Unconventional Compute Architectures for Enabling the Roll-Out of Deep Learning

Michaela Blott
Distinguished Engineer, Xilinx Research

Background

> Xilinx

- Fabless semiconductor company
- Founded in Silicon Valley in 1984
- Today:
  - – 3,500 employees
  - – $2.5B revenue
- Invented the FPGA

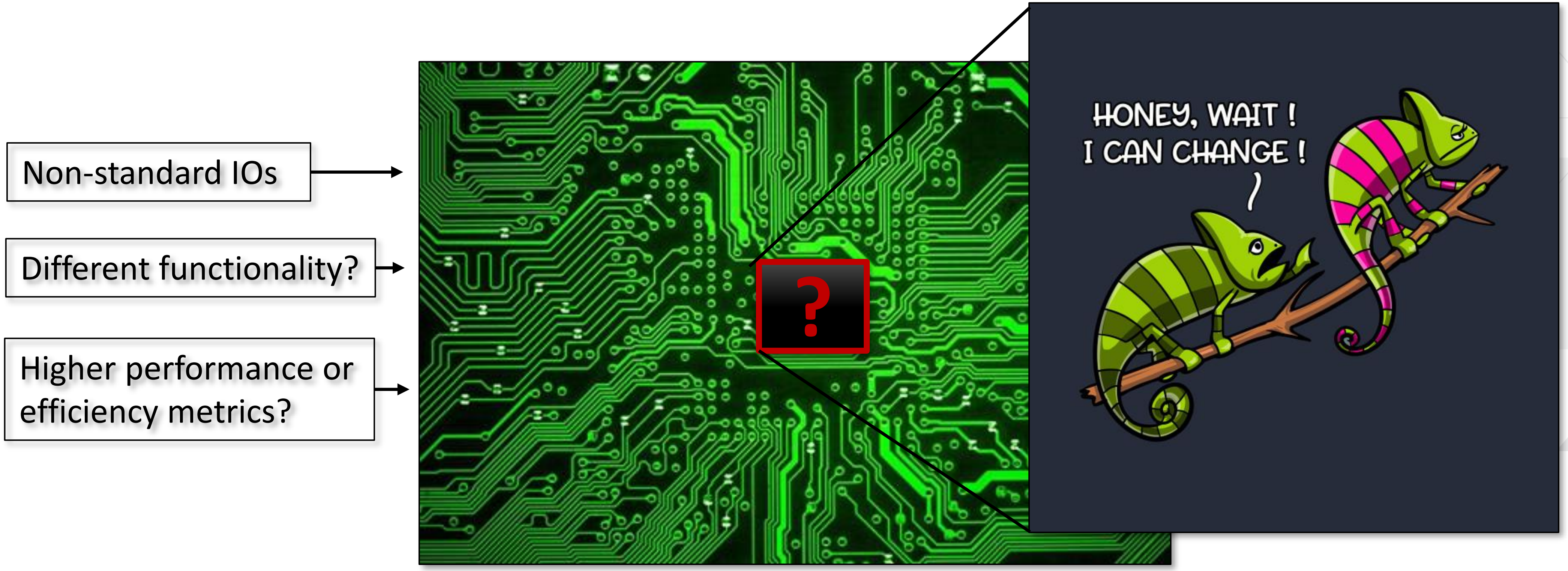
1st FPGA in 1985: XC2064
128 3-input LUTs

What are FPGAs?

Customizable, Programmable Hardware Architectures

- **The chameleon amongst the semiconductors…**

  - Customizes IO interfaces, compute architectures, memory subsystems to meet the application
- **Classic use case: Nothing else works, and you want to avoid ASIC implementation**
- **Recent use cases: Custom hardware architecture for performance or efficiency required**



Non-standard IOs

Different functionality?

Higher performance or efficiency metrics?

Trends Meet Technological Reality

HERE COMES THE SPORTS CAR AT 200 MILES PER HOUR!
HERE COMES A CEMENT TRUCK! LOOK OUT!
AND HERE COMES AN INFLAMMABLE CHEMICAL TRUCK! OH NO!!
THIS OUGHT TO BE GOOD.
12-18
WATTERSON

Roll-out of machine learning
"Potential to solve the unsolved"

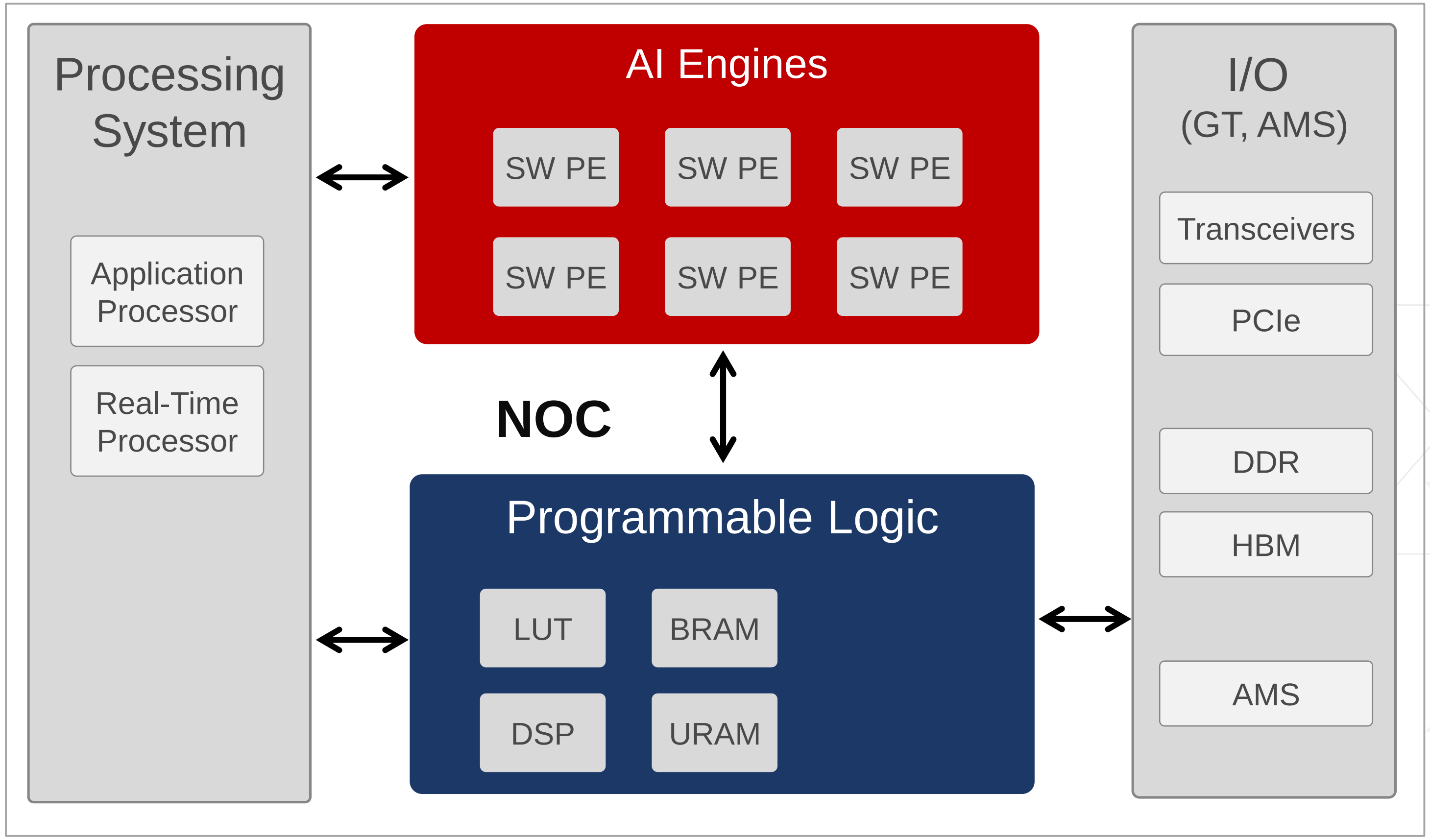
Explosion of Data
"genomical"

End of Moore's Law
End of Dennard Scaling

Era of heterogeneous computing has begun

- **Diversification of increasingly heterogeneous "unconventional" devices**
- **Moving away from van Neumann architectures**
- **Archit5ectural and algorithmic innovation is needed**

Increasingly Heterogeneous Devices
From the Xilinx World: Evolution of FPGAs to ACAPs

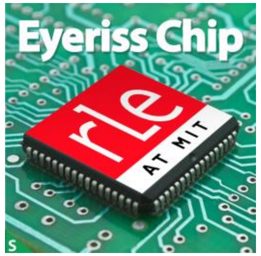


More Unconventional:
Customized Hardware for AI
DPU: Deep Learning Processing Unit

