Performance Scaling with Innovative Computing Architectures and FPGAs

Michaela Blott
Distinguished Engineer, Xilinx Research

XILINX

Agenda

- Background
- Motivation
- **Innovative Computing Architectures**

Background

XILINX

Background

> **Xilinx**
>> Fabless semiconductor company
>> Founded in Silicon Valley in 1984
>> Today:
– 4,200 employees
– 20k customers
– $3B revenue

> **Invented the FPGA**
>> From 128LUTs to millions LUTs

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

LUT (lookup table):

X(2..0)	y
000	0
001	1
010	0
011	1
100	0
101	0
110	0
111	1

Sea of LUTs, FFs
+
Programmable interconnect
+ IO

XILINX

What are FPGAs?

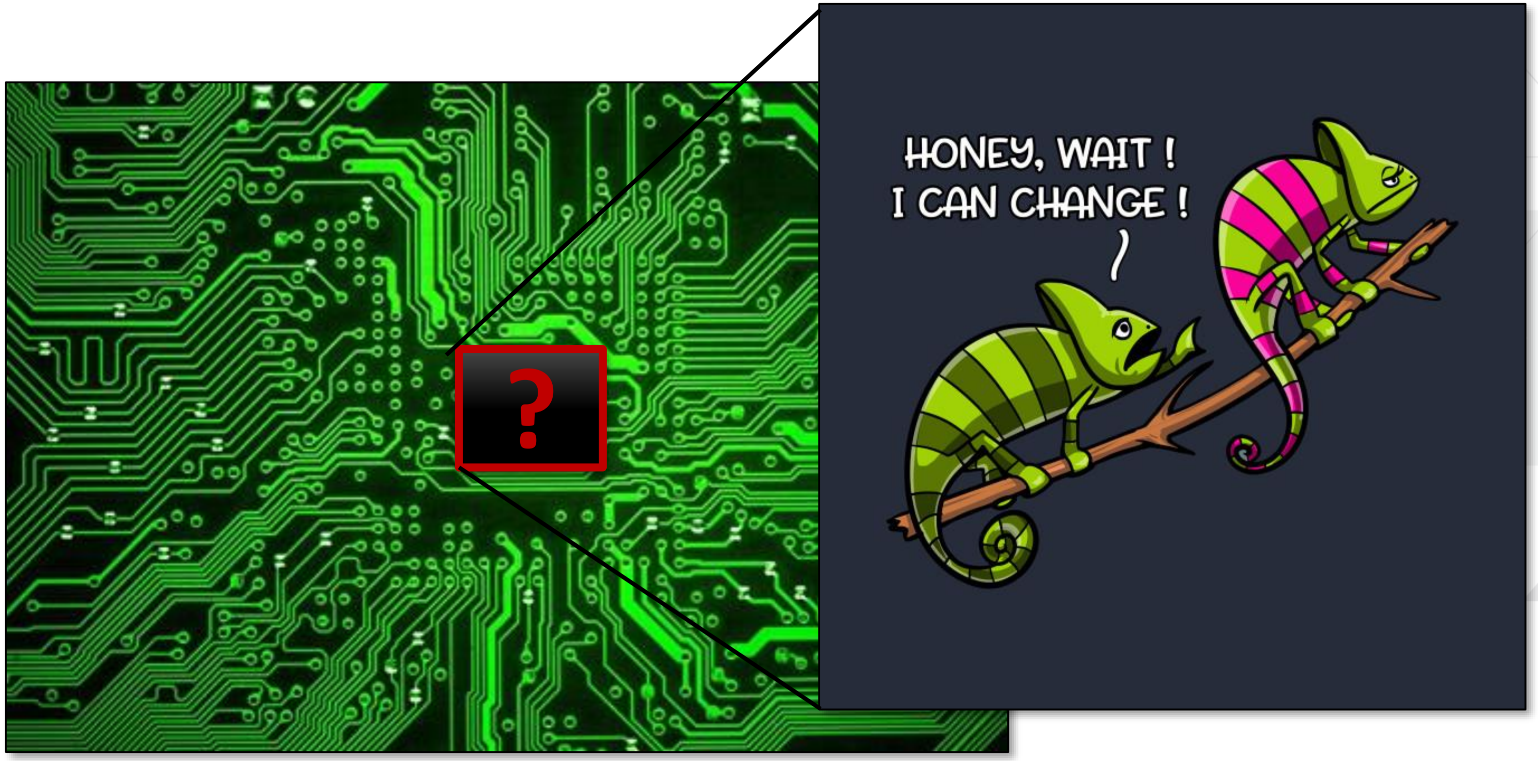
Customizable, Programmable Hardware Architectures

- **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

Xilinx Research - Ireland

- **Established over 13 years ago**
- **Slowly expanding**
- **Increasingly leveraging external funding (IDA, H2020)**

IDA Ireland

XILINX

Current Xlabs Dublin Team

Plus 2 in University Program
(Cathal McCabe, Katy Hurley)

Lucian Petrica, Giulio Gambardella, Alessandro Pappalardo,
Ken O’Brien, me, Nick Fraser, Yaman Umuroglu, Peter Ogden, Giuseppe Natale (from left to right)

Current Focus

- **Quantifying value proposition for FPGAs in Machine Learning**
  - Architectural exploration
  - Algorithmic optimizations
  - Benchmarking

- **In collaboration with**
  - Universities
  - Startups
  - Customers

Plus a Very Active Internship Program

- On average 4-6 interns at any given time
  - From top universities all over the world

- **Overall**
  - 70+ interns since 2007
  - Many collaborations have come from this
  - Many found employment

KIT Karlsruher Institut für Technologie

Technische Universität Kaiserslautern

NTNU Norwegian University of Science and Technology

Politecnico Milano 1863

University of Cambridge

ETH Zürich

Technische Universität Dresden

The University of Sydney

Imperial College London

Motivation

Trends meet Technological Reality

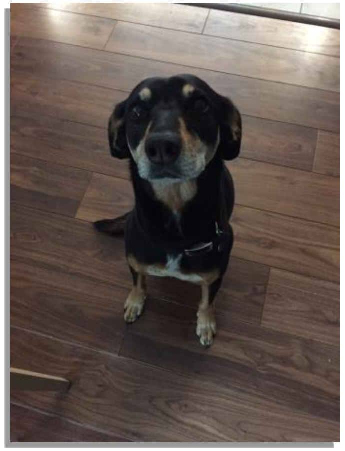
Mega-Trend:
The Rise of the Machine (Learning Algorithm)

1.	1800	2.	1900	3.	2000	4.
Mechanical Steam powered mechanical production		**Electrical** Mass production with electrical energy		**Digital** Automated production		**Virtual** **Machine Learning** Industrial Internet

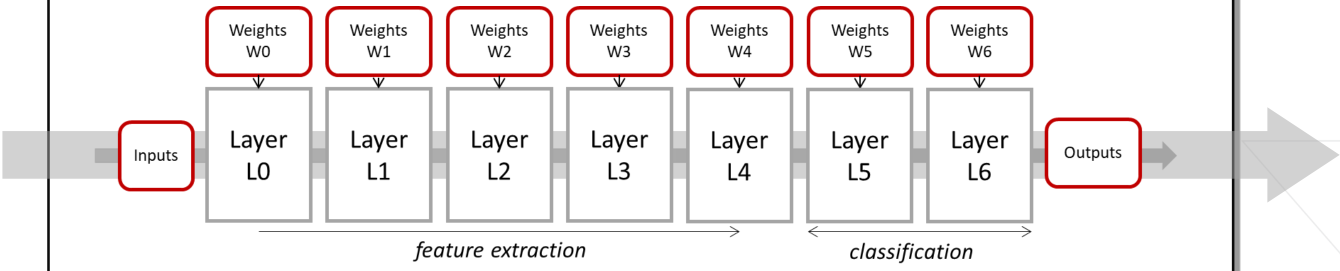
- **Potential to solve the unsolved problems**
  - Reverse engineering the brain (Jeff Dean, Google Brain 2017)
- **Requires little or no domain expertise**
- **NNs are a "universal approximation function"**
- **If you make it big enough and train it enough, can outperform humans on specific tasks**

Mega-Trend:
Enormous Compute and Memory Requirements

Input Image



Neural Network



Neural Network

Cat?

For ResNet50:

