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

Michaela Blott
Principal Engineer
Oct. 2018

XILINX

Background

> Xilinx

- Fabless semiconductor company
- Founded in Silicon Valley in 1984
- Today:
  - – 3,500 employees
  - – $2.25B revenue
- Invented the FPGA

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

What are FPGAs?

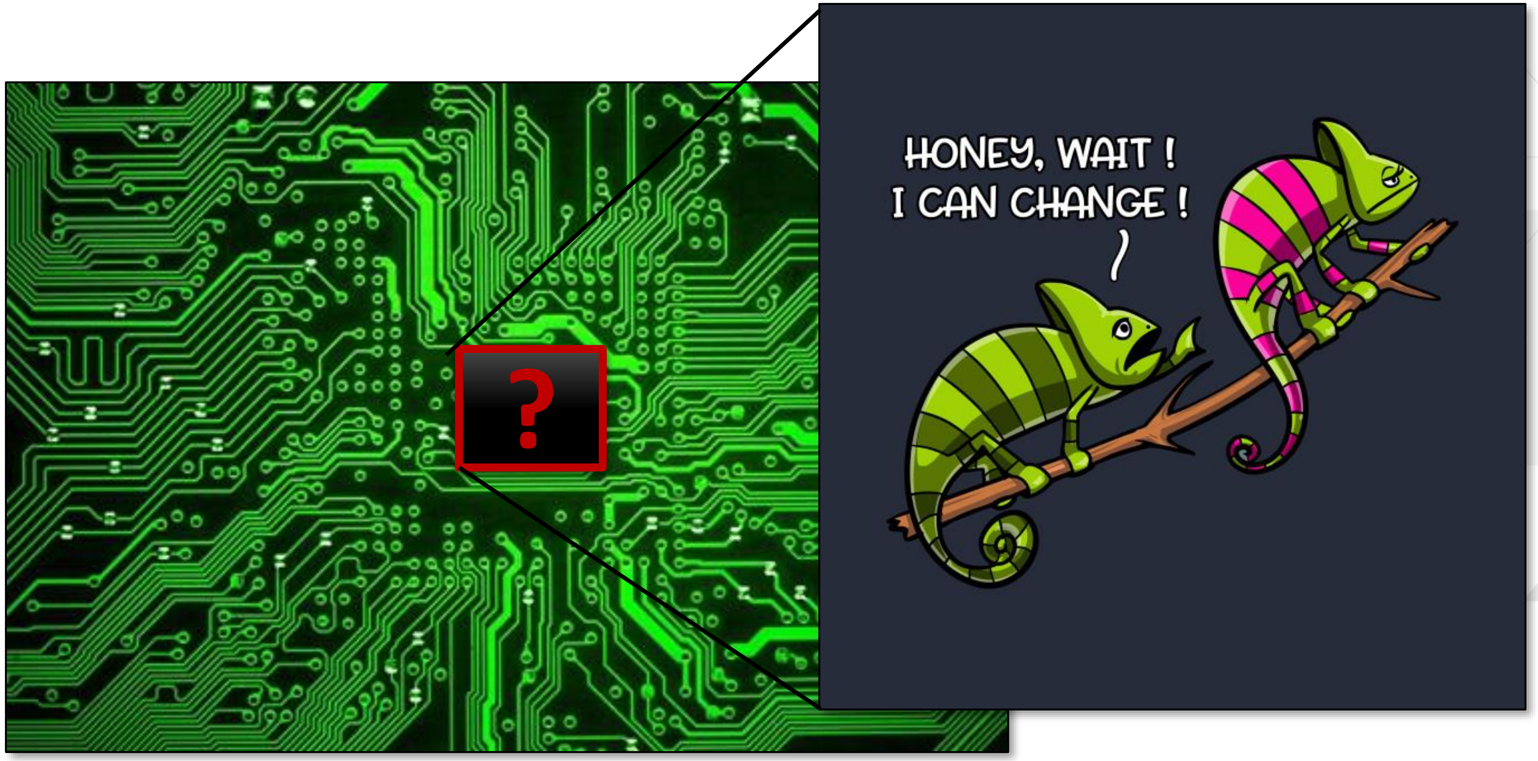
Customizable, Programmable Hardware Architectures

- **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?



XILINX

Context Machine Learning

Trends meeting Technological Reality

Mega-Trend:
The Rise of the Machine (Learning Algorithm)

1800 — 1900 — 2000

1. Mechanical
Steam powered mechanical production

2. Electrical
Mass production with electrical energy

3. Digital
Automated production

4. Virtual
Machine Learning
Industrial Internet

> **Potential to solve the unsolved problems**

   > Making solar energy economical, reverse engineering the brain (Jeff Dean, Google Brain 2017)

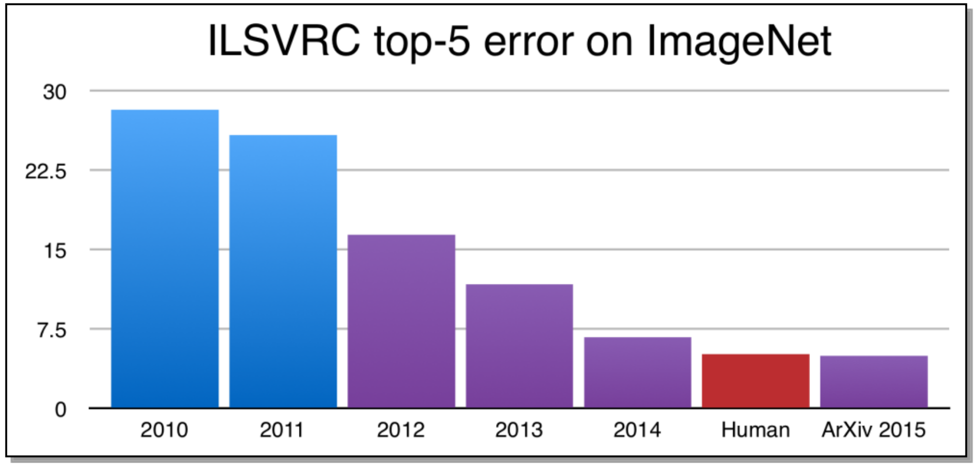
> **How can we computer architects help to enable the roll-out of these algorithms?**

XILINX

Convolutional Neural Networks (CNNs)

Why are they so popular?

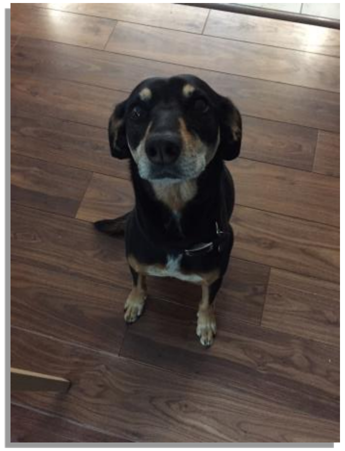
- Requires little or no domain expertise
- NNs are a "universal approximation function"
- If you make it big enough and train it enough
  - Can outperform humans on specific tasks



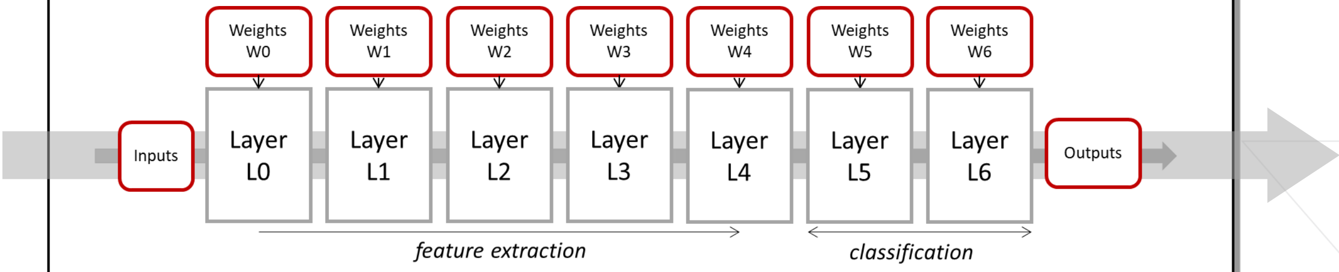
- Will increasingly replace other algorithms
  - unless for example simple rules can describe the problem
- Solve problems previously unsolved by computers
- And solve completely unsolved problems

Convolutional Neural Networks:
Forward Pass (Inference)

Input Image



Neural Network



Neural Network

Cat?

For ResNet50:

70 Layers
7.7 Billion operations
25.5 millions of weight

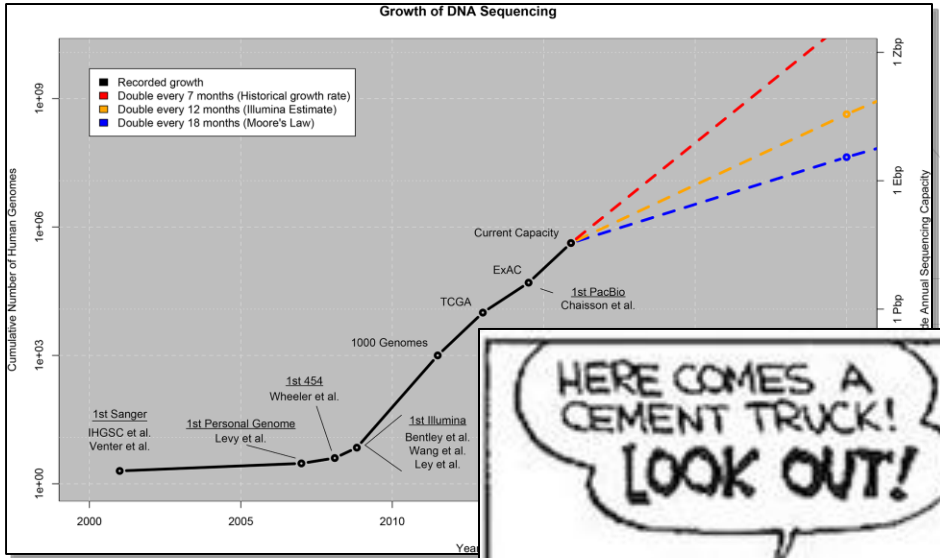
Basic arithmetic, incredible parallel but
Huge Compute and Memory Requirements

Mega-Trend:
Explosion of Data

> **Computing shifts towards cloud computing**

> **Data storage requirements explode**
>> #users
>> Photos => videos
>> DNA!

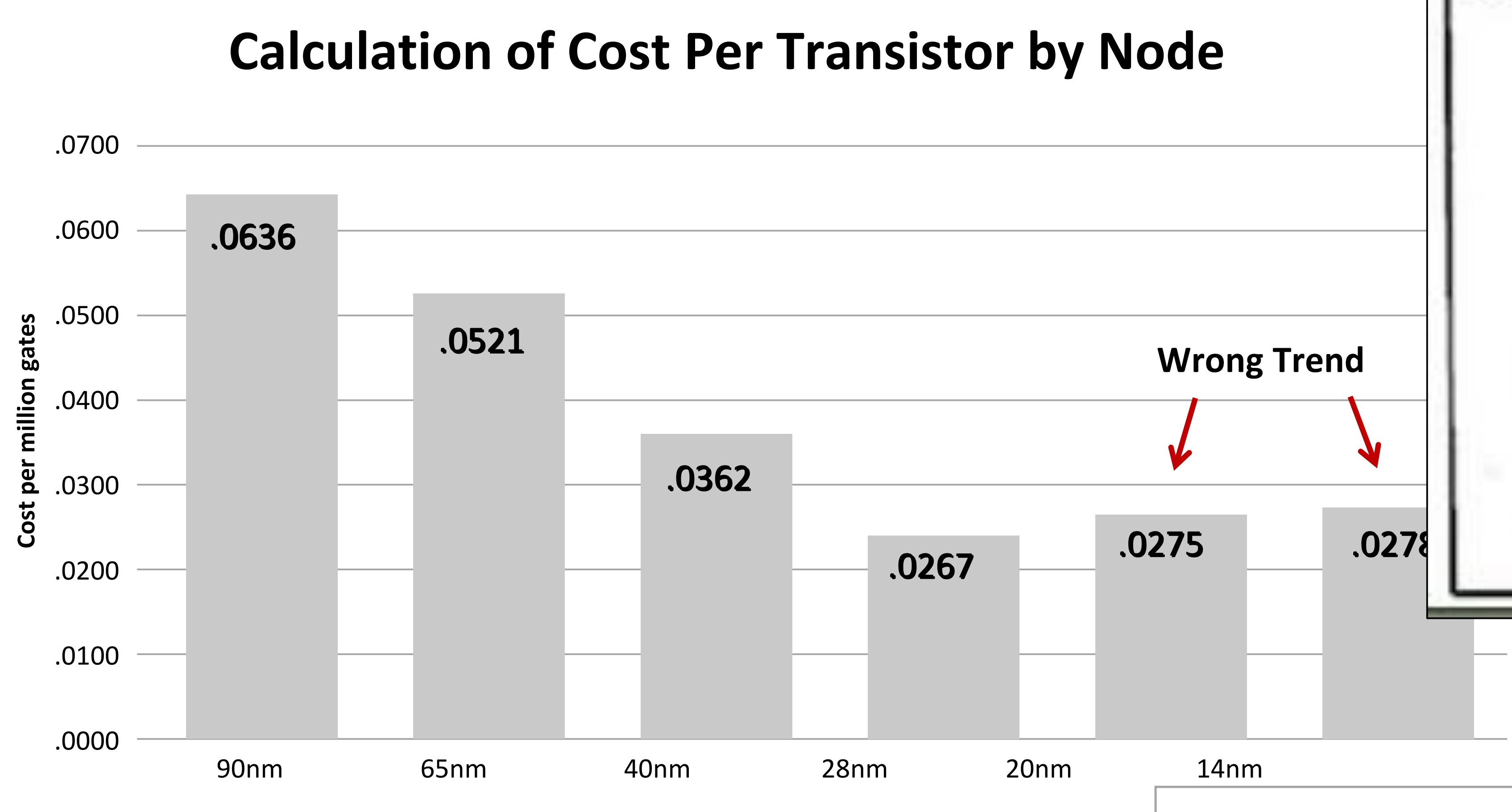
> **Big data problem:**
>> Gaining intelligence out of vast amounts of unstructured data using machine learning algorithms



Stephens, Zachary D., et al. "B genomical?." PLoS biology 13.

XILINX

Technology:
End of Moore's Law

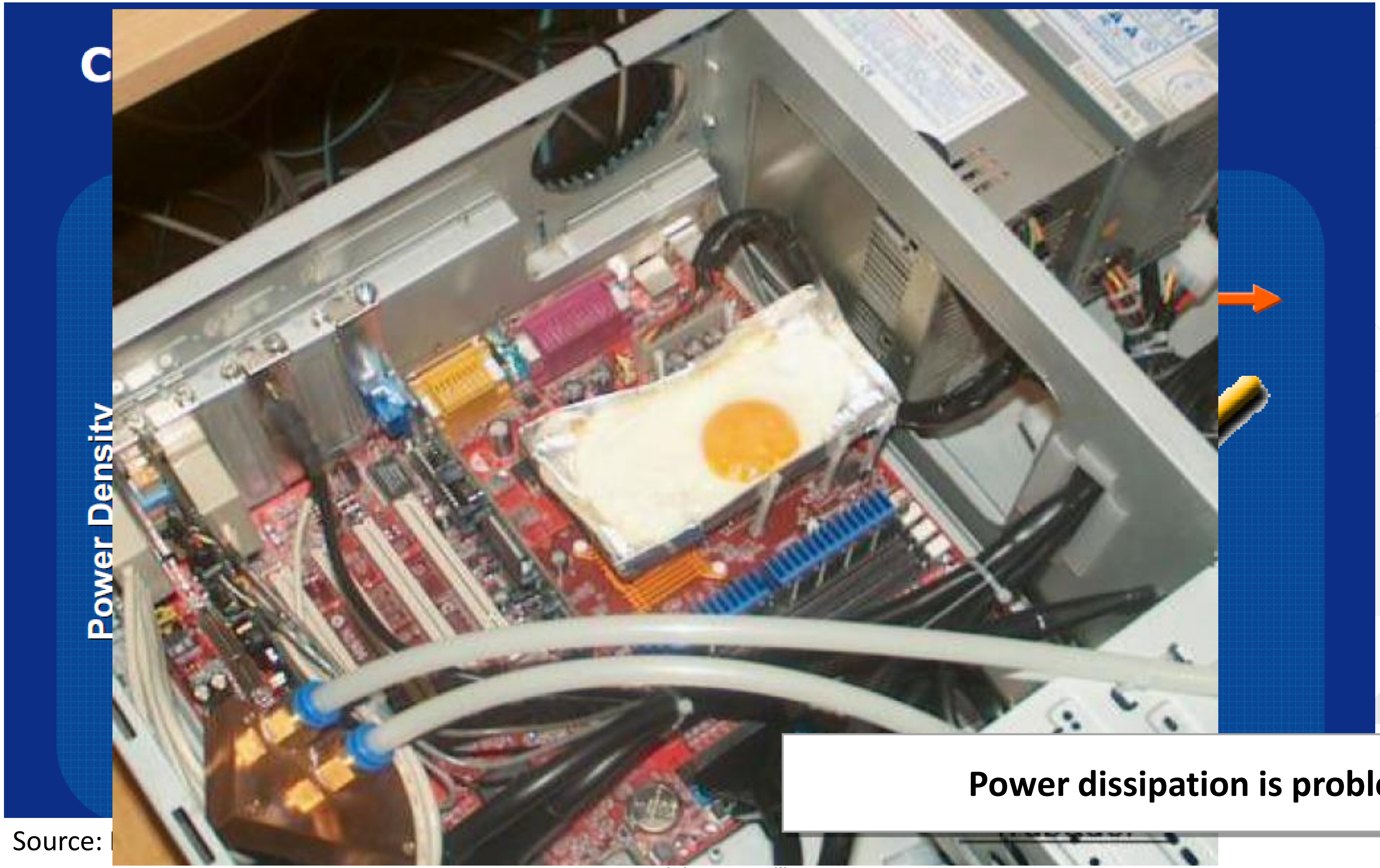


Source: IBS

Economics become questionable

XILINX

Technology:
End of Dennard Scaling



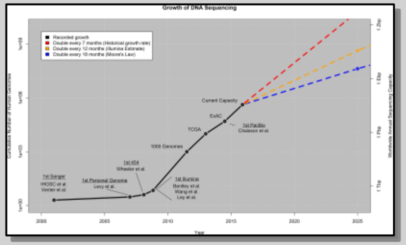
Power dissipation is problematic

Source:

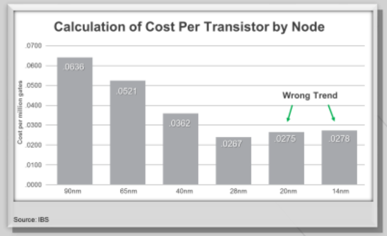
XILINX

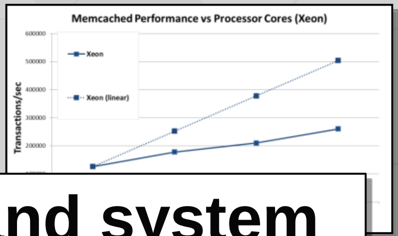
Era of Heterogeneous Compute using Accelerators

Trends



Technology



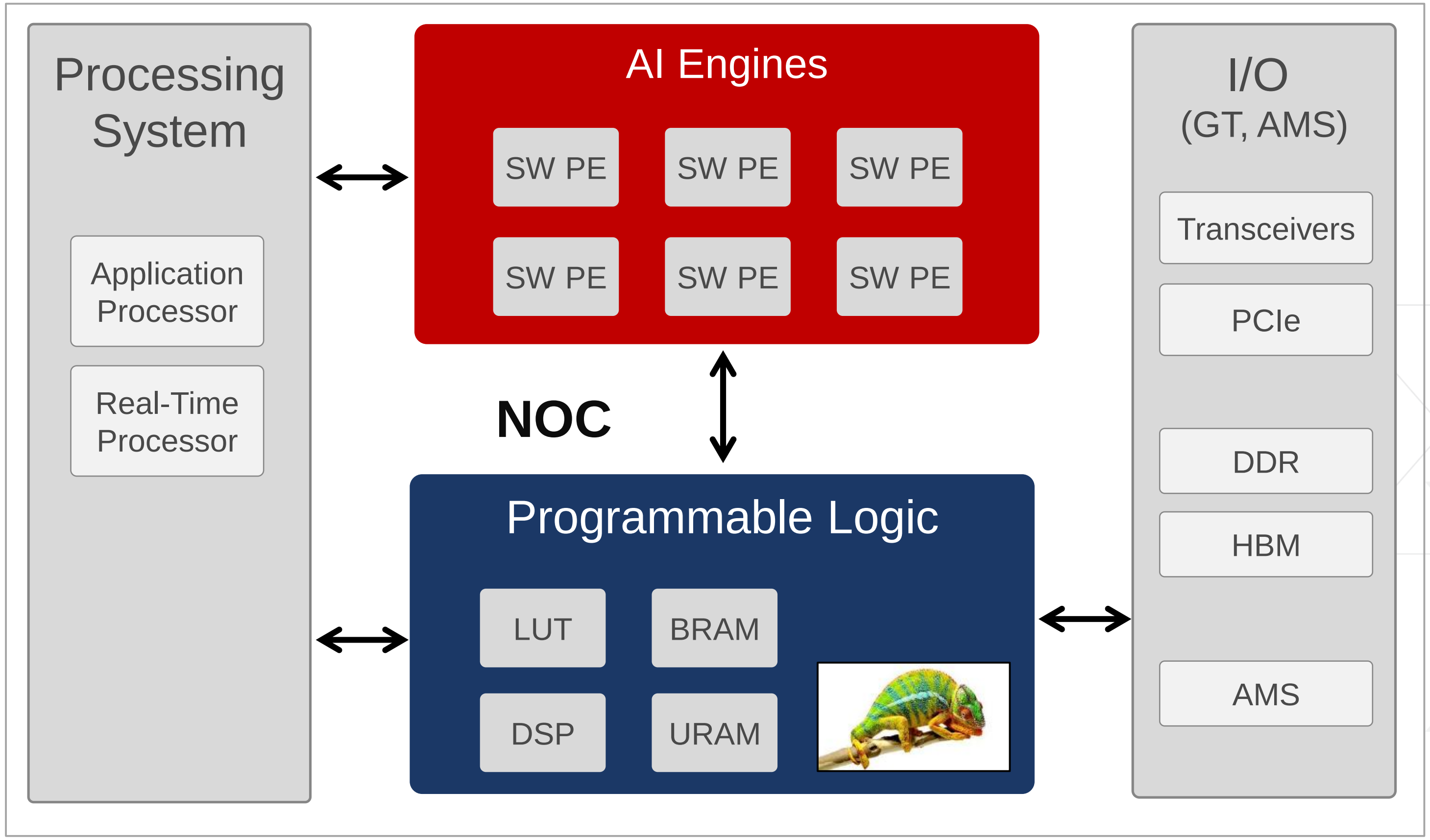


> Diversification of increasingly heterogenous devices and system

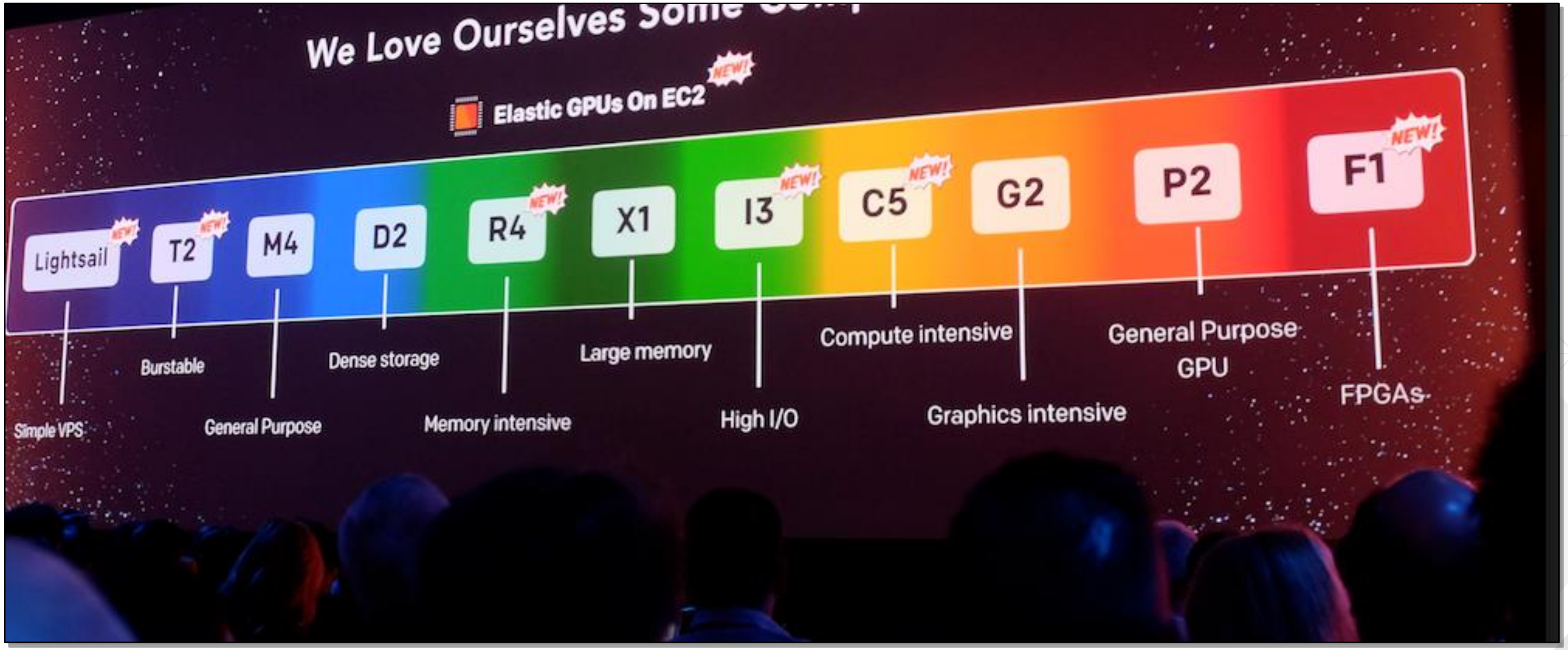
> Moving away from standard van Neumann architectures

> Architectural innovation

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



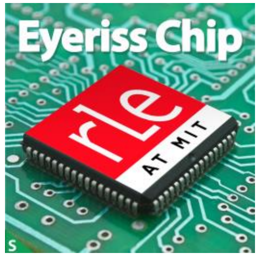
Towards Heterogeneous Cloud: AWS



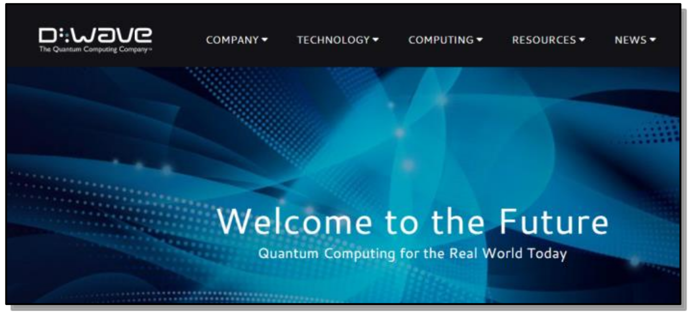
Insight 2016: AWS adding FPGA instances

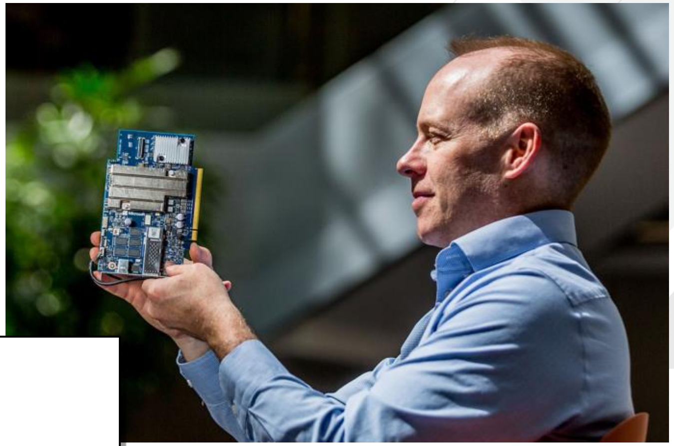
Pretty unconventional: Customized Hardware for AI *DPU: Deep Learning Processing Unit*

> Custom AI Silicon



> Quantum computing



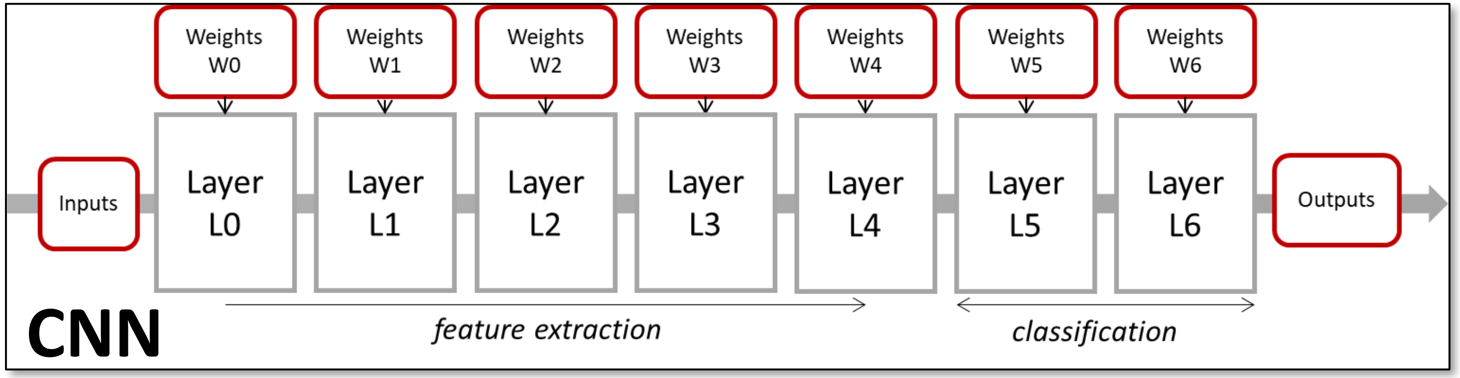


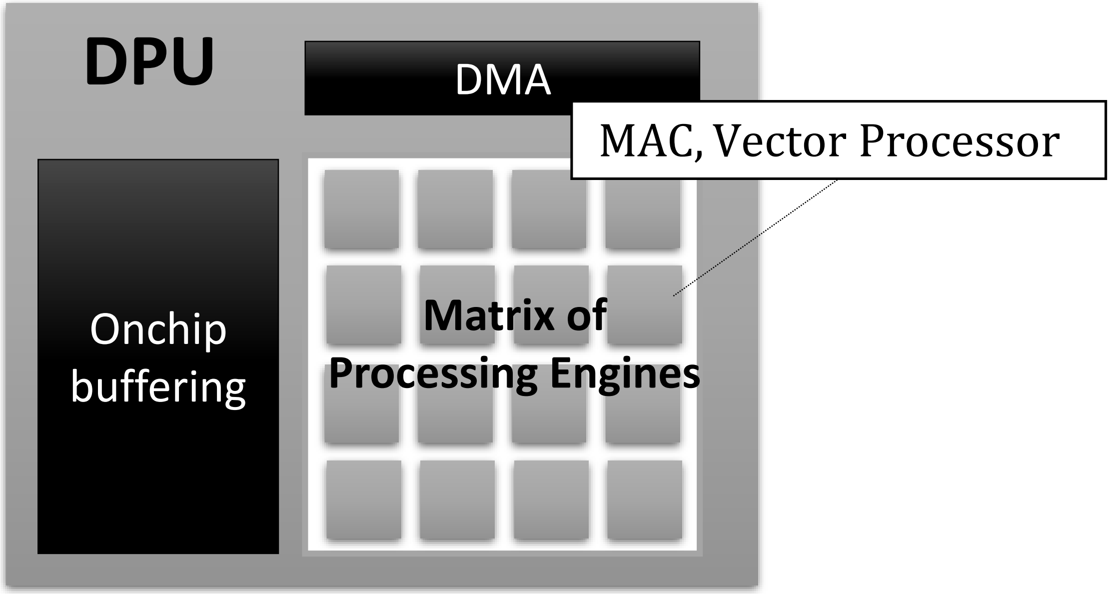
Microsoft Brainwave

> Both soft and hard DPUs

XILINX

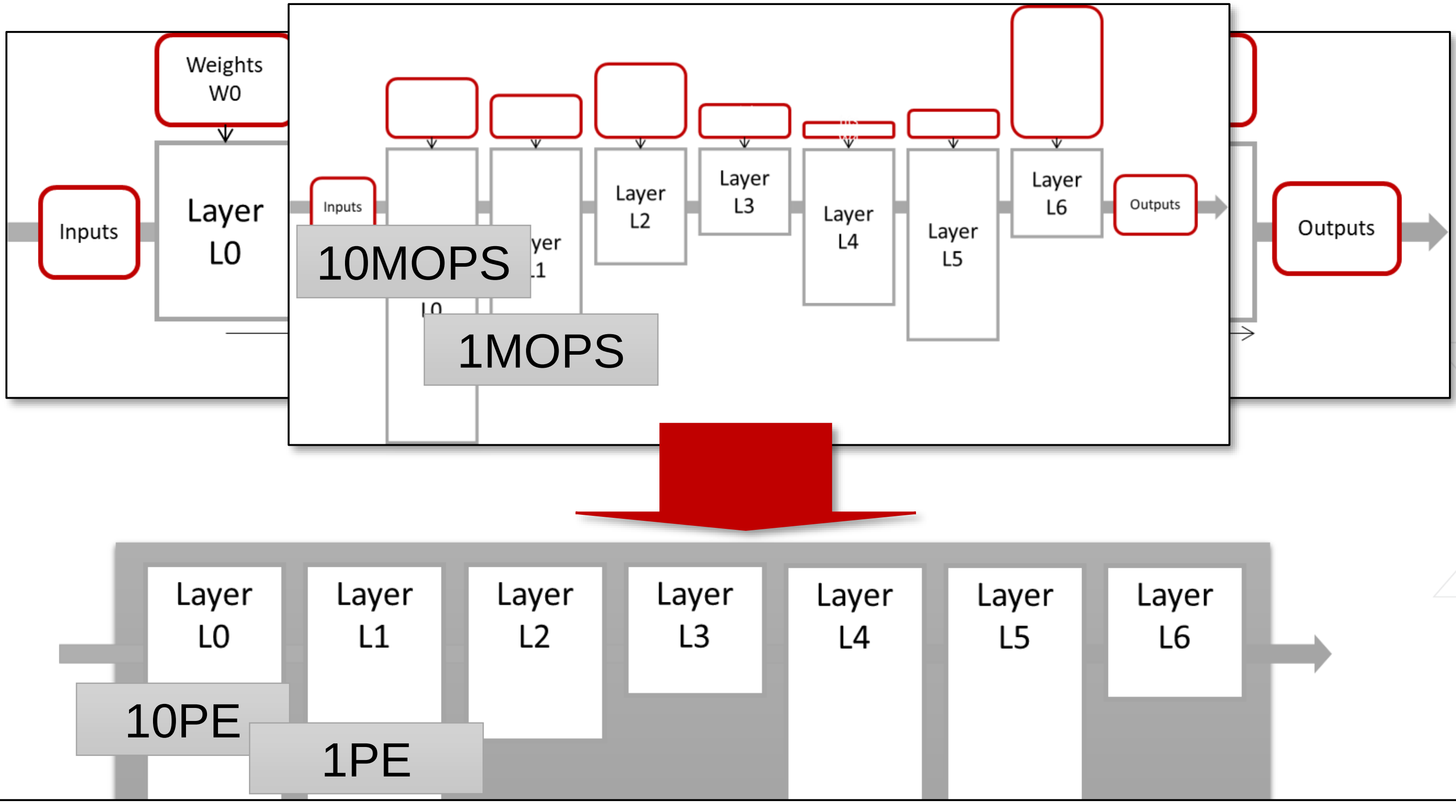
Popular DPU Architecture





"Layer by layer compute"

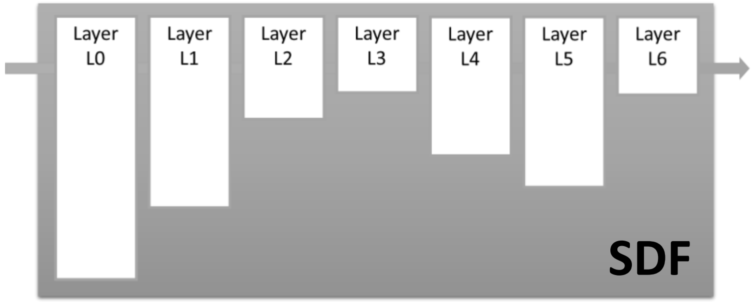
Even more unconventional:
Custom-Tailored Hardware Architectures (Macro-Level)
Synchronous Dataflow

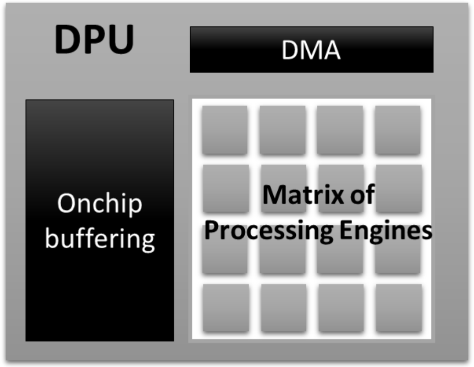


> ***Hardware Architecture Mimics the NN Topology***

> **Customized feed-forward dataflow architecture to match network topology & performance targets**

Synchronous Dataflow (SDF) vs Matrix of Processing Elements (MPE)

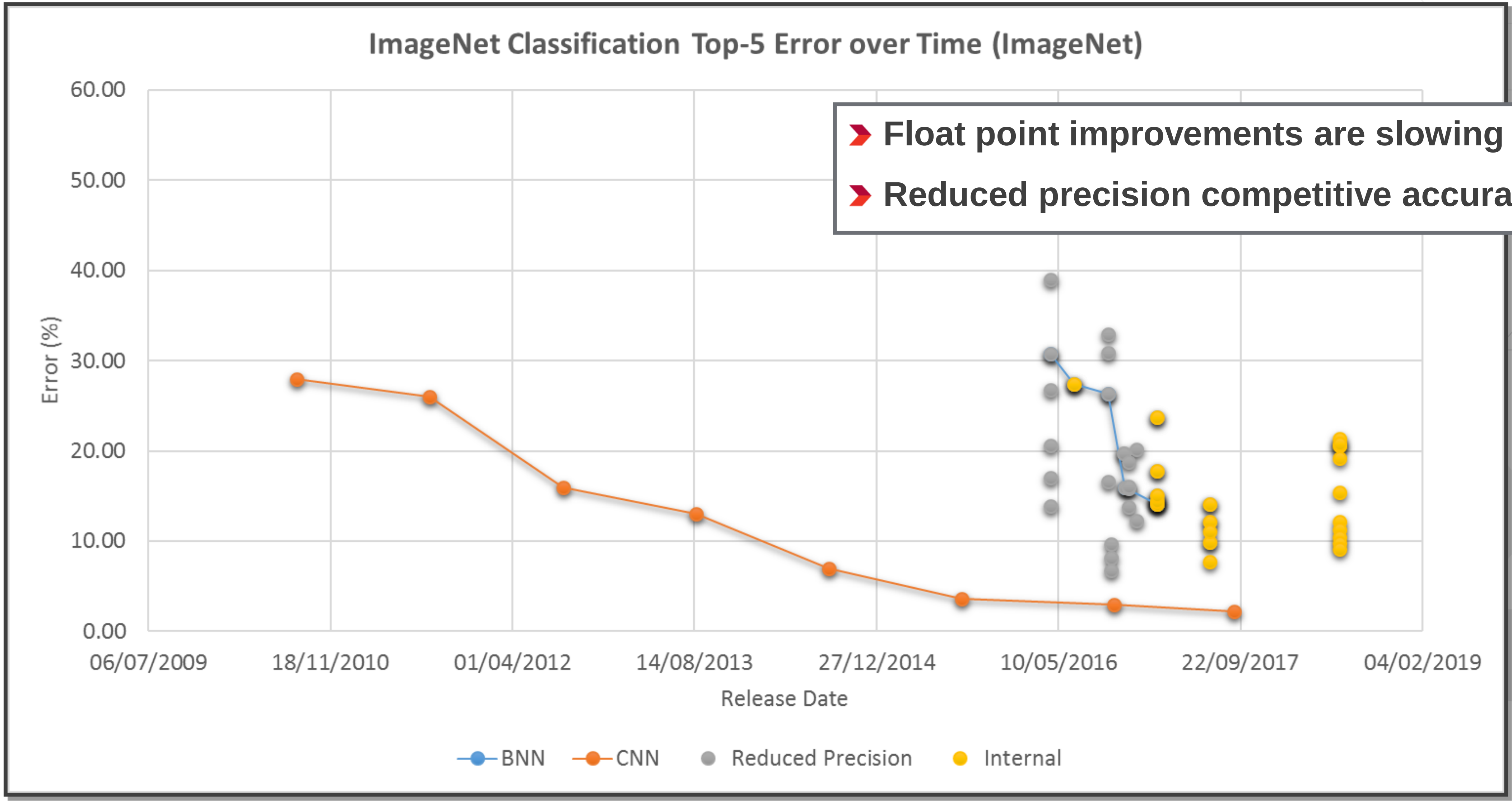




- Higher compute and memory efficiency due to custom-tailored hardware design
- Less flexibility
- No control flow (static schedule)

- Efficiency depends on how well balanced the topology is
- Scales to arbitrary large networks
- Compute efficiency is a scheduling problem

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



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

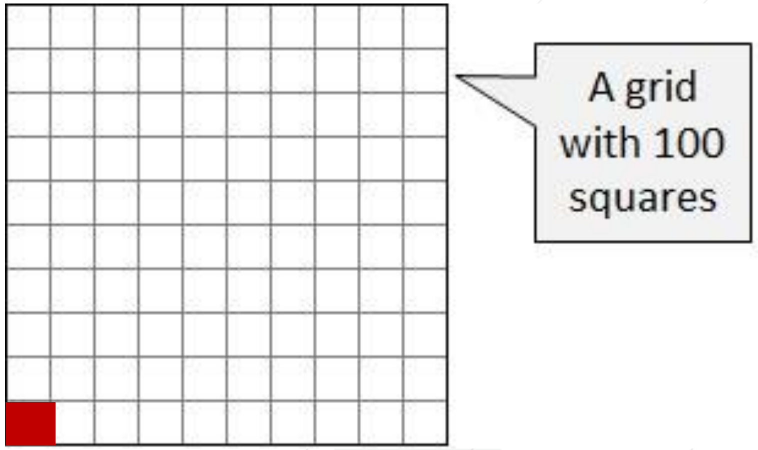
XILINX

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

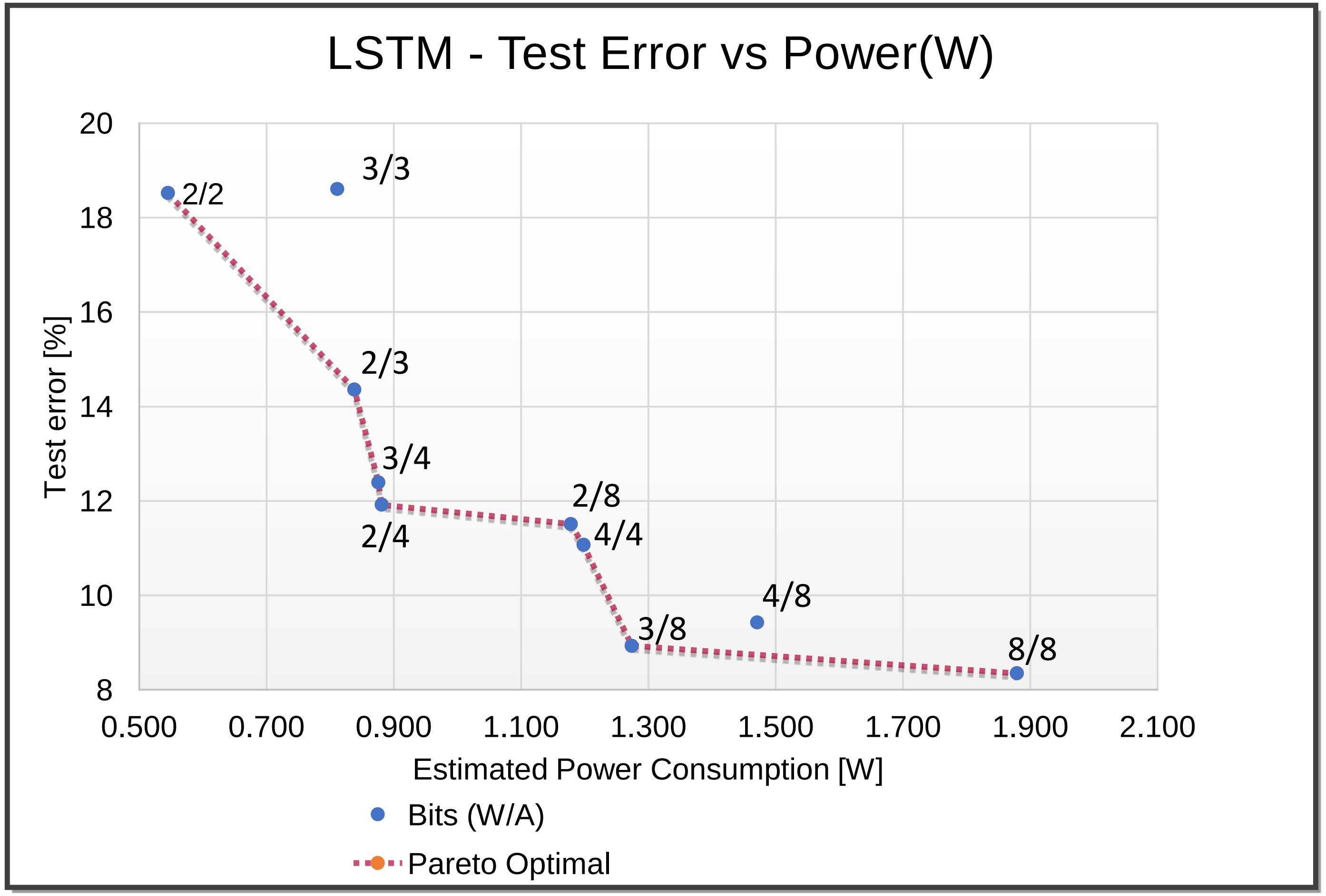
Precision	Modelsize [MB] (ResNet50)
1b	3.2
8b	25.5
32b	102.5



XILINX

Reducing Precision Inherently Saves Power

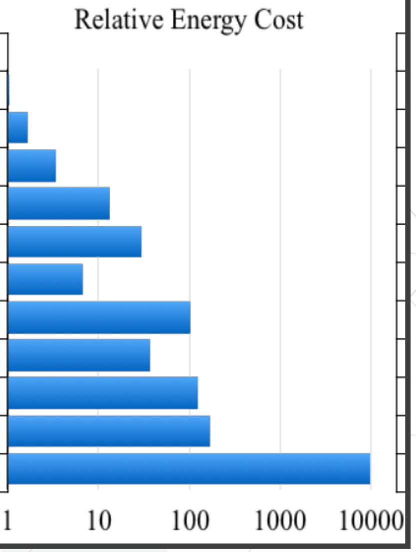
FPGA:



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:

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

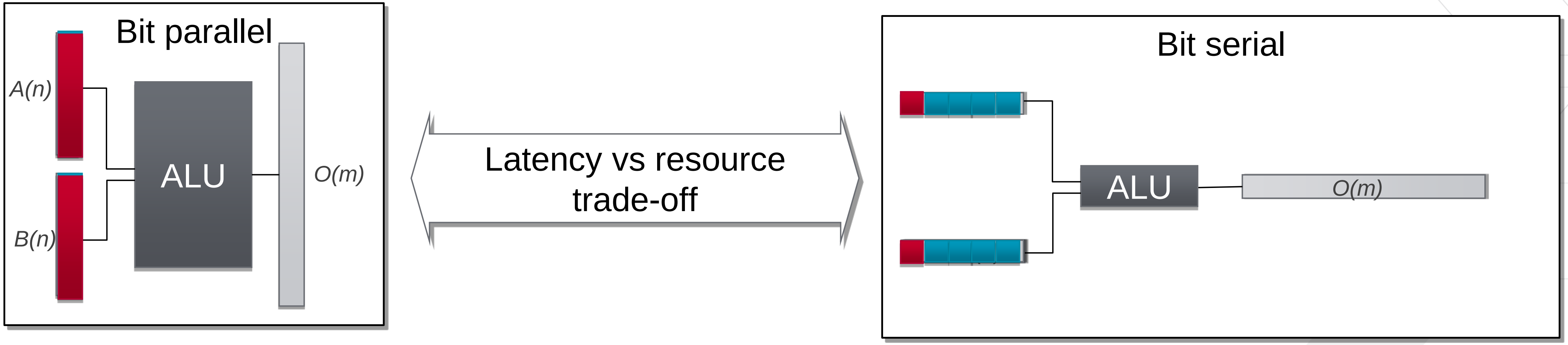
TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

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

Taking unconventional one step further still:
Bit-Parallel vs Bit-Serial

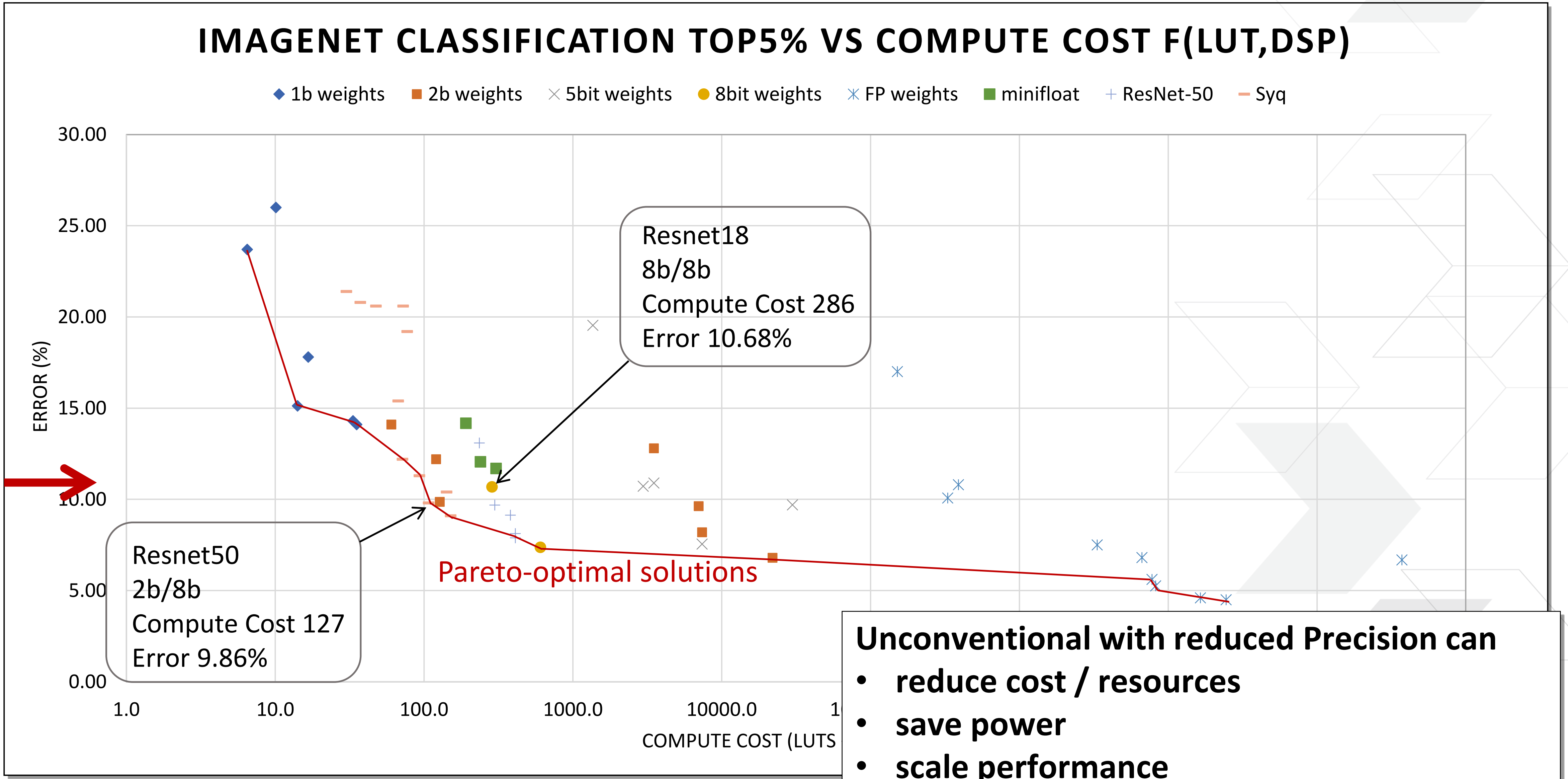
> **Parallelize across the bit precision**



> **FPGA: provides equivalent bit-level performance at chip-level for low precision* + flexible for run-time programmable precision**

Umuroglu, Rasnayake, Sjalander"BISMO: A Scalable Bit-Serial Matrix Multiplication Overlay for Reconfigurable Computing." FPL'2018
https://arxiv.org/pdf/1806.08862.pdf

Design Space Trade-Offs



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
- Leveraging customized dataflow architectures and precisions, these provides dramatic performance scaling and energy efficiency benefits
- Providing new exciting trade-offs within the design space

THANK YOU!

Adaptable.
Intelligent.

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

XILINX