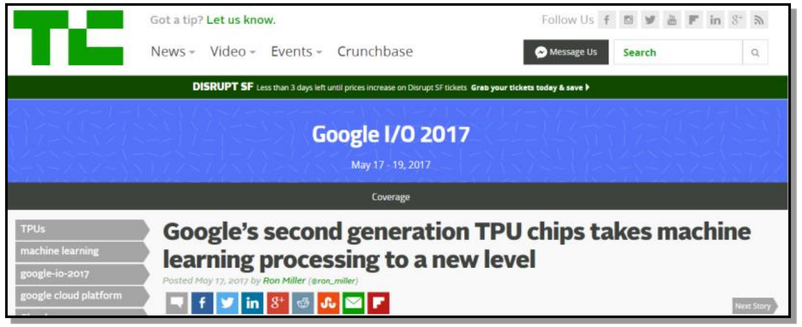
> Custom AI Accelerators (soft in FPGA and hard in ASICS)

WAVE® COMPUTING

cerebras



nervana


Got a tip? Let us know.

Follow Us

News Video Events Crunchbase

Message Us

Search

DISRUPT SF Less than 3 days left until prices increase on Disrupt SF tickets Grab your tickets today & save

Google I/O 2017

May 17 - 19, 2017

Coverage

TPUs

machine learning

google-io-2017

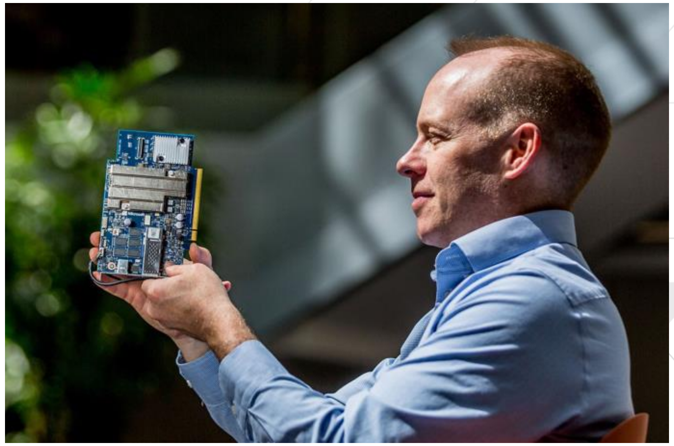
google cloud platform

Google's second generation TPU chips takes machine learning processing to a new level

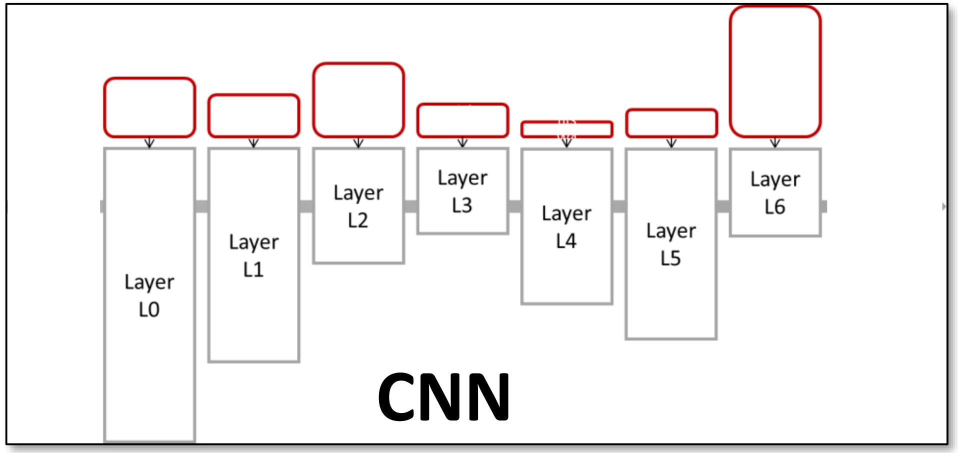
Posted May 17, 2017 by Ron Miller (@ron_miller)

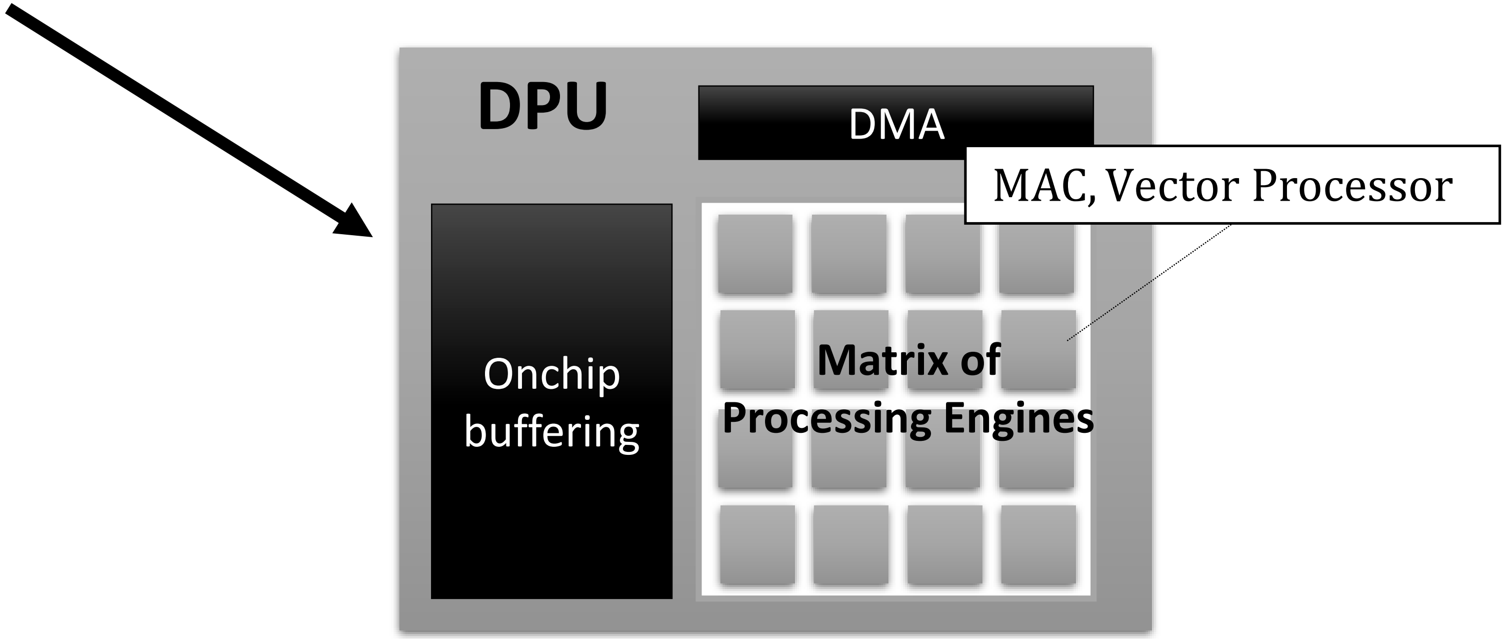
Next Story

KNUPATH


Microsoft Brainwave

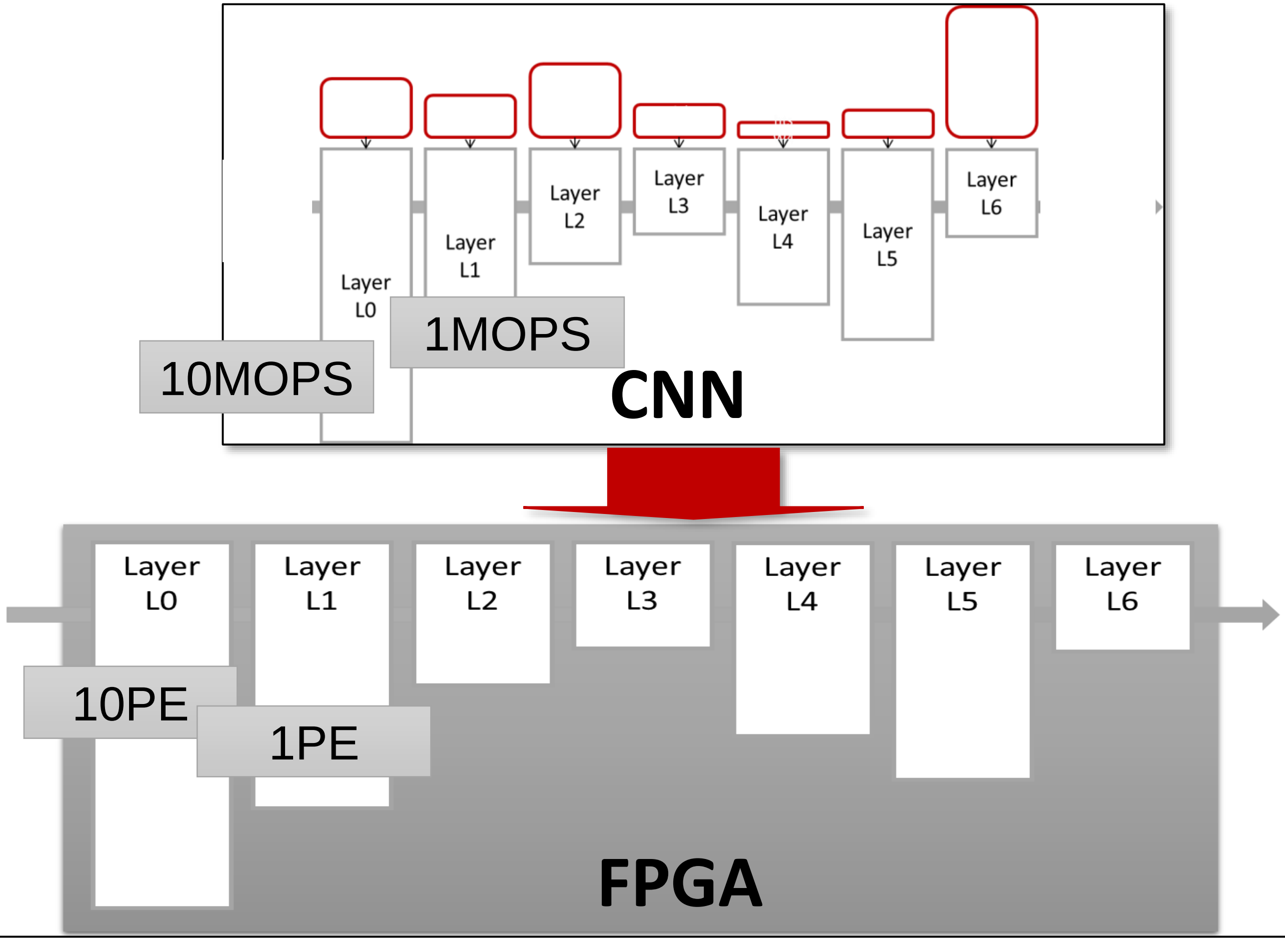
Popular DPU Architecture





“Layer by layer compute”

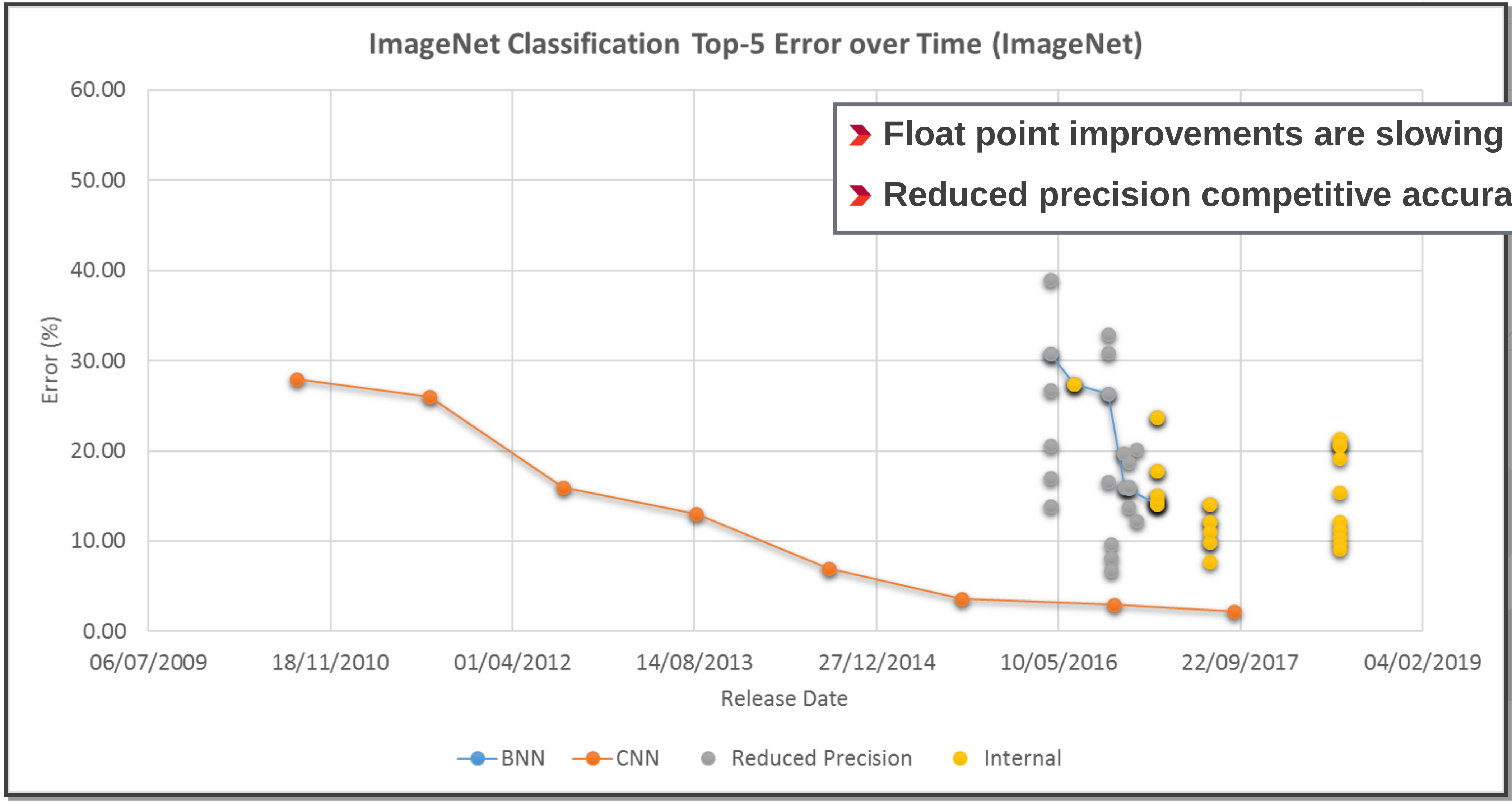
Even more unconventional: Custom-Tailored Hardware Architectures (Macro-Level) *Synchronous Dataflow*



"Hardware Architecture Mimics the NN Topology"

- **Customized feed-forward dataflow architecture to match network topology**
- **Higher compute and memory efficiency**

Further unconventional at the Micro-Architecture, leveraging
Floating Point to Reduced Precision Neural Networks

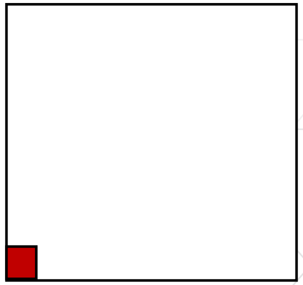


- Float point improvements are slowing down
- Reduced precision competitive accuracy

Reducing Precision

Scales Performance & Reduces Memory

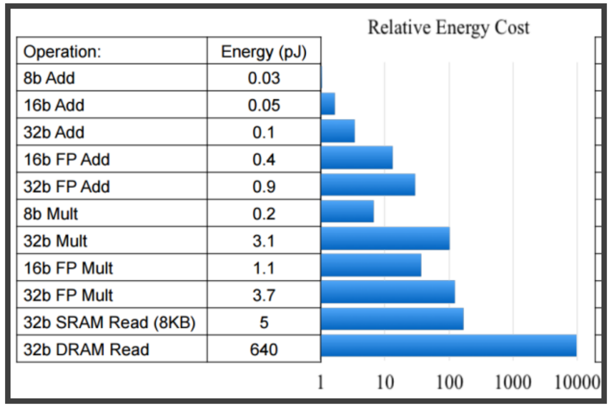
- **Reducing precision shrinks hardware cost**
  - Instantiate **100x** more compute within the same fabric
  - Thereby scale performance **100x**



- **Potential to reduce memory footprint**
  - NN model can stay on-chip => no memory bottlenecks

Precision	Modelsize [MB] (ResNet50)
1b	3.2
8b	25.5
32b	102.5

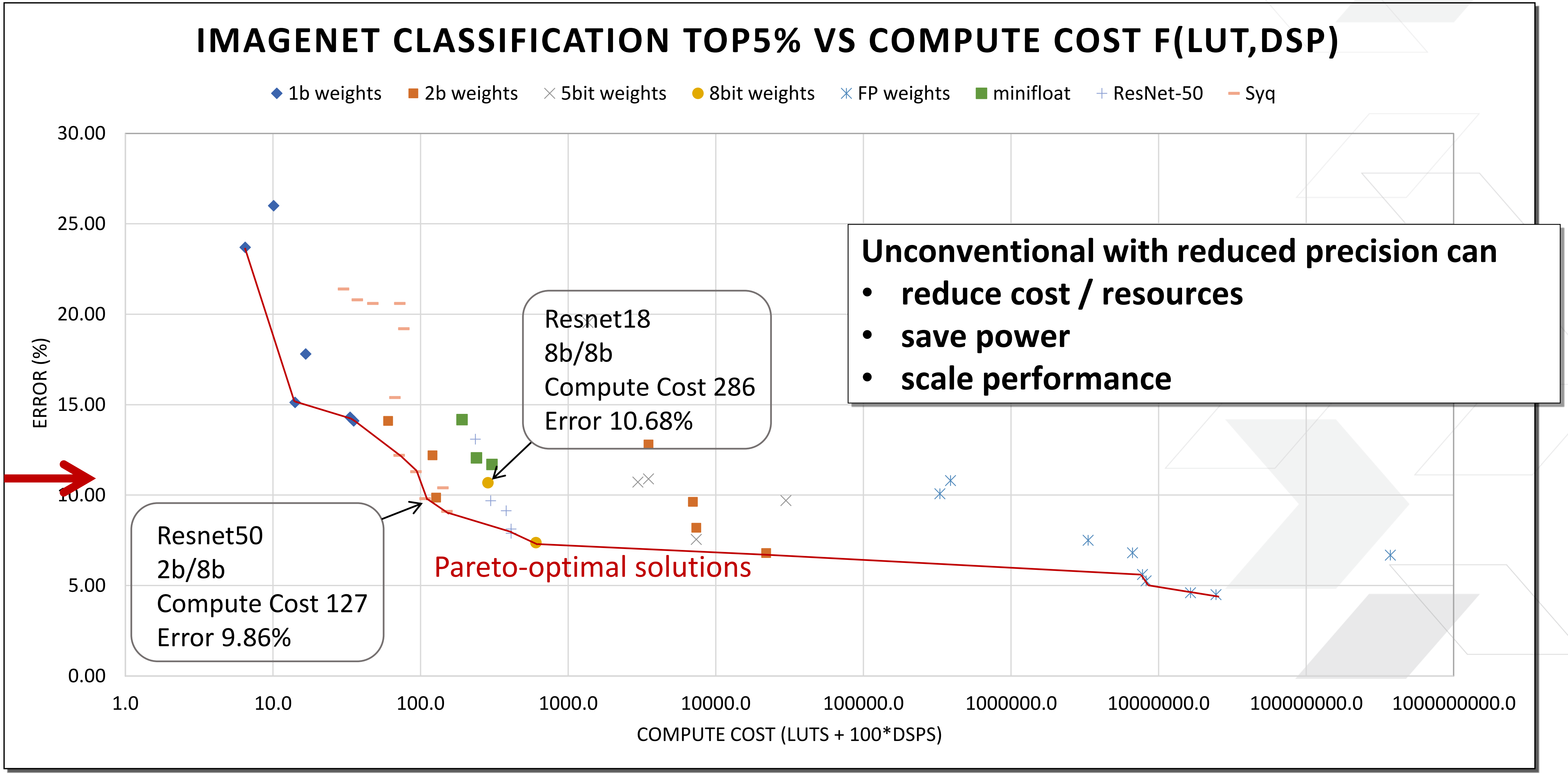
- **Reducing precision inherently saves power**



Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

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

Design Space Trade-Offs



Resnet18
8b/8b
Compute Cost 286
Error 10.68%

Resnet50
2b/8b
Compute Cost 127
Error 9.86%

Pareto-optimal solutions

Unconventional with reduced precision can

- **reduce cost / resources**
- **save power**
- **scale performance**

Summary

- Unconventional computing architectures emerge to help with the roll-out of deep learning
- Customized dataflow architectures and precisions provide dramatic performance scaling and energy efficiency
- Providing new exciting trade-offs within the design space

More information can be found at:
http://www.pynq.io/ml

Adaptable.
Intelligent.

More information can be found at:
http://www.pyng.io/ml

XILINX