td>
</tr>
<tr>
<td>Ubuntu-based Linux</td>
<td></td>
</tr>
<tr>
<td>ARM A9 / A53</td>
<td>ZU+ Fabric</td>
</tr>
</tbody>
</table>
Python Productivity for Zynq

Jupyter notebooks, browser-based interface

PYNQ enables JupyterLab on Zynq and ZU+

FPGA designs delivered as Python packages

<table>
<thead>
<tr>
<th>Jupyter web server</th>
<th>Overlays/designs</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPython kernel</td>
<td>ZU+ Fabric</td>
</tr>
<tr>
<td>Ubuntu-based Linux</td>
<td></td>
</tr>
<tr>
<td>ARM A9 / A53</td>
<td></td>
</tr>
</tbody>
</table>

Delivered as SD Card image
PYNQ Community – ML, Non-ML & Academic Partners
PYNQ Community – ML, Non-ML & Academic Partners
PYNQ Community – ML, Non-ML & Academic Partners
Xilinx open source engagements related to today’s TVM meeting
Xilinx open source engagements related to today’s TVM meeting

University of Washington

UC San Diego

UC Berkeley

Xilinx Research

MicroPython

© Copyright 2018 Xilinx
Finally, Xilinx & building new open source communities...

Get started with Alveo accelerator card applications today on the Nimbix Cloud

Nimbix has partnered with Xilinx to provide developers and engineers a trial account that provides up to 100 hours of free time on the Nimbix Cloud using Xilinx Tools and Accelerators.

- pynq.io/community
- DAC2019 Design Contest
- OpenHW Design Contest

Cloud Free Trials
Summary

Xilinx
Great for exploring and deploying inference

Xilinx Open Source
We’re actively engaging with TVM and other communities

Email: graham.schelle@xilinx.com
Visit: Boulder, Colorado
Adaptable. Intelligent.
Edge to Cloud Inference – Automotive
Edge to Cloud Inference – Automotive

ADAS/AD Centra Module
Edge to Cloud Inference – Automotive

- Surround-View Camera Back
- Short-Range Radar
- Surround-View Camera Right
- Surround-View Camera Front
- Short-Range Radar
- Long-Range Lidar
- Surround-View Camera Left
- Forward-Looking Camera
- Drive Monitor Camera

ADAS/AD Central Module
Edge to Cloud Inference – Xilinx Platforms

Edge Devices
Custom I/O, ARM CPUs

Cloud Platforms
Power Efficient, PCIe, Networking
Example IIoT Control Rates

Natural Frequency (Hz)

Control Rate

Gas Turbine
Wind Turbine
HVDC
Wind Converter
MRI Inverter

Difficult to control
Unreasonable / Uncontrollable

Example IIoT Latency/Data Example
Example IIoT Control Rates

Distance NYC to LA: 2,800 miles
Speed of light: 186,000 miles/s
Round trip: $2 \times \frac{2800}{186000} = 30$ms

Required Control Rate = 10ms
Edge to Cloud Inference – IIoT Latency/Data Example

Example IIoT Control Rates

- Distance NYC to LA: 2,800 miles
- Speed of light: 186,000 miles/s
- Round trip: \(2 \times \frac{2800}{186000} = 30\text{ms}\)
- Required Control Rate = 10ms

E.g. Power Plant @ 8TB/Month