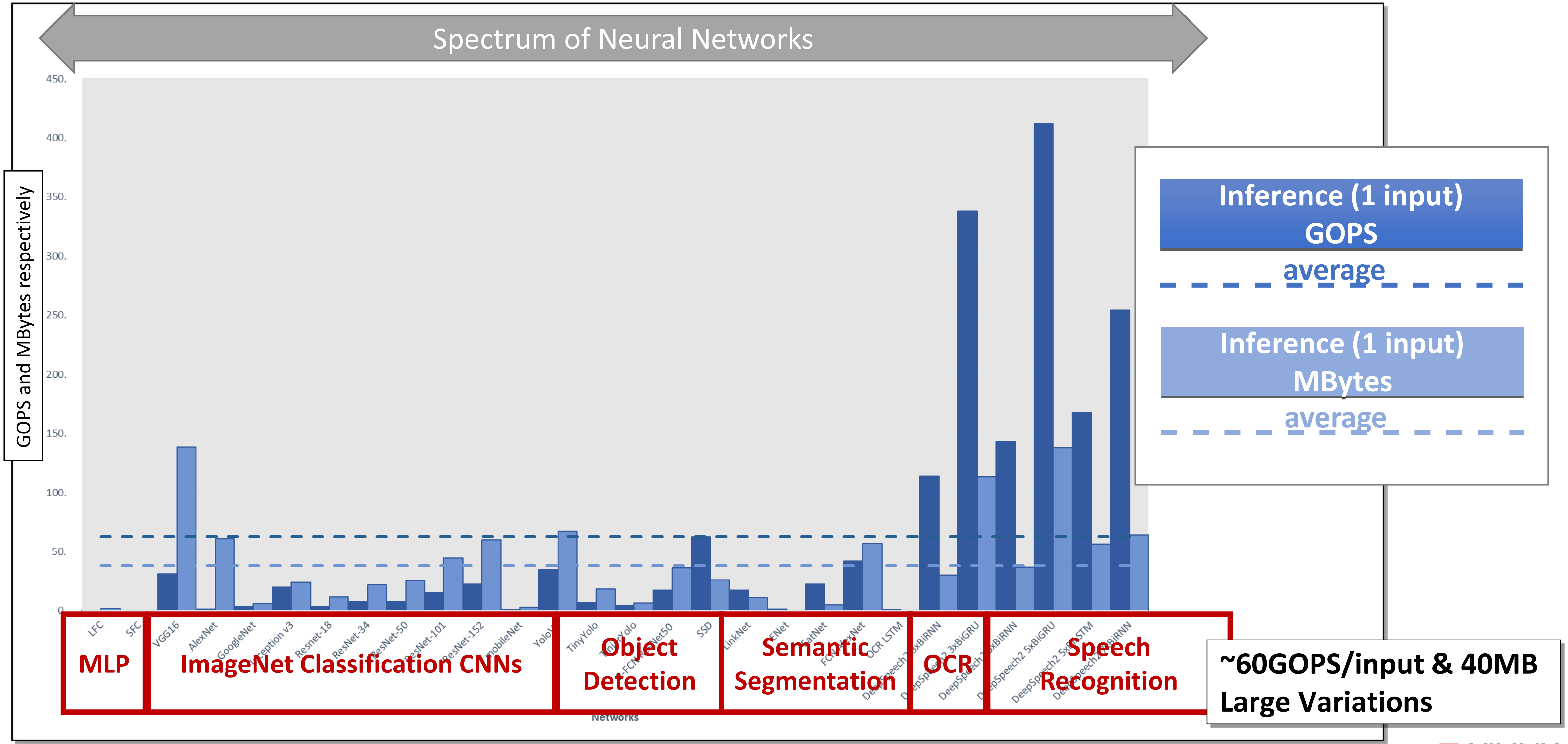
7.7 Billion operations

25.5 millions of weight

Basic arithmetic, incredible parallel but huge compute and memory requirements

Mega-Trend:
Compute and Memory for Inference

*architecture independent
**1 image forward
*** batch = 1
**** int8

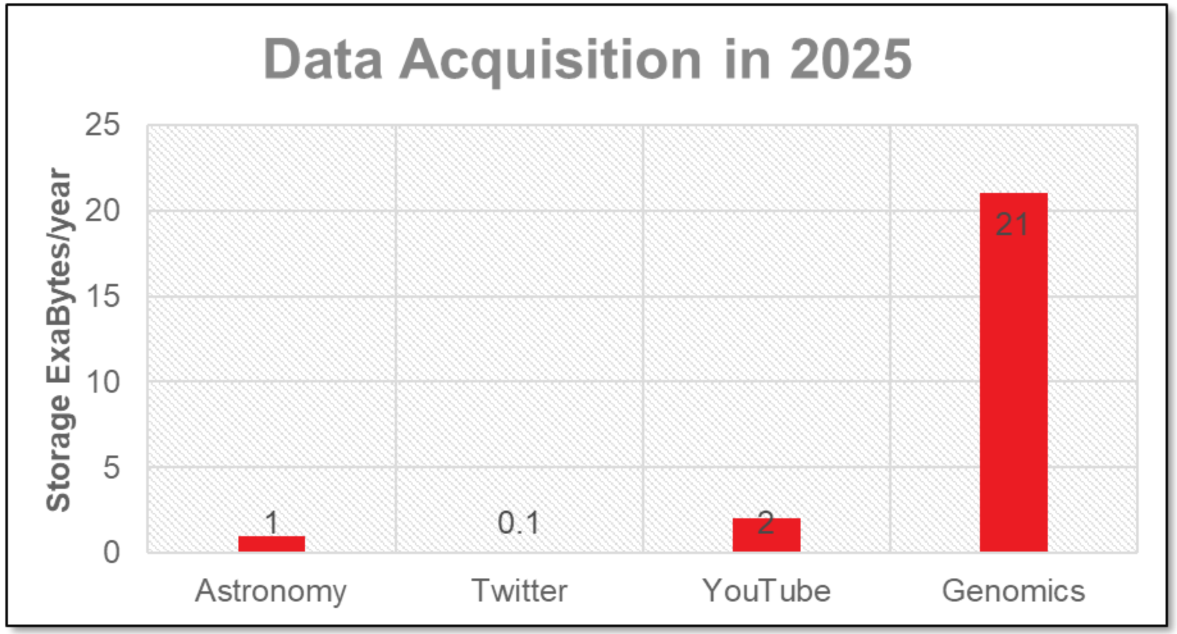


~60GOPS/input & 40MB
Large Variations

XILINX

Mega-Trend:
Explosion of Data

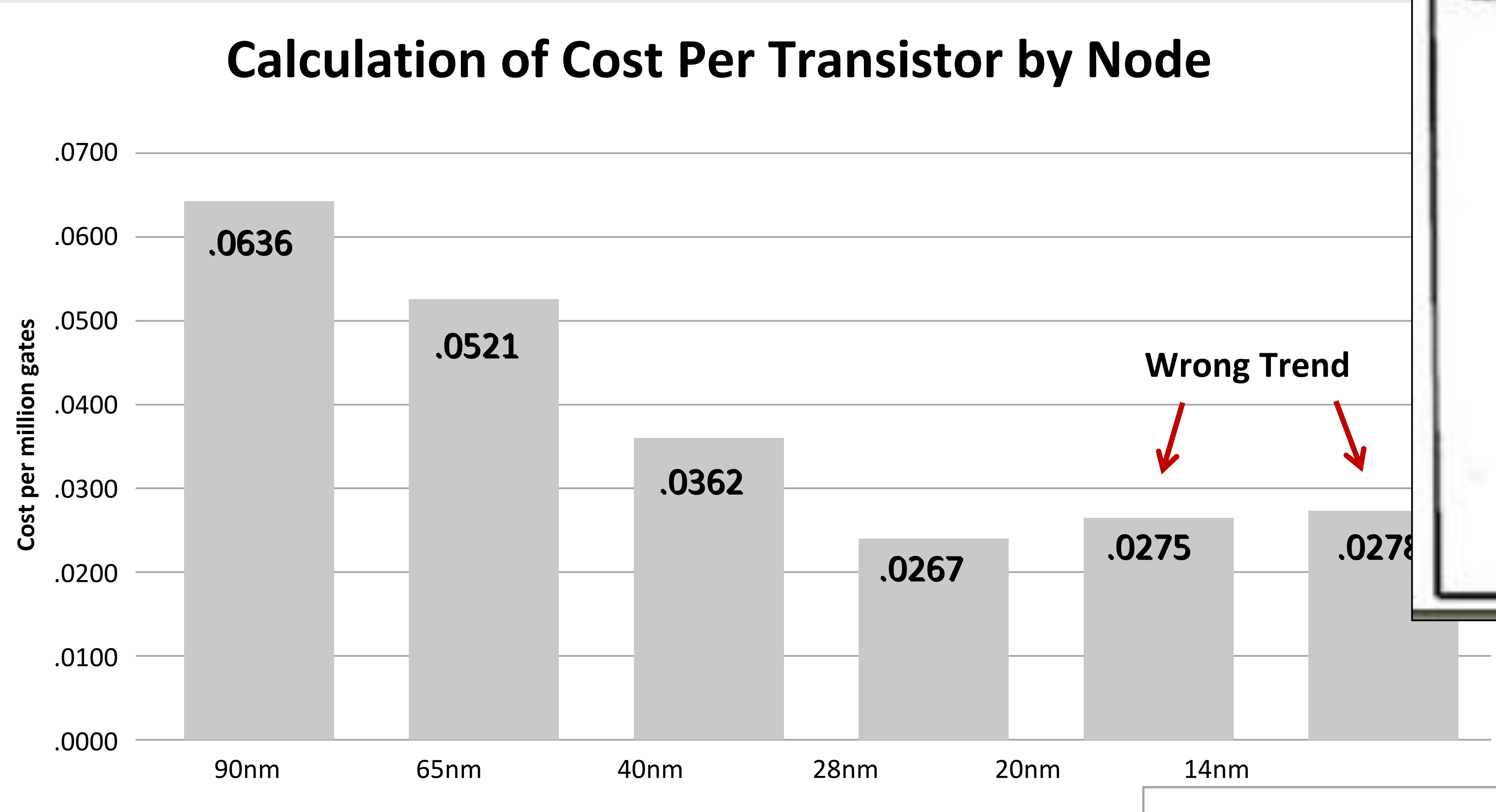
> **#Sensors, #users, videos, DNA!**



Stephens, Zachary D., et al. "Big data: astronomical or genomical?." PLoS biology 13.7 (2015): e1002195.

XILINX

Technology:
End of Moore's Law

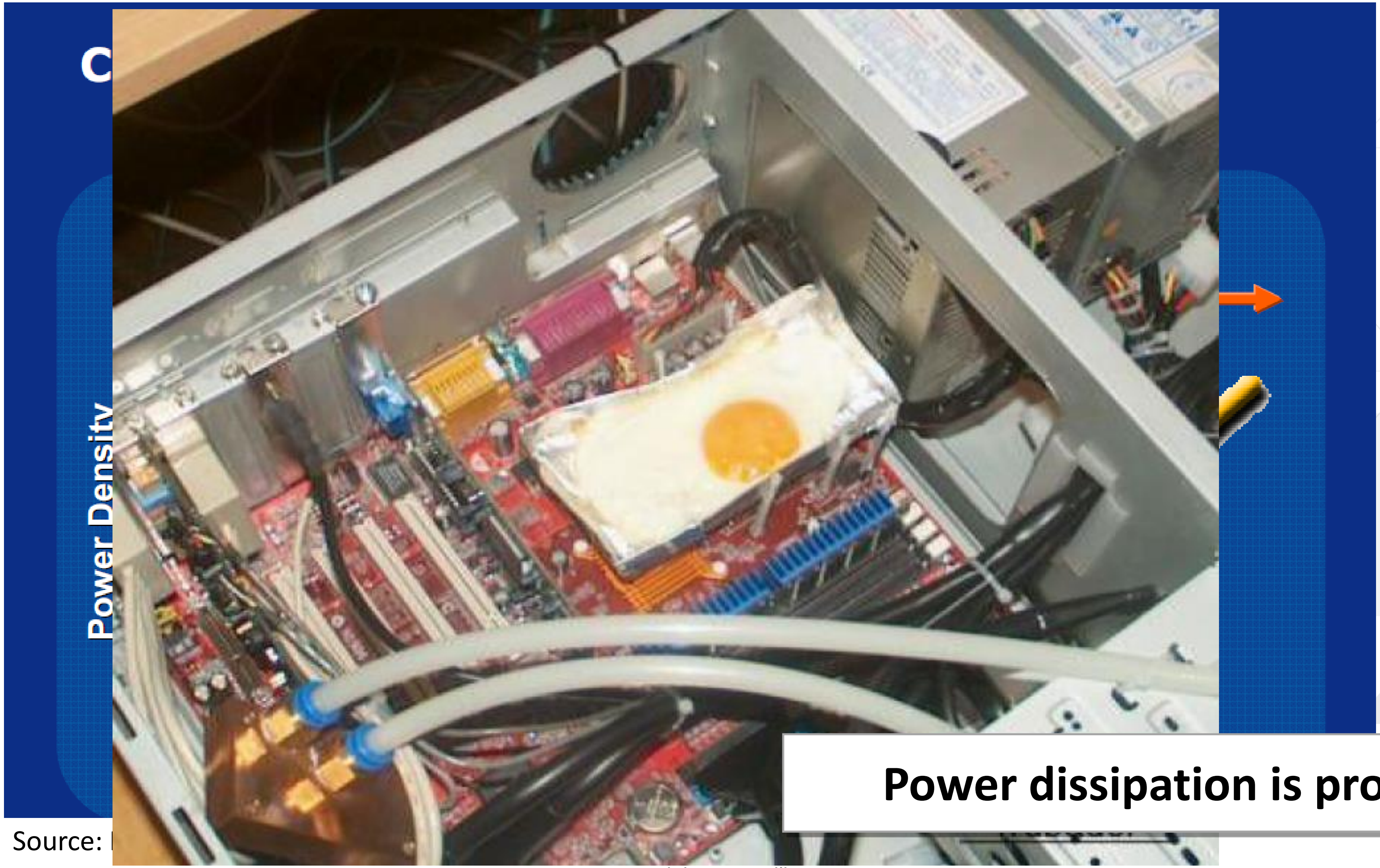


Source: IBS

Economics become questionable

XILINX

Technology:
End of Dennard Scaling



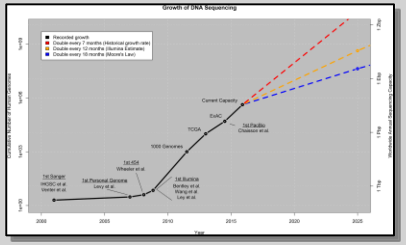
Source: I

Power dissipation is problematic

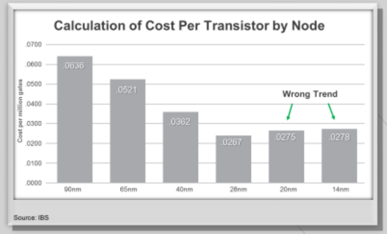
XILINX.

Era of Heterogeneous Compute Using Accelerators

Trends



Technology

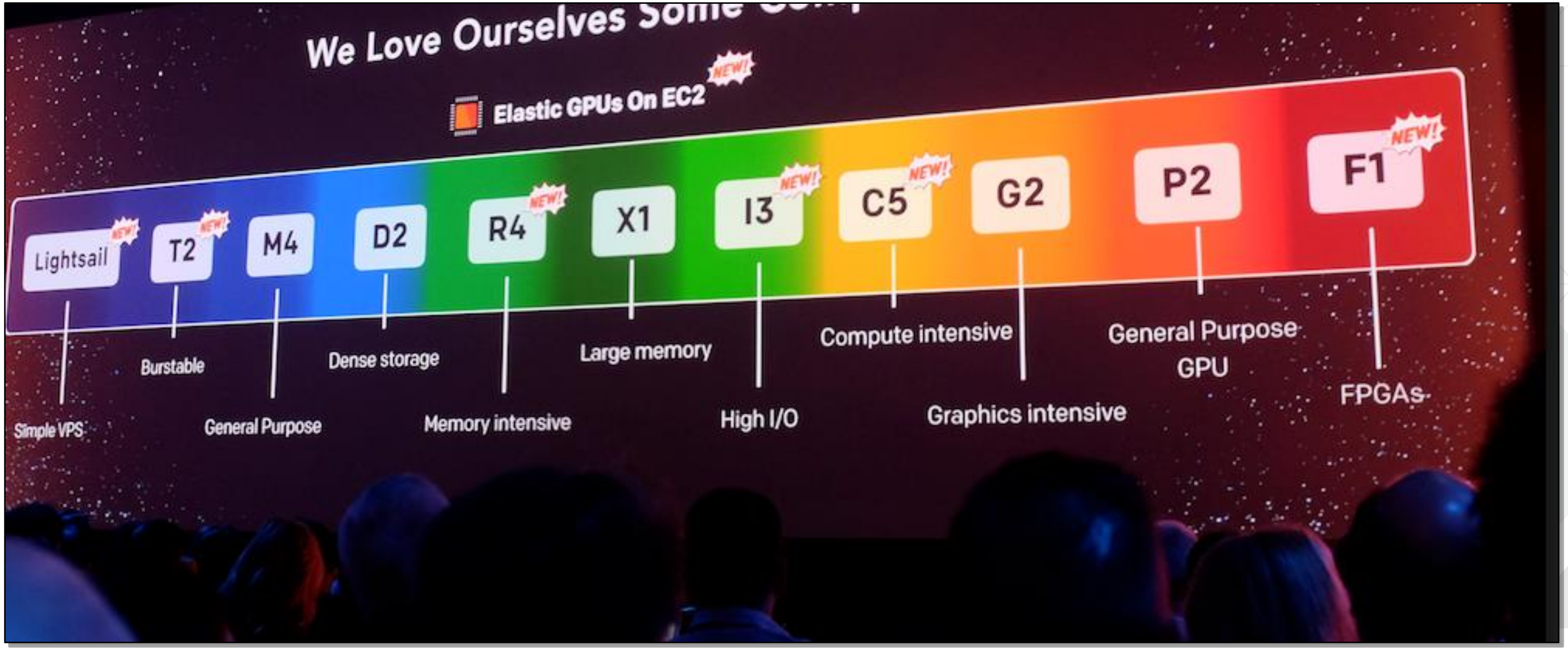


- > **Moving away from standard van Neumann architectures**
- > **Diversification of increasingly heterogenous devices and system**
- > **Algorithmic & architectural innovation is paramount**

XILINX

Diversification of Increasingly Heterogenous Devices and Systems

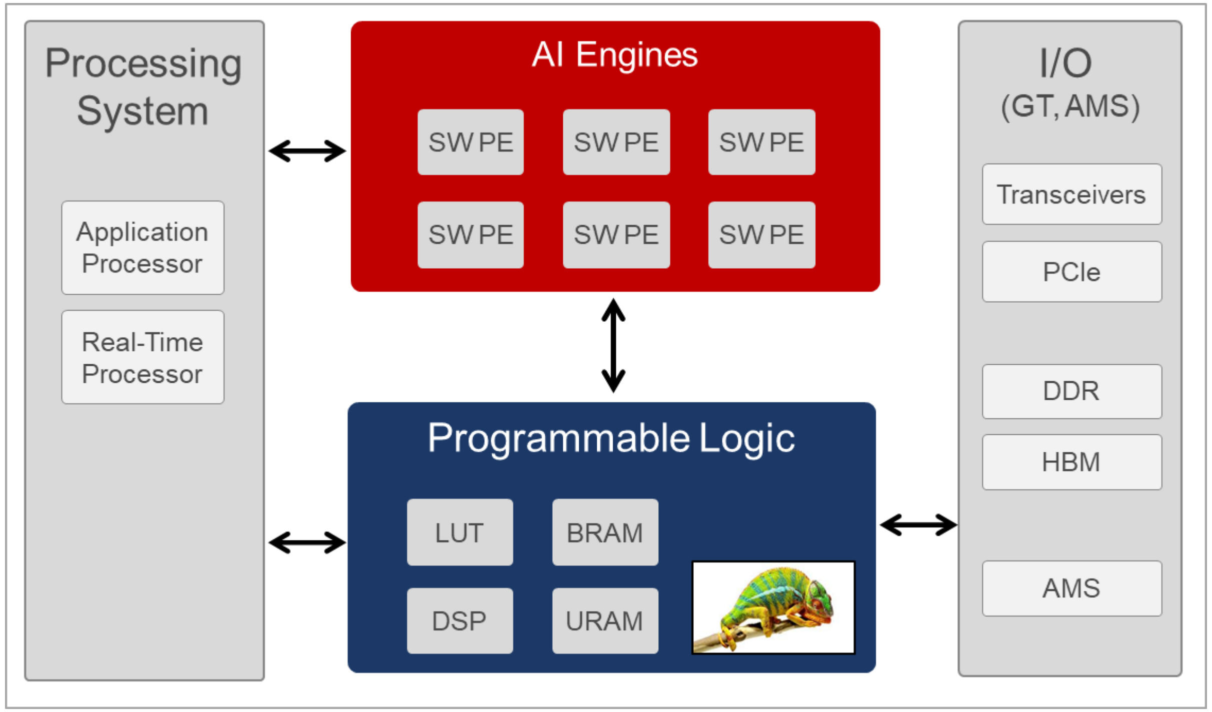
> **Example: AWS Heterogeneous Cloud**

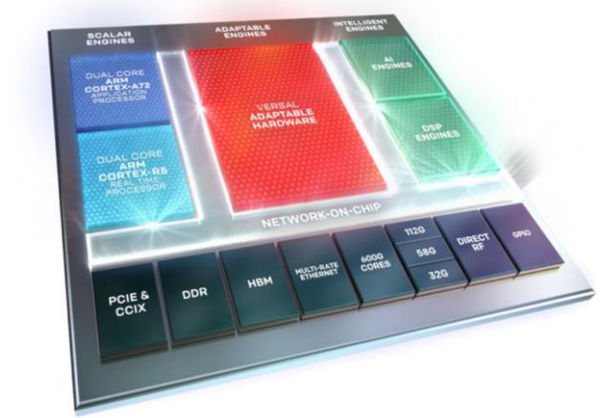


Insight 2016: AWS adding FPGA instances

Diversification of Increasingly Heterogenous Devices and Systems

> Example from the Xilinx World: Evolution of FPGAs to **ACAPs**





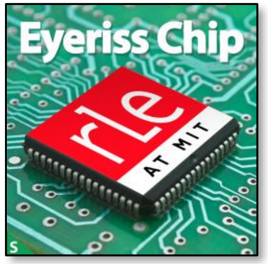
Nowhere it is as extreme as here…
Example: Customized Hardware for AI

KNUPATH

cerebras

nervana

WAVE COMPUTING



TESLA

habana

ASPINITY

DPU: Deep Learning Processing Unit

XILINX

Algorithmic and Architectural Innovation is no longer optional

XILINX

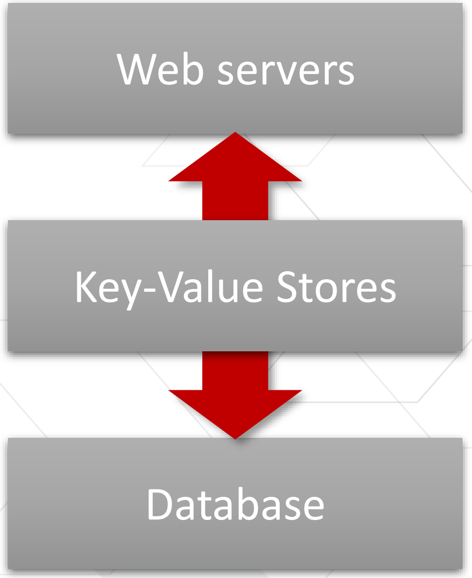
Key-Value Stores

Customized Compute & Memory Subsystem

XILINX

Key Value Stores

- **Many popular websites share a similar basic architecture:**
  - Tier of web servers
  - Disk-based SQL database
  - Caching tier to relief access bottleneck on database
    - Most popular and most recent database contents are cached in main memory of a tier of server platforms



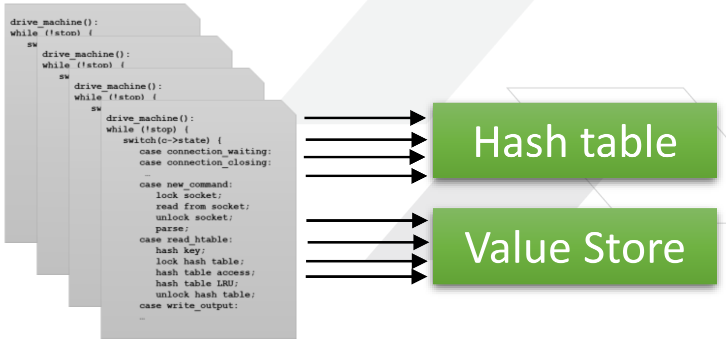
- **Key value stores are content addressable memory (CAM)**
  - Present the query as “key”
  - And the query response is the “value”

Typical Implementations

> **Multithreaded implementation (pthreads)**
   - All threads execute drive_machine(), processes connections from one state to next
   - Receive & parse - hash lookup - value store access - format & transmit
   - Share data structures (hash tables, value store,…)

```
drive_machine():
while (!stop) {
   switch(c->state) {
          …
      case new_command:
         lock socket;
         read from socket;
         unlock socket;
         parse;
      case read_htable:
         hash key;
         lock hash table;
         hash table access;
         hash table LRU;
         unlock hash table;
      case write_output:
      …
```

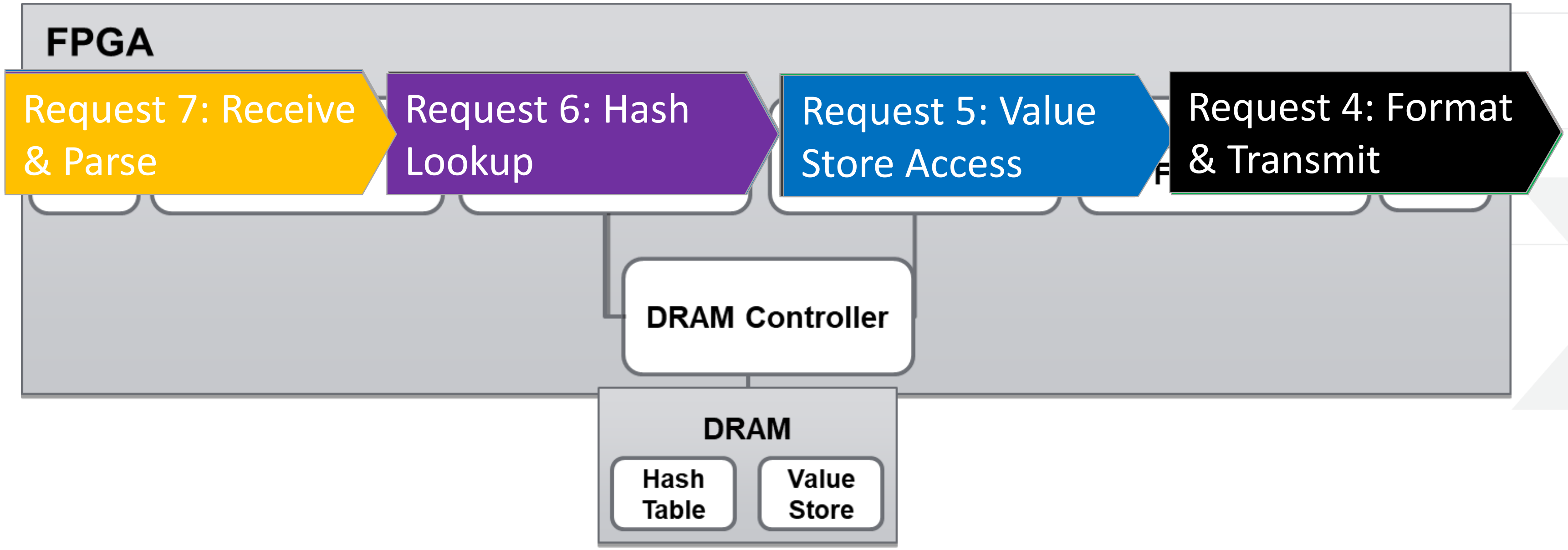
> **Bottlenecked by**
   - Synchronization overhead
     – Threads stall on memory locks, serializing execution for x86s
   - Last level cache ineffective due to random-access nature of the application (miss rate 60% - 95% on x86)



> **Performance significantly below 10Gbps line rate**
     – Intel Xeon (8cores): 1.34MRps, 200-300usec

Dataflow Architectures to Scale Performance & Reduce Latency

- Streaming architecture with flow-controlled series of processing stages which manipulate and pass through packets and their associated state
- Numerous requests are processed in parallel



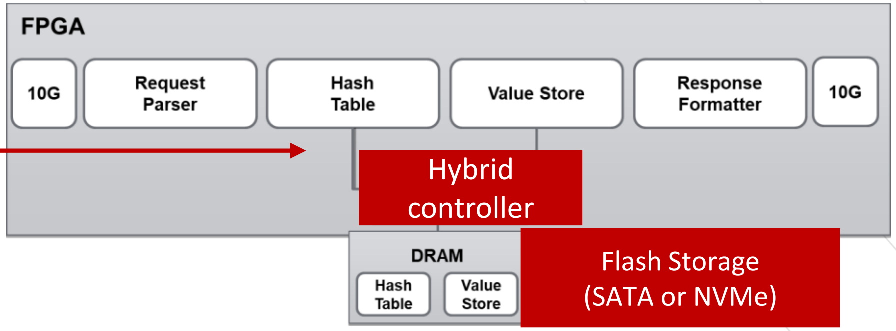
High Compute Performance By Exploiting Task-Level Parallelism

Customized Memory Architecture

> **No wasted caches**

> **Static memory access schedule**
>> Perfect overlap of compute and memory access

> **Leveraging flash to scale capacity with hybrid memory controller**
>> For large values



13MRps demonstrated with a 64b data path @ 156MHz using 3% of FPGA resources with 3-5useconds latency
80Gbps can be achieved by using a 512b @ 156MHz pipeline for example
40TB of value store

Source: Blott et al: Achieving 10Gbps line-rate key-value stores with FPGAs; HotCloud 2013
Blott et al: Scaling towards 80Gbps line-rate 40TB key-value stores with FPGAs; HotStorage 2014

Key-Value Stores

Customized Network Stack & System Architectures

XILINX

Another Bottleneck: TCP/IP

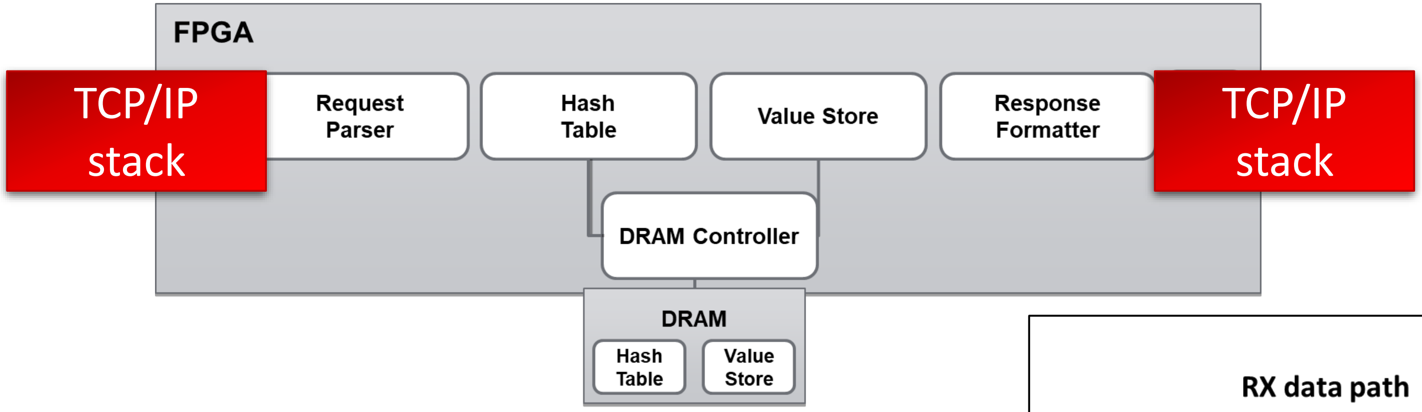
- **CPU intensive**
  - 114% system cycles vs 45% user space out of 800% (8-core Xeon processor)

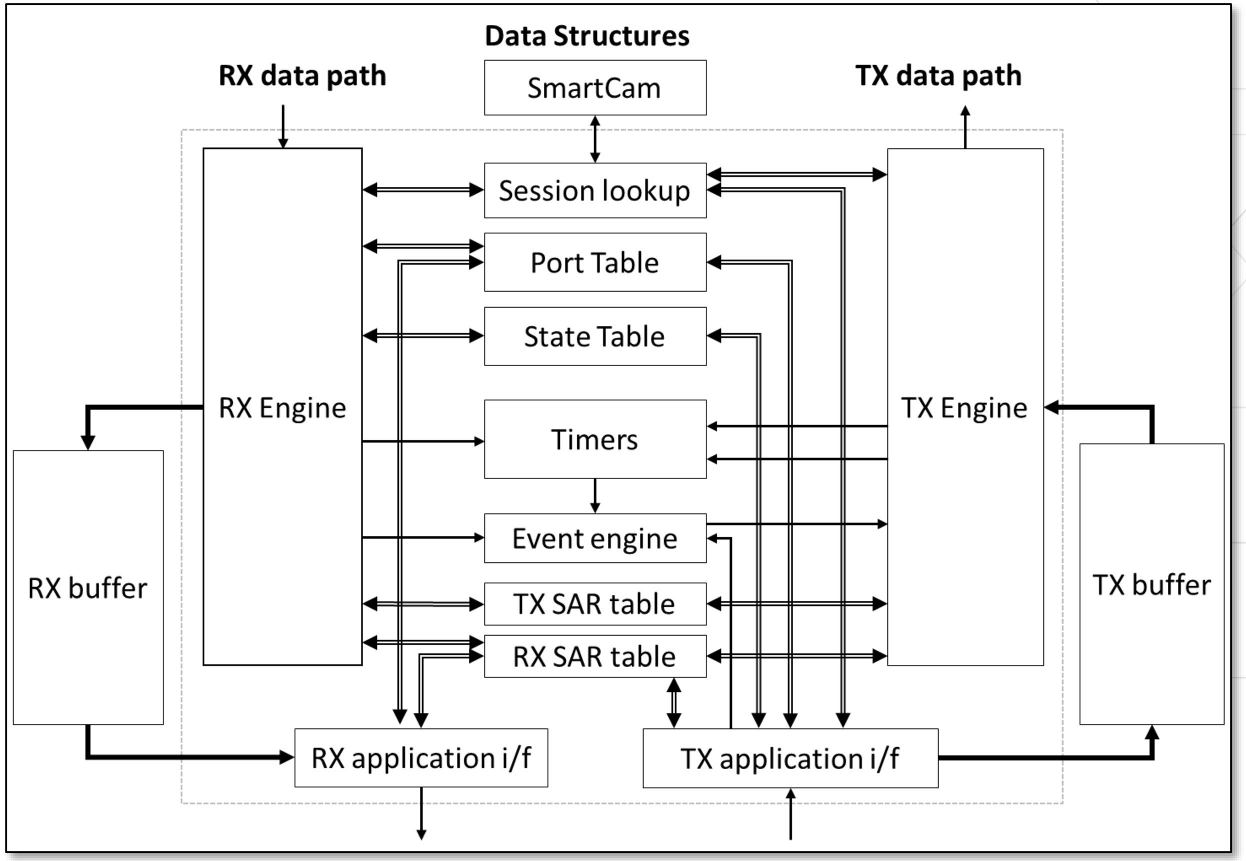
- **Large footprint which leads to high rate of instruction cache misses (up to 160 MPKI)**

- **Frequent interrupts**
  - Causes poor branch predictability (stalling superscalar pipeline) on x86

- **CPI = 2.5**

Network Processing Offloaded





- **In collaboration with ETH Zurich**
- **10G Ethernet line-rate**
- **100G release imminent**
- **Available open source**

- Scalable 10 Gbps TCP/IP Stack Architecture for Reconfigurable Hardware; David Sidler, Gustavo Alonso, Michaela Blott, Kimon Karras, Kees Vissers, and Raymond Carley; FCCM 2015, Vancouver, 2015

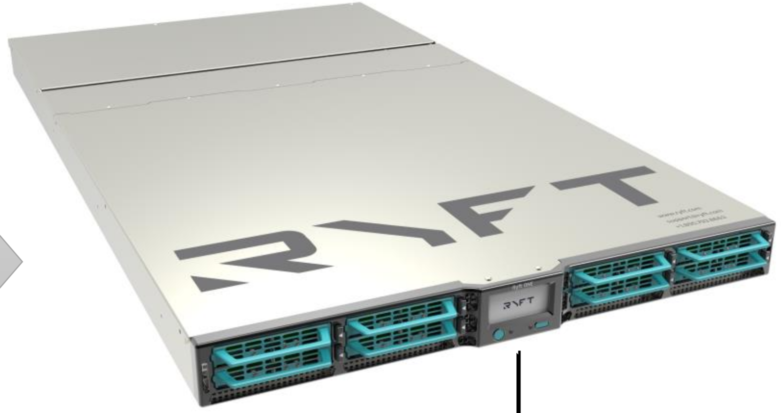
XILINX

Towards Customized System Architectures

Hosted Accelerator

AMD EPYC Server

Customized Appliances (unhosted) “in-network processing”



Web servers

Database

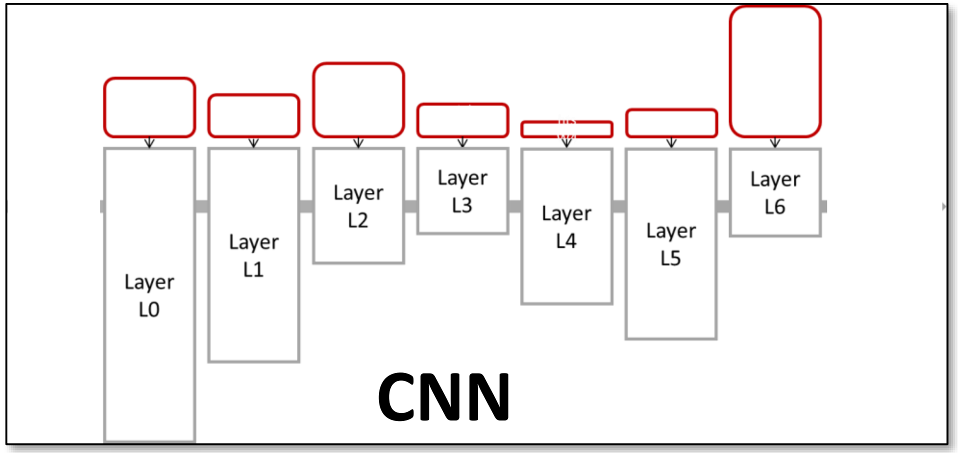
Improves energy efficiency further by removing host system
Reduces required rack space

XILINX

Deep Learning
Customizing Compute

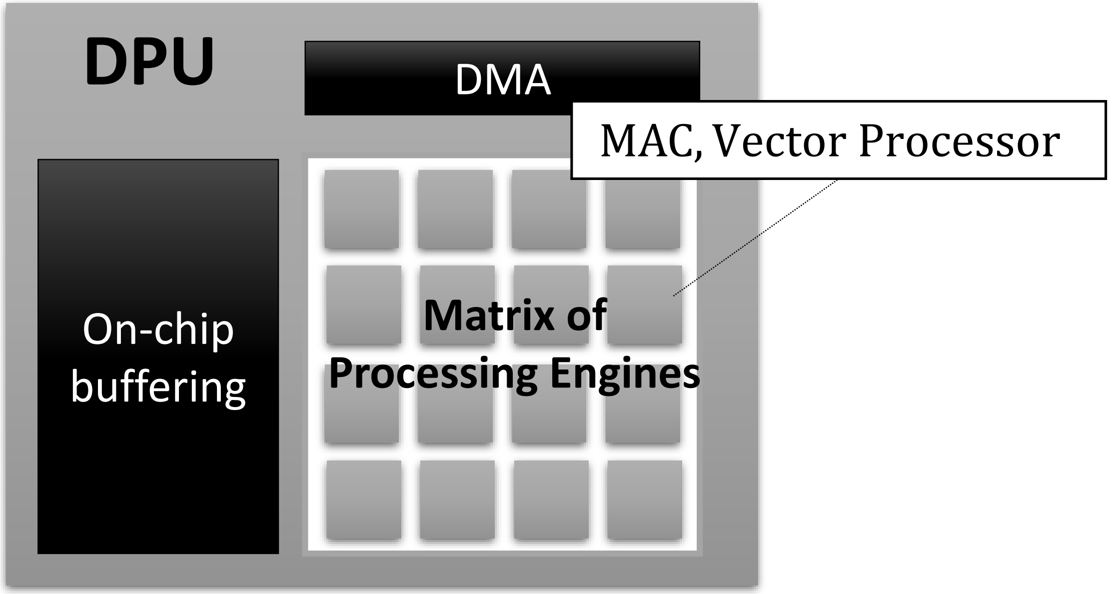
XILINX®

Customized Compute for Machine Learning Workloads



DPU: Deep Learning Processing Unit

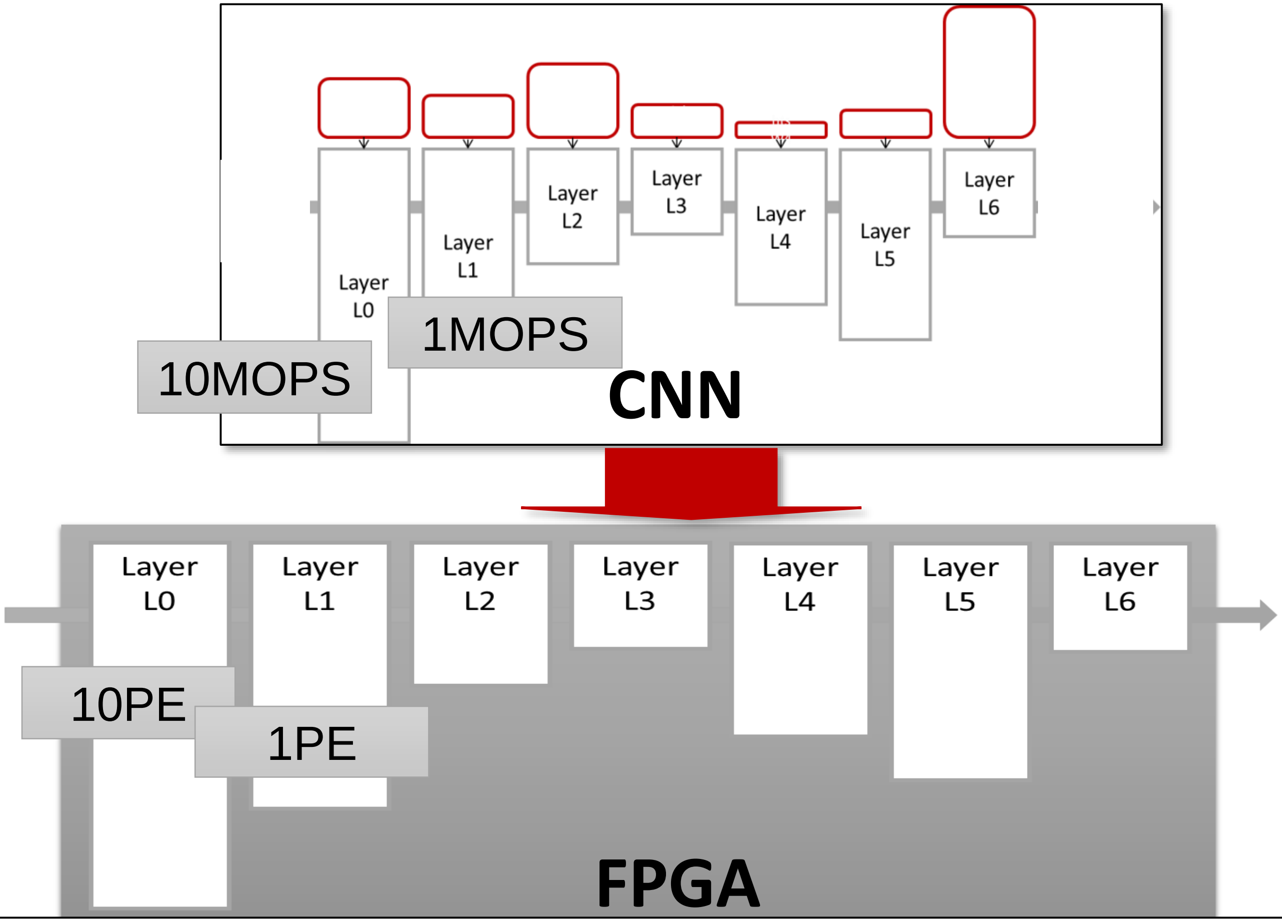
"Layer by layer compute"



Popular DPU Architecture

Custom-Tailored for Specific Topologies

Synchronous Dataflow

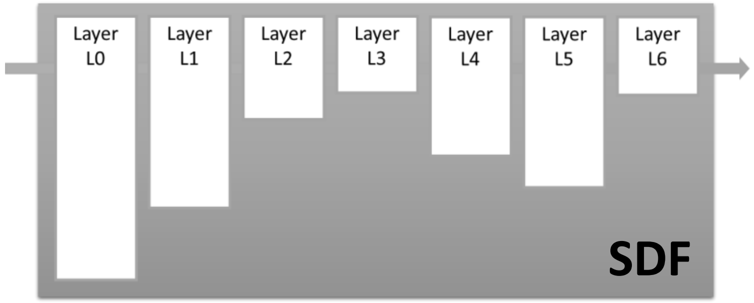


"Hardware Architecture Mimics the NN Topology"

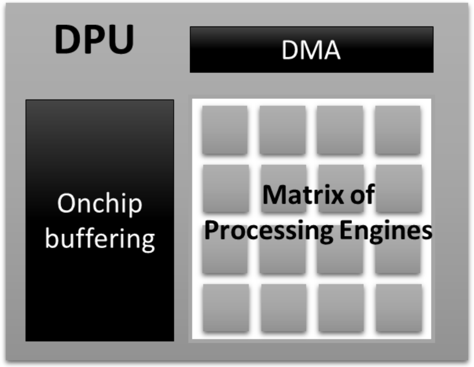
> **Customized feed-forward dataflow architecture to match network topology**

XILINX

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



Spectrum of Options



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

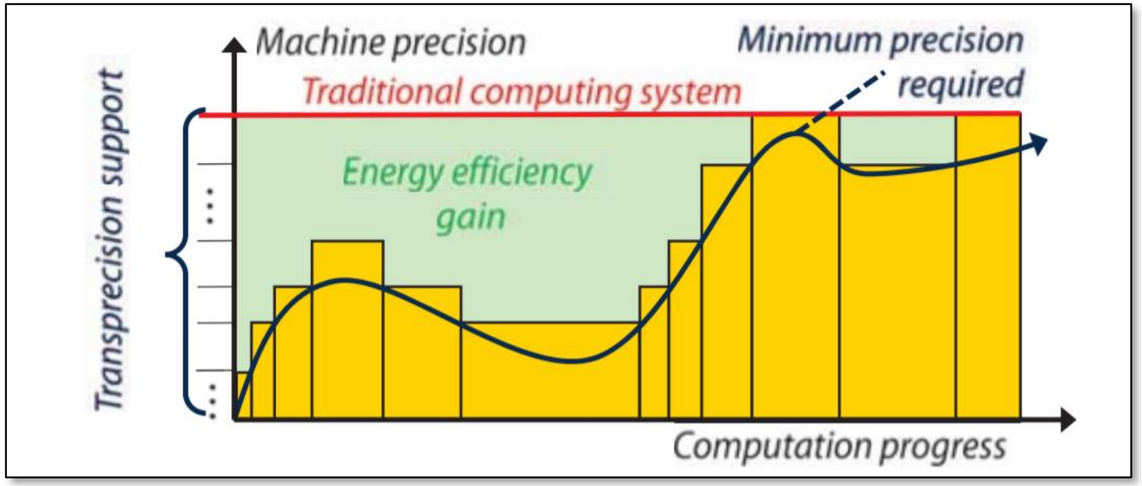
- **Efficiency depends on how well balanced the topology is**
- **Scales to arbitrary large networks**
- **Compute efficiency is a scheduling problem**

Deep Learning
Customizing Arithmetic

XILINX®

Transprecision Computing

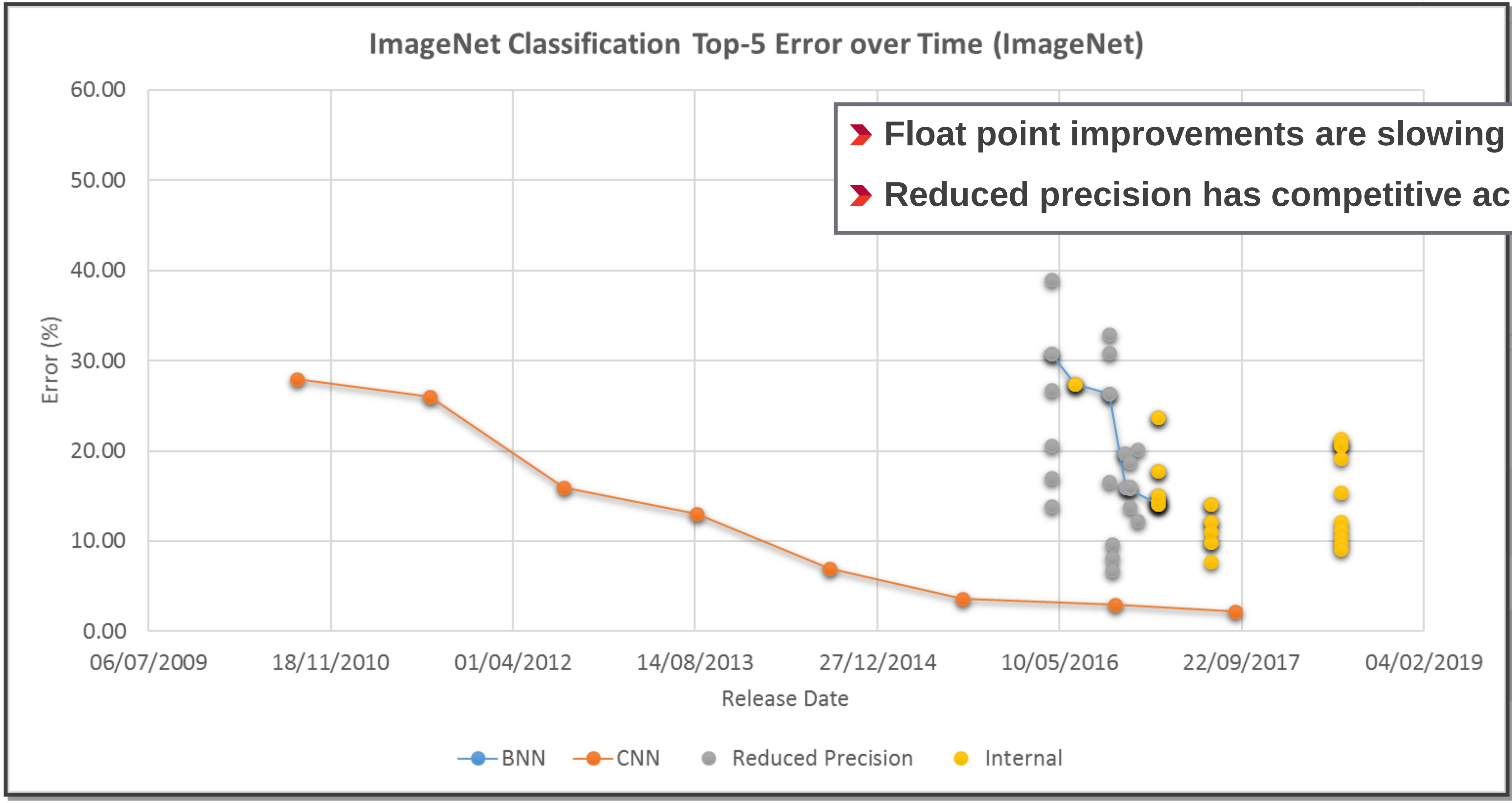
- Application precision requirement can change over time
- Adapt precision to what is required to save energy and/or increase performance
- Numerous applications: PageRank, KNN, stencils, deep learning...



[Malossi et al., The Transprecision Computing Paradigm: Concept, Design, and Applications, DATE’18]

- **Executing everything in the same precision can be wasteful**

Customized Precision
Deliver Competitive Accuracy



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

XILINX

Reducing Precision

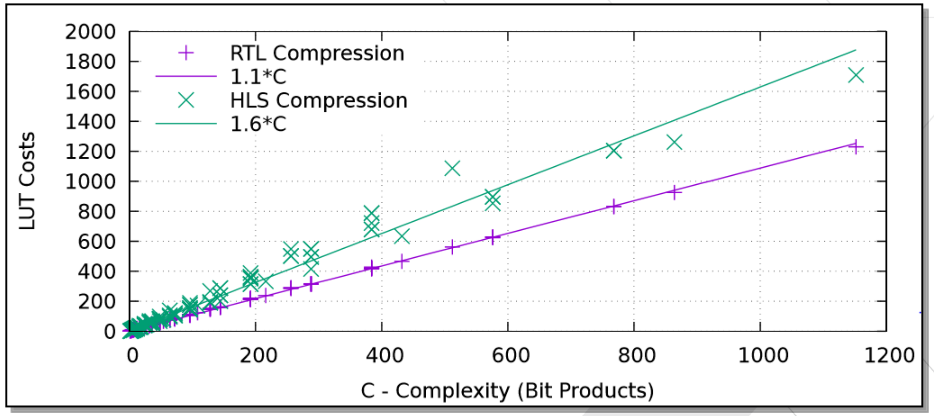
Scales Performance & Reduces Memory

> **Reducing precision shrinks LUT cost**
>
> >> Instantiate **100x** more compute within the same fabric

> **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

Precision	Cost per Op [LUT]	Cost per Op [DSP]
1b	2.5	0
8b	45	0
32b	178	2

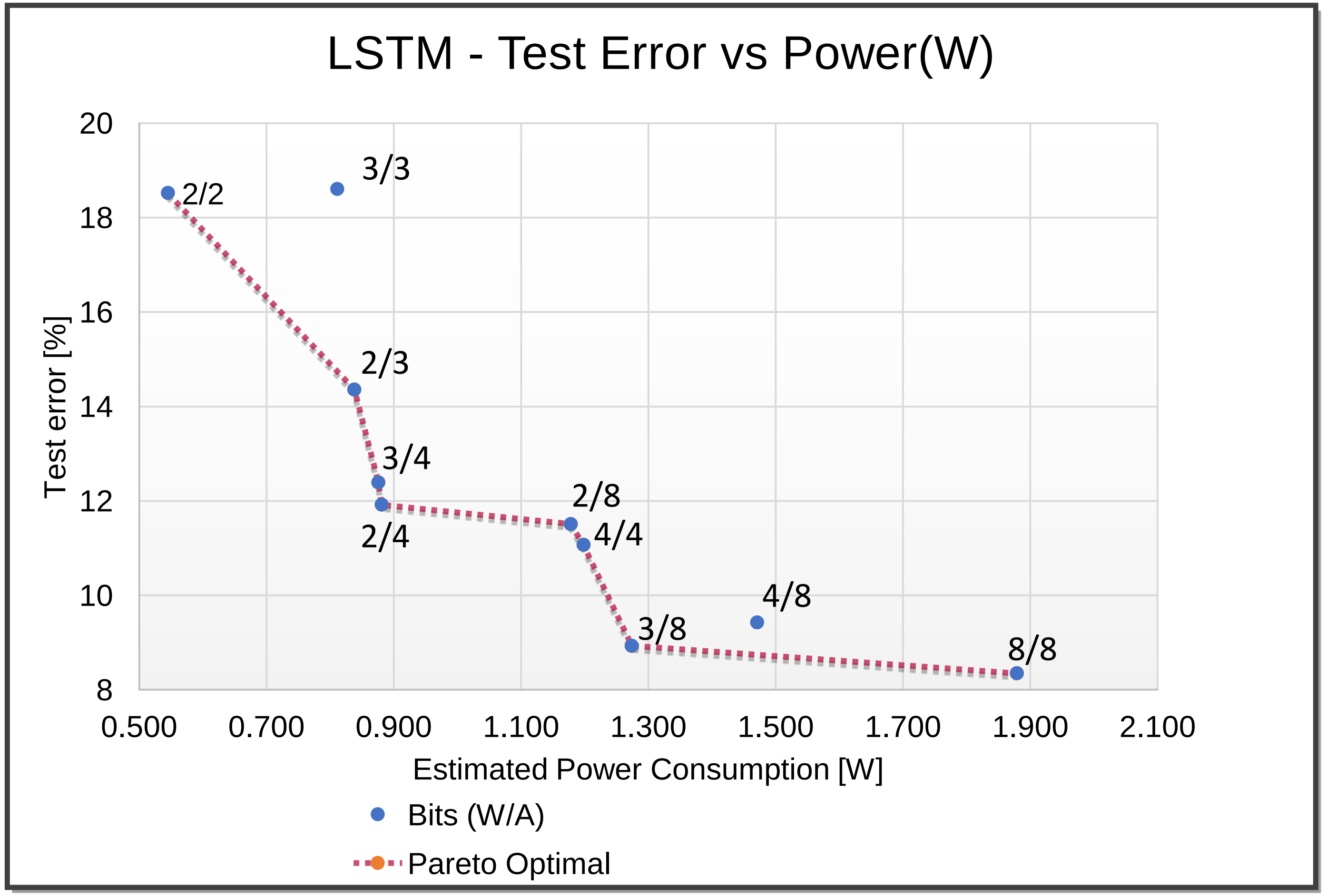


C= size of accumulator * size of weight * size of activation

XILINX

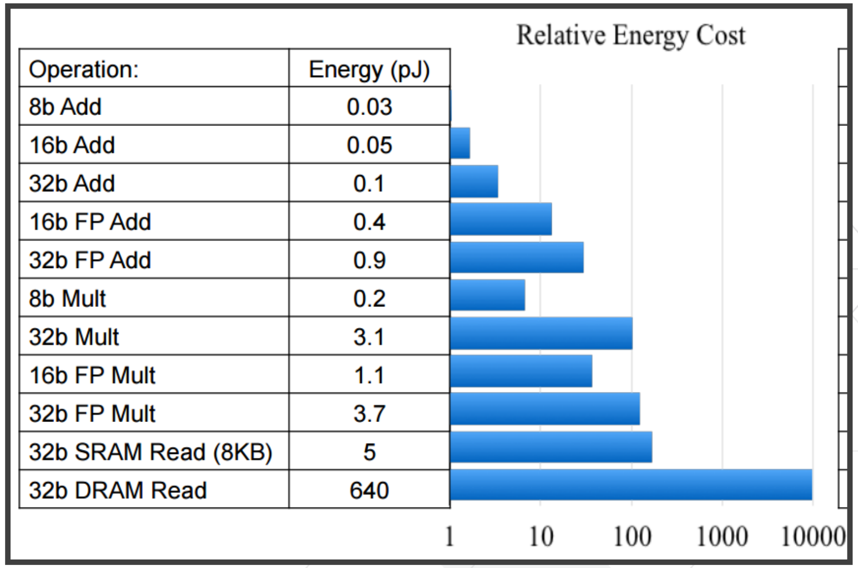
Reducing Precision Inherently Saves Power

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:

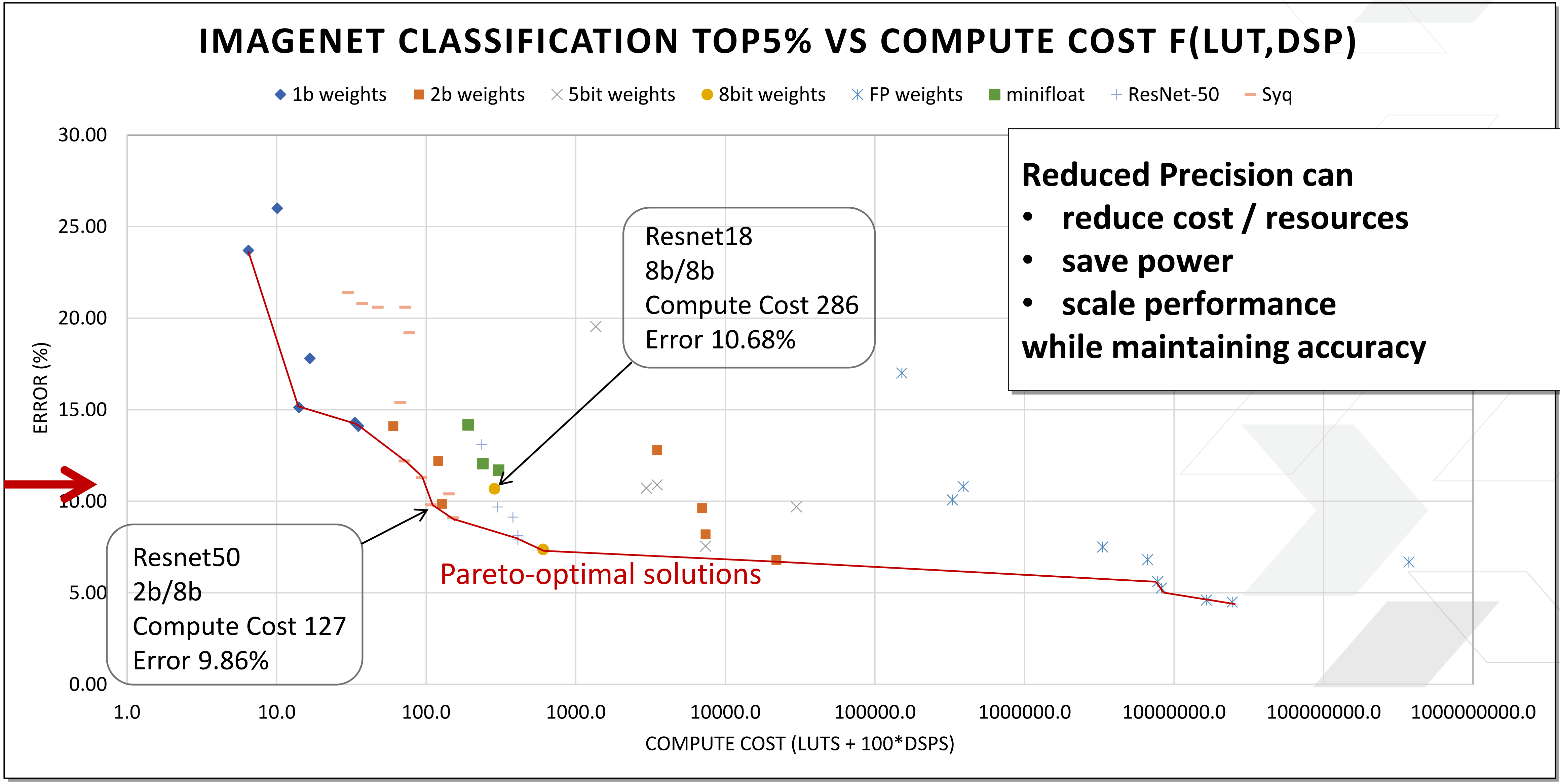


Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

[Adapted from Horowitz. Computing's Energy Problem (and what we can do about it), ISSCC'14]

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."

Design Space Trade-Offs

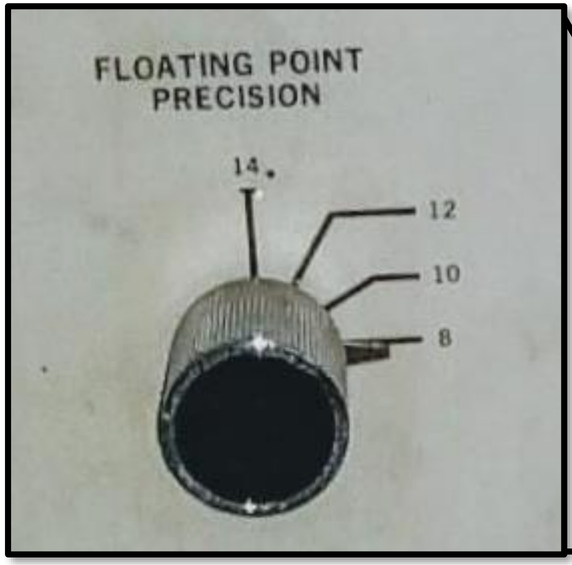


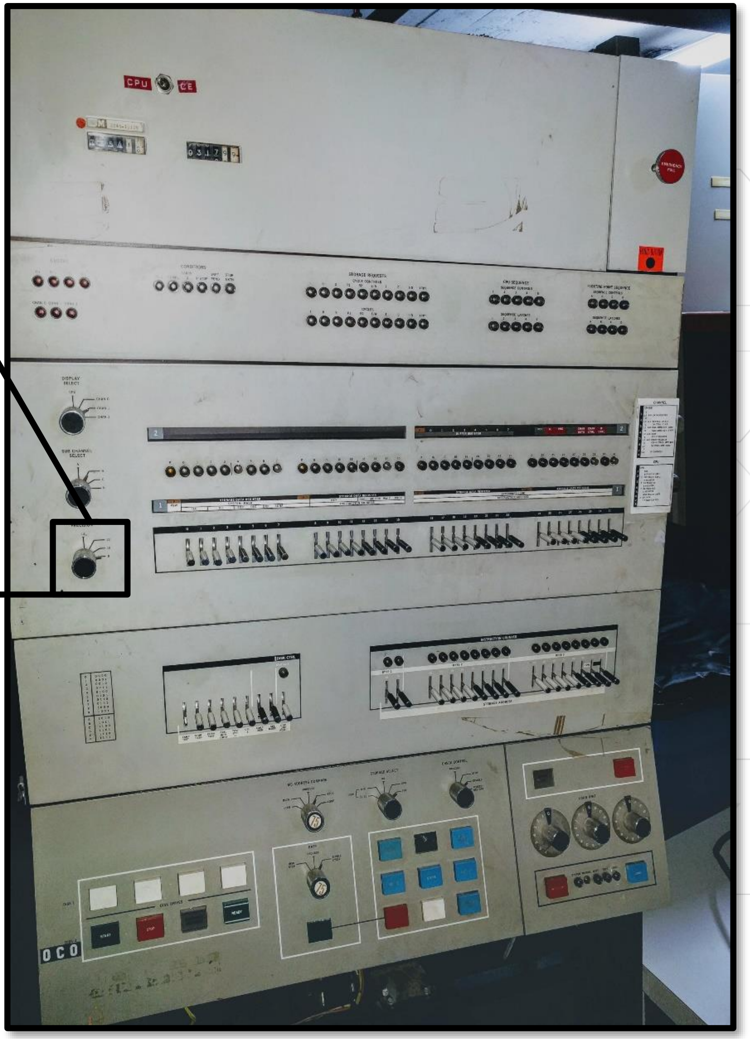
Reduced Precision can
- **reduce cost / resources**
- **save power**
- **scale performance**

while maintaining accuracy

Even More Extreme:
Run-Time Programmable Precision

IBM System/360 Model 44 (announced 1965)



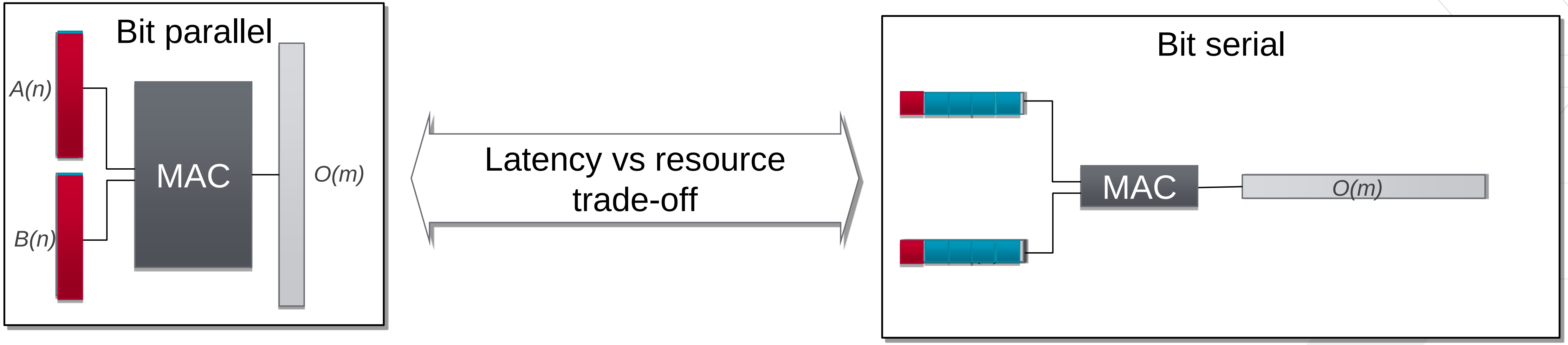


"One unusual feature of the Model 44's console was a rotary knob to ***select floating point precision; reducing the precision increased speed."***

[Source: http://www.righto.com/2019/04/iconic-consoles-of-ibm-system360.html]

Bit-serial Architectures Can Provide Run-time Programmable Precision with Fixed Architecture

Comparison to Traditional Bit-Parallel:



> **FPGA Evaluation for Matrix Multiply: Flexibility comes at almost no cost and provides equivalent bit-level performance at chip-level for low precision***

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

Summary

Summary

- **Innovative architectures emerge to address the needs of latest technology trends**
- **Customized memory subsystems, dataflow architectures and precisions can provide dramatic**
  - **Performance scaling**
  - **Latency reductions**
  - **Energy savings**

Challenges

- **Programming complex systems**
- **Benchmarking heterogeneous systems for specific applications**
  - That are fundamentally differently programmed

It is already.
Computer Architecture has never been as exciting.

XILINX

THANK YOU!

Adaptable.
Intelligent.

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

XILINX