Unconventional Compute Architectures with Reconfigurable Devices in the Cloud

Michaela Blott
Principal Engineer
Sep. 2018

XILINX

Agenda

Background

Motivation

Heterogeneous Hardware Platforms

- System to Device-Level
- Unconventional Examples

Background

Xilinx Research - Ireland

Ivo Bolsens
CTO

- Since 13 years
- Part of the worldwide CTO organization (8 out of 36)
- AI Lab expansion part-financed through

IDA Ireland

Kees Vissers
Fellow

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 (from left to right)

Plus a Very Active Internship Program

- **On average 4-6 interns at any given time**
  - From top universities all over the world
  - We are always looking for talent ;-)

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

Mission: Application-Driven Technology Development

- Identify strategic applications (for the last 6 years within data centers)
- Derisk emerging technologies (HBM, OpenCL, HLS)
- In partnership with universities, customers, and partners
- **Current Focus: Quantifying value proposition for FPGAs in Machine Learning**
  - Prototyping, testdriving, benchmarking

Motivation

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

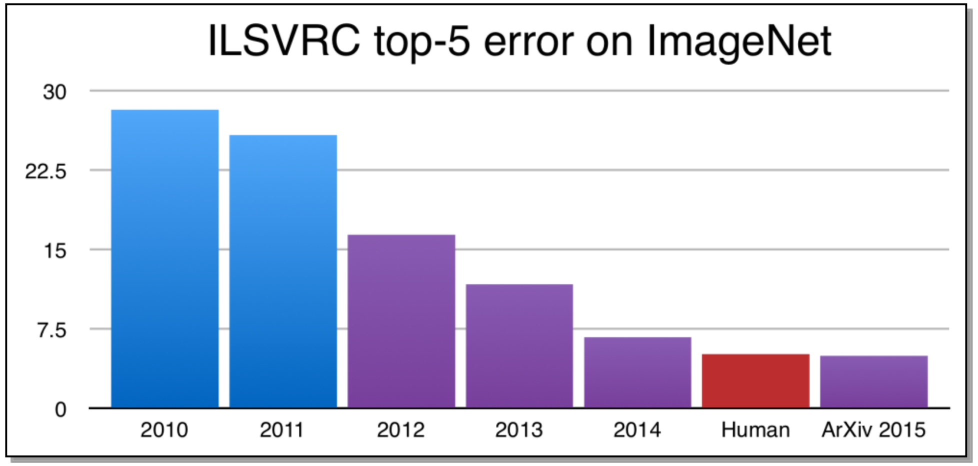
> **What is the potential and the challenge?**

XILINX

Convolutional Neural Networks (CNNs)

Why are they so popular?

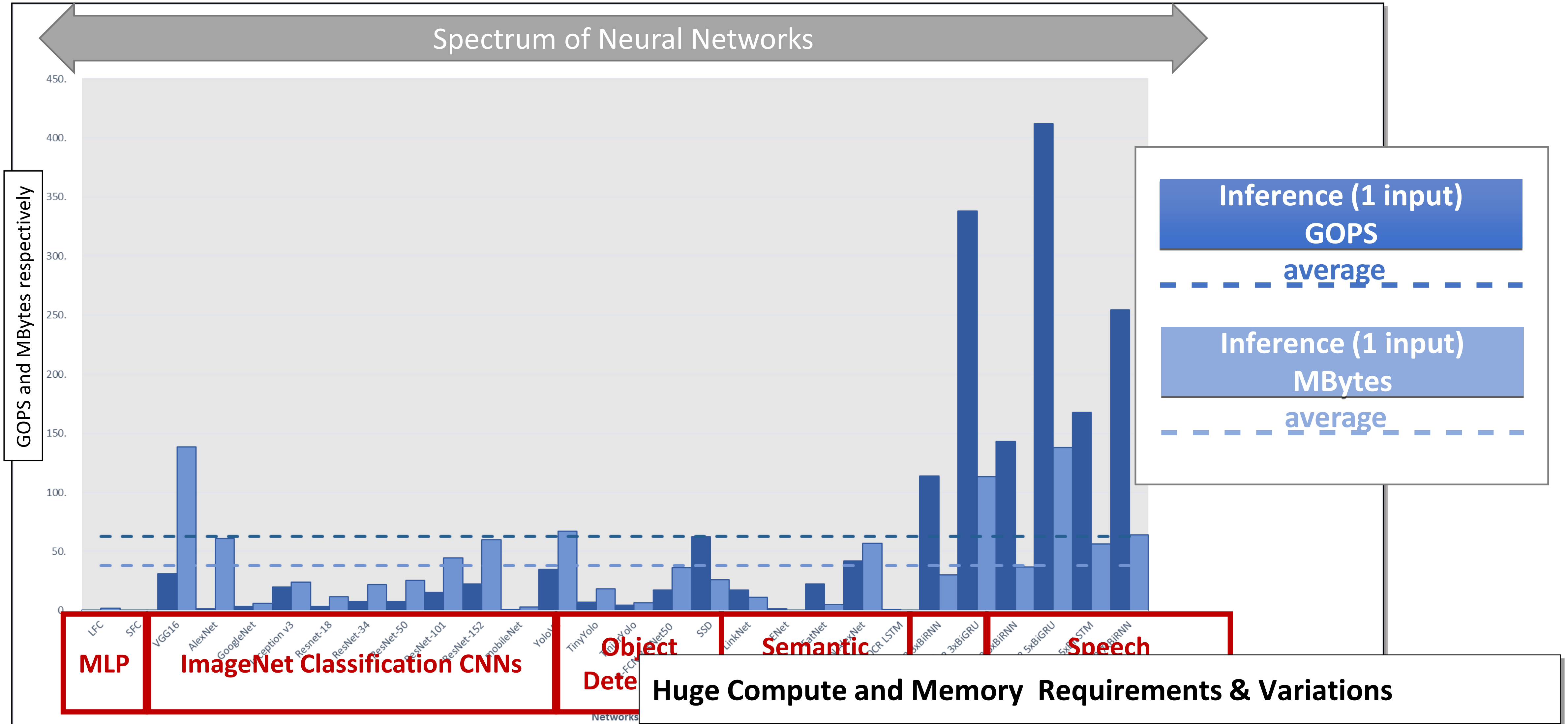
- **Requires little or no domain expertise**
- **NNs are a “universal approximation function”**
- **If you make it big enough and train it enough**
  - Can outperform humans on specific tasks



- **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**

Compute and Memory for Inference

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



Huge Compute and Memory Requirements & Variations

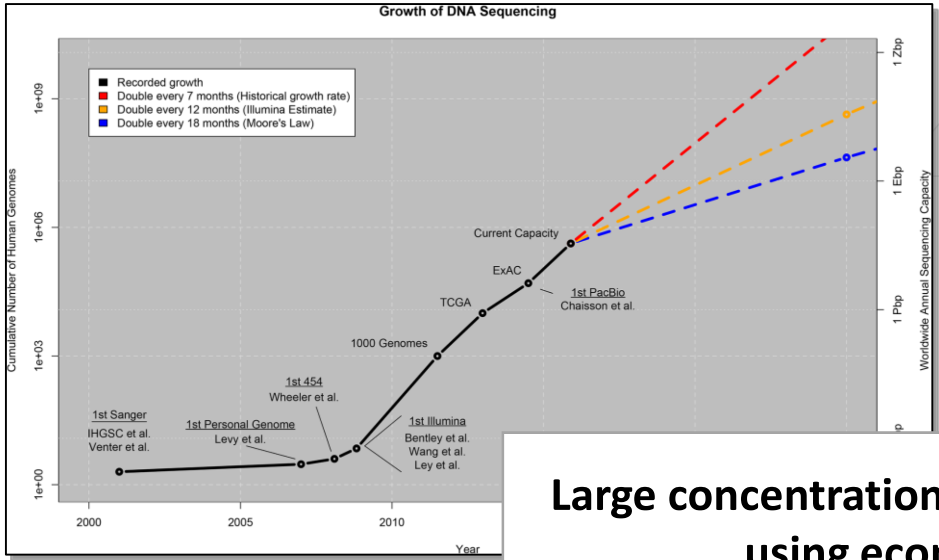
Trend:
Explosion of Data

> **Computing shifts towards cloud computing**

> **Data storage requirements explodes**

>> Photos => videos

>> DNA!

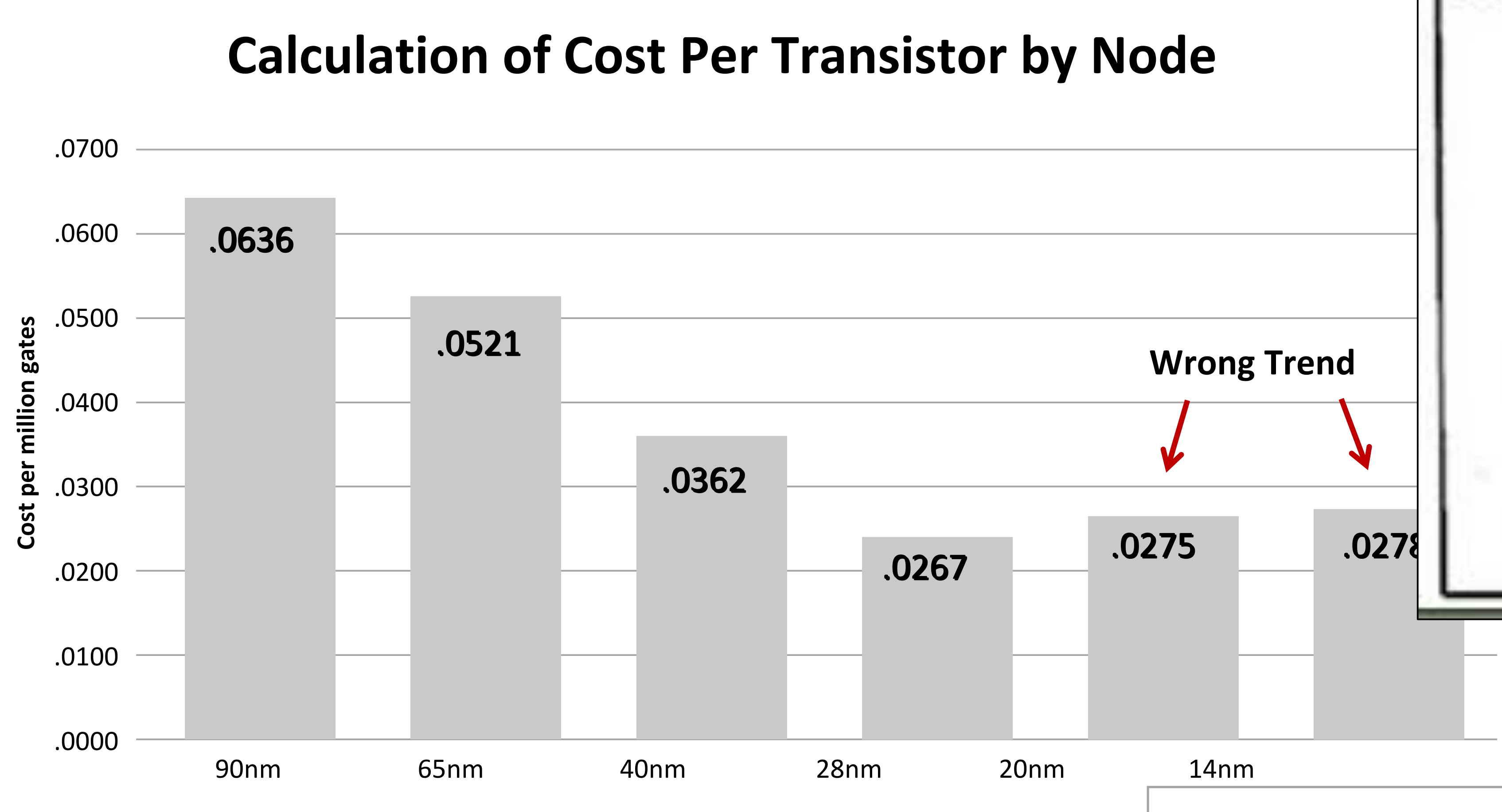


Stephens, Zachary D., et al. "Big data: astronomical or genomical?." PLoS biology 13.7 (2015): e1002195.

Large concentration of compute and storage using economics of scale

XILINX.

Technology:
End of Moore's Law

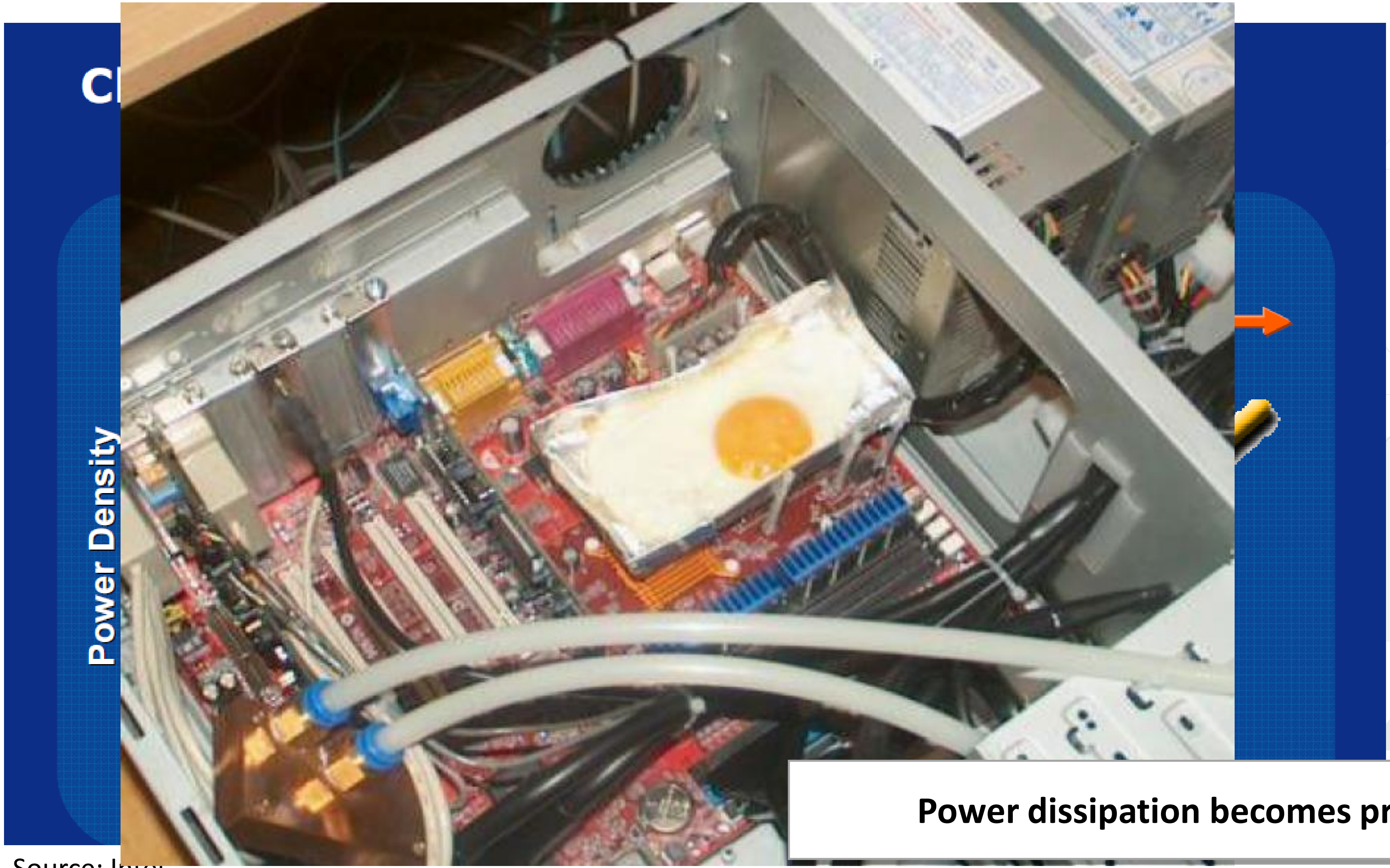


Source: IBS

Economics become questionable

XILINX

Technology:
End of Dennard Scaling

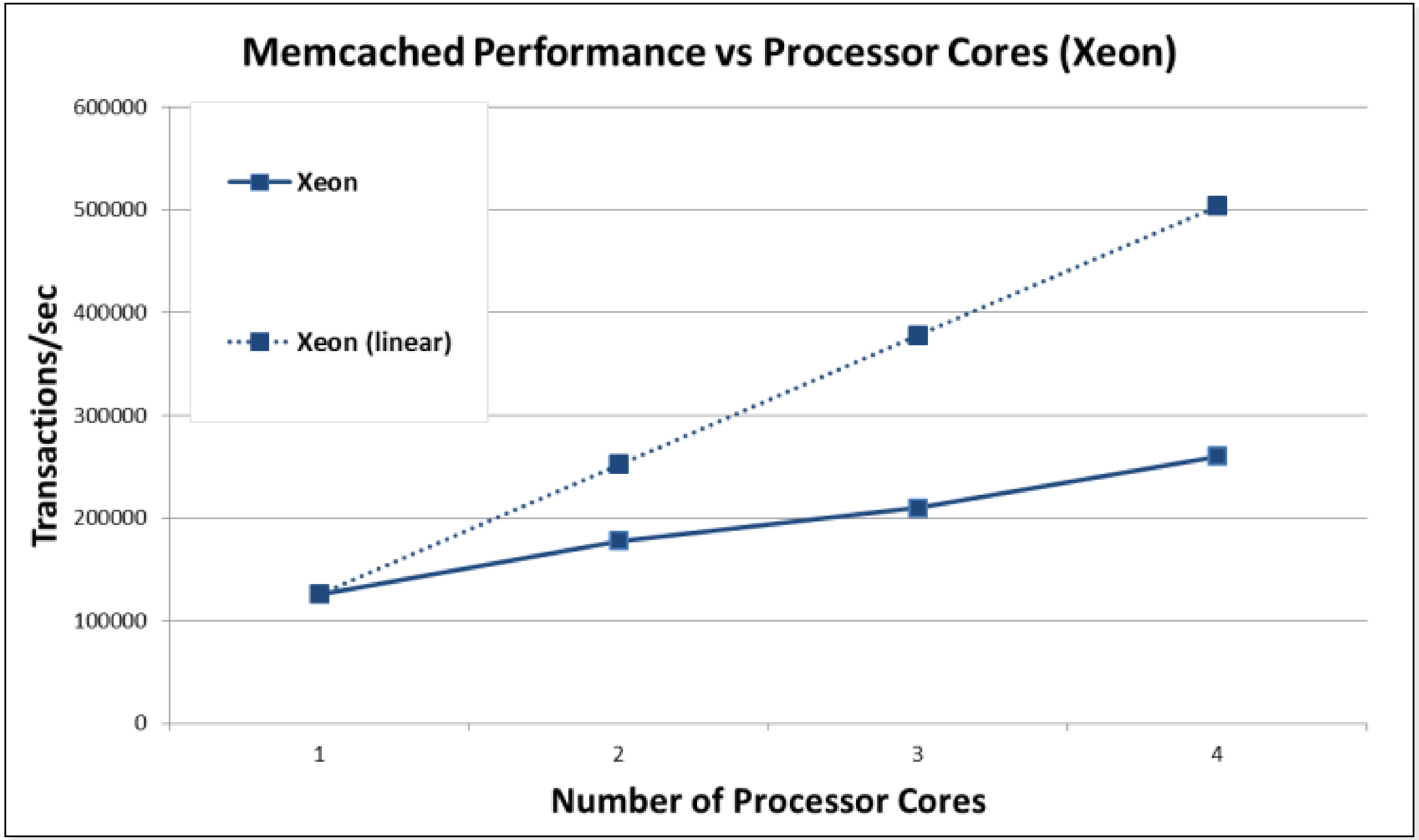


Power dissipation becomes problematic

Source: Intel

XILINX

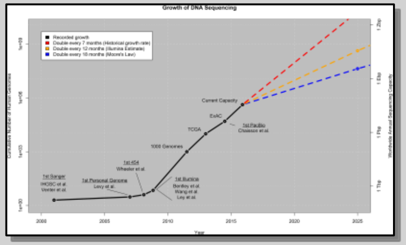
Technology:
Traditional Compute Architectures Are Not Scalable



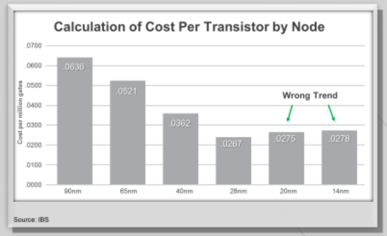
XILINX

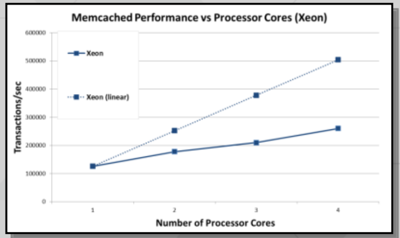
Era of Heterogeneous Compute & Accelerators Has Arrived

Trends



Technology

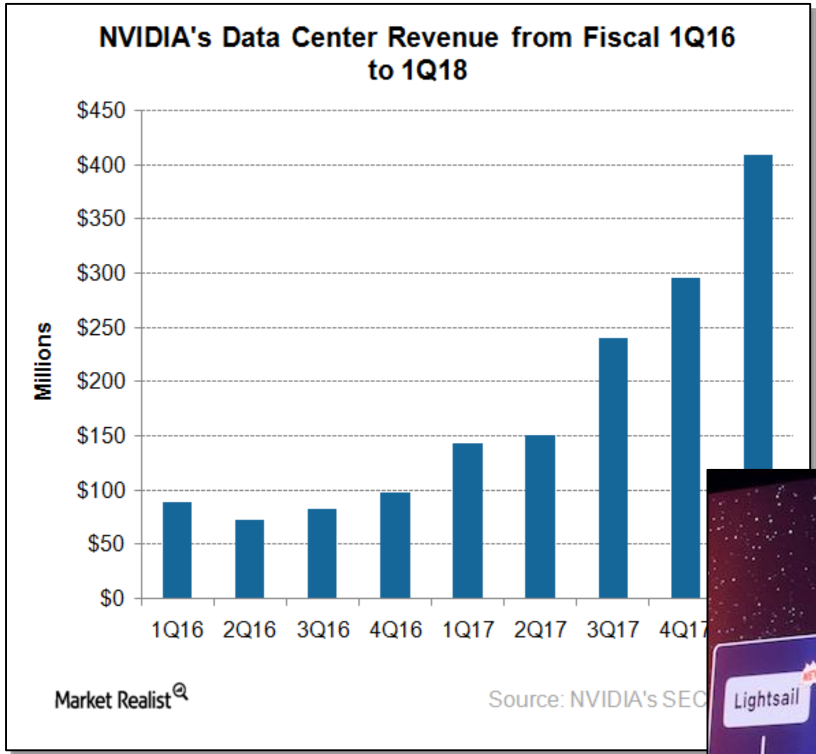


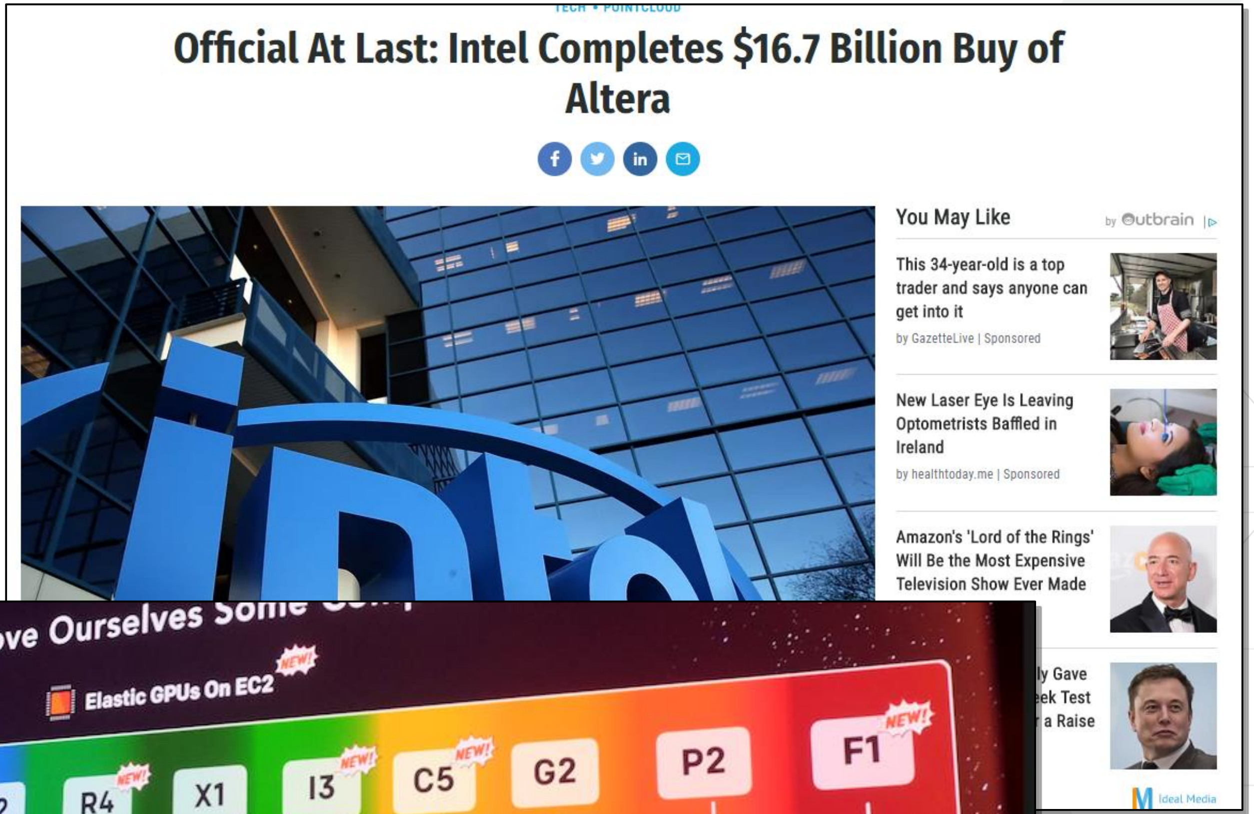


> Diversification of increasingly heterogenous devices and system

XILINX.

Evidence



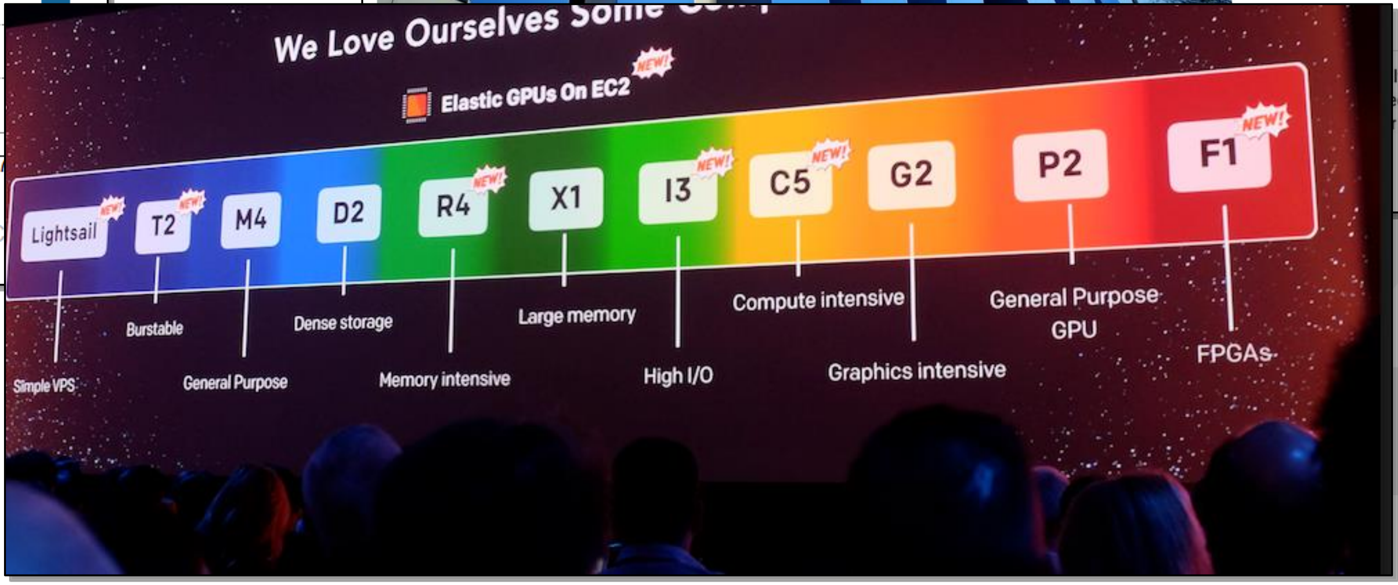

Official At Last: Intel Completes $16.7 Billion Buy of Altera

You May Like

This 34-year-old is a top trader and says anyone can get into it
by GazetteLive | Sponsored

New Laser Eye Is Leaving Optometrists Baffled in Ireland
by healthtoday.me | Sponsored

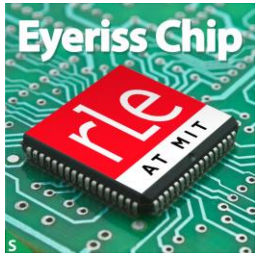
Amazon's 'Lord of the Rings' Will Be the Most Expensive Television Show Ever Made



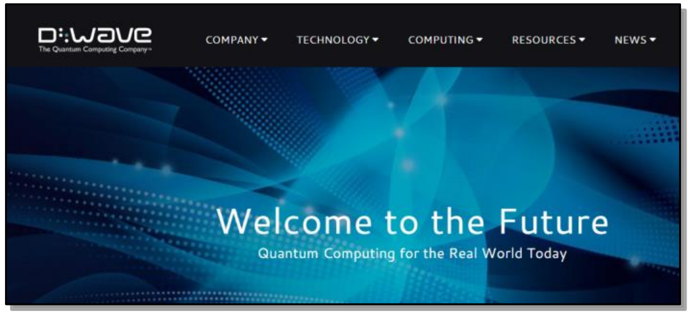
Insight 2016: AWS adding FPGA instances

Wave of Customized Hardware for AI

> Custom AI Silicon

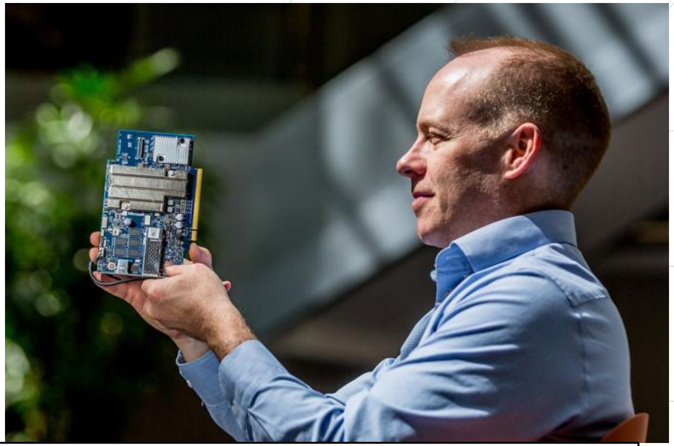


> Quantum computing



> FPGA solutions

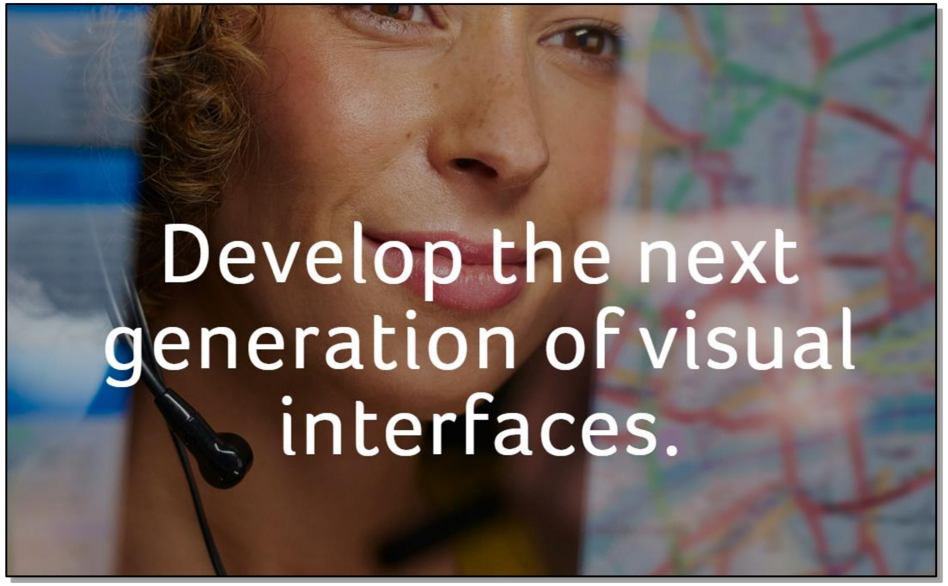
>> Microsoft Catapult and Bainwave



> Cloud economics enable adoptions

XILINX.

FPGAs in Video Processing - Skreens



Cloud-Based API

RESTful API allows service and content providers to integrate the power of Skreens video processing across their solutions.

One Encoder per Person

No set top box necessary. Skreens' innovative cloud encoding allows for limitless personalization of digital media.

Interactive Multi-Layering

Combine, customize and arrange any video, web content, advertising or interactive inputs using our patented multi-layering technology.

Machine Learning

Apply off-the-shelf or custom algorithms to analyze video frame-by-frame, taking action as appropriate to prioritize content or dynamically update displays.

Powered by Xilinx

The best video processing, now on the cloud. Skreens makes the most recent programmable-chip acceleration accessible for any service without hardware

FPGAs in Crypto-Currency Mining

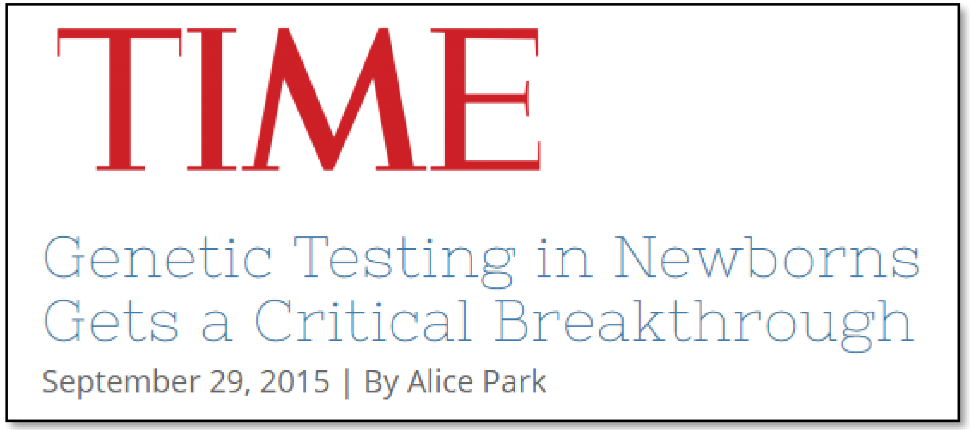
- FPGAs are very good at hash algorithm which require bit-level fiddling
- Net benefit of mining is determined by the used energy cost
  - FPGAs are typically more energy-efficient than GPUs and CPUs

XILINX

FPGAs in Genomics Acceleration Example

- Important compute problem for personalized medicine
  - >> https://www.youtube.com/watch?v=u6Q4_L2_ZnA

- FPGA value
  - Dramatic speed-up to enable real-time diagnosis
  - Analysis reduced from days to 20 minutes


TIME

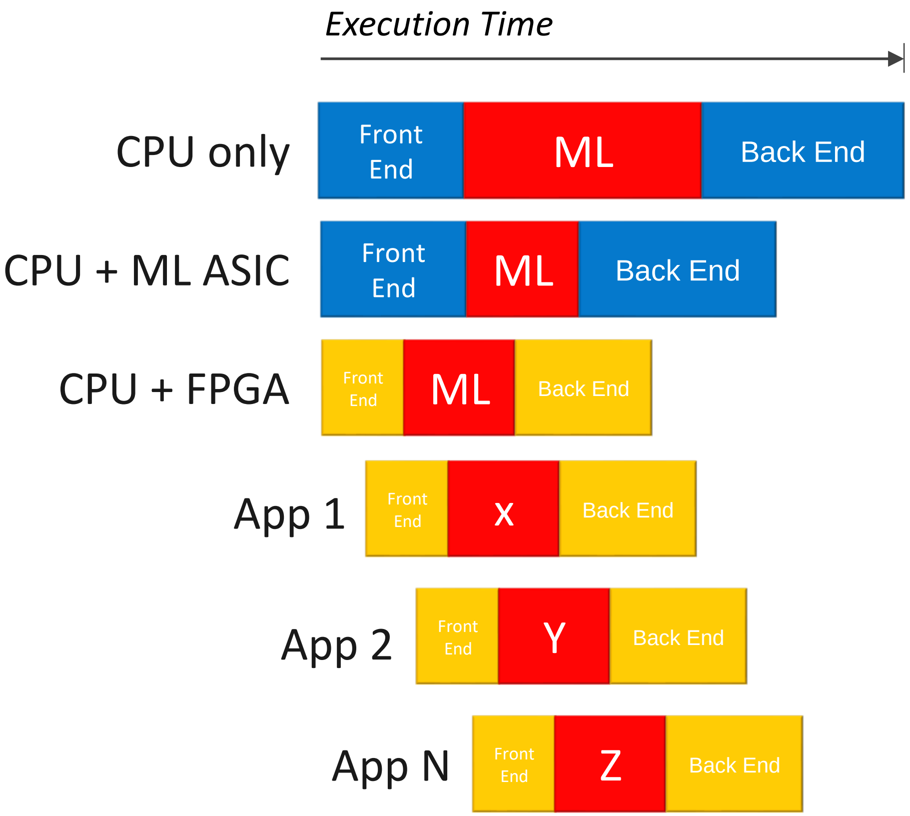
Genetic Testing in Newborns Gets a Critical Breakthrough

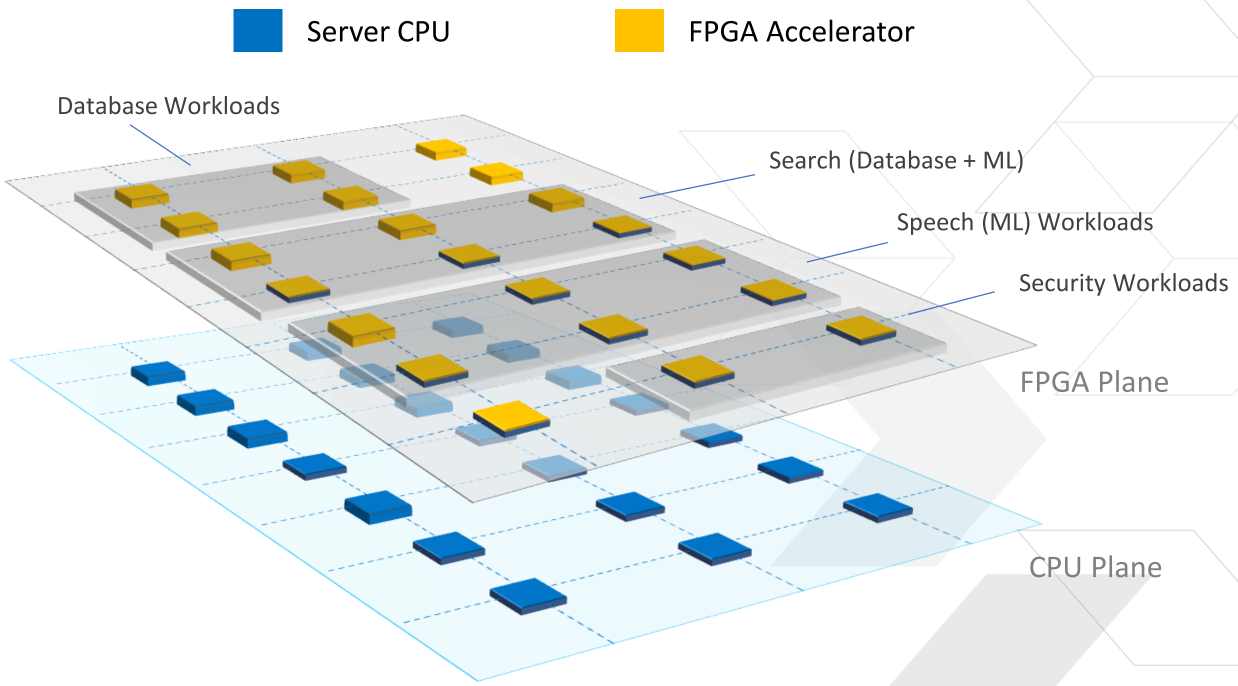
September 29, 2015 | By Alice Park

Application Acceleration with DSAs on FPGA Platforms

> **DSAs for different applications – dynamic optimizations for changing workloads**

> **When networked, opportunity to directly scale-out**





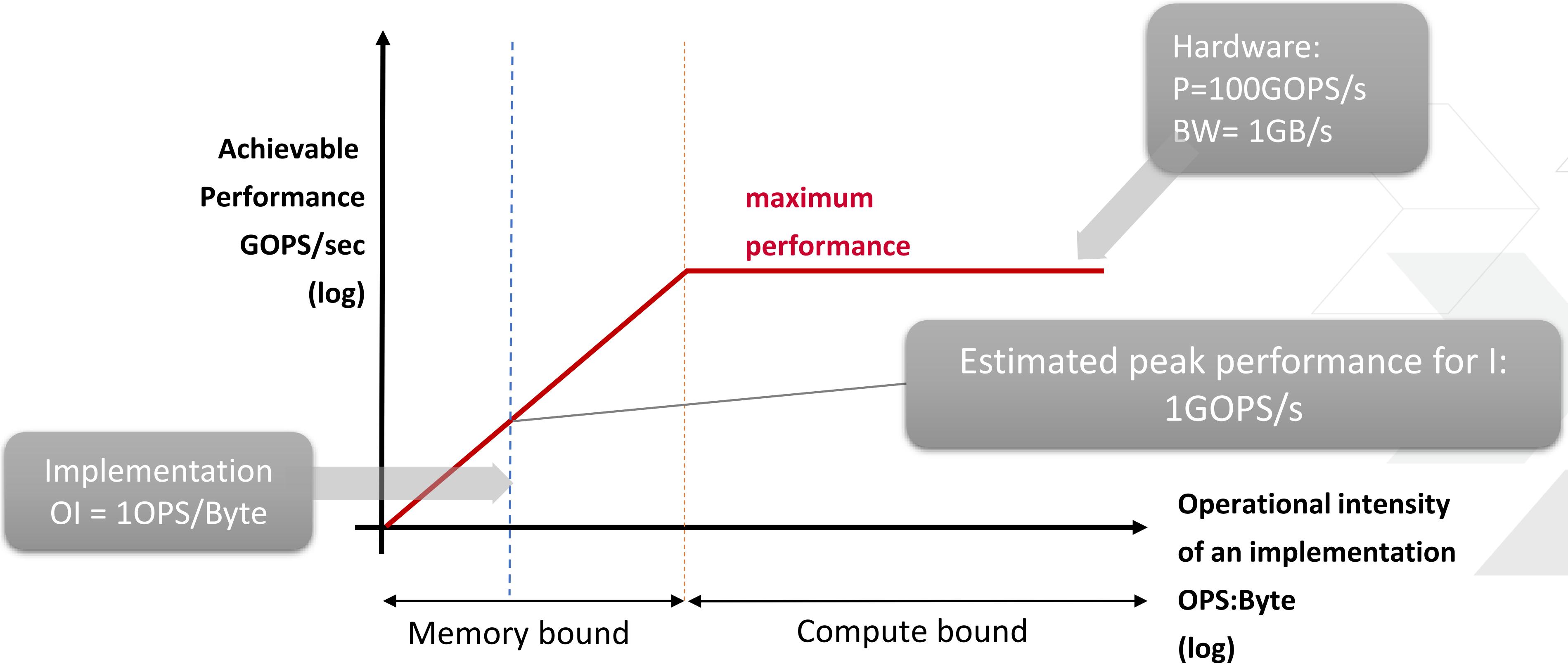
Hyperscale Data Centers with FPGA Accelerators

>Beyond FPGAs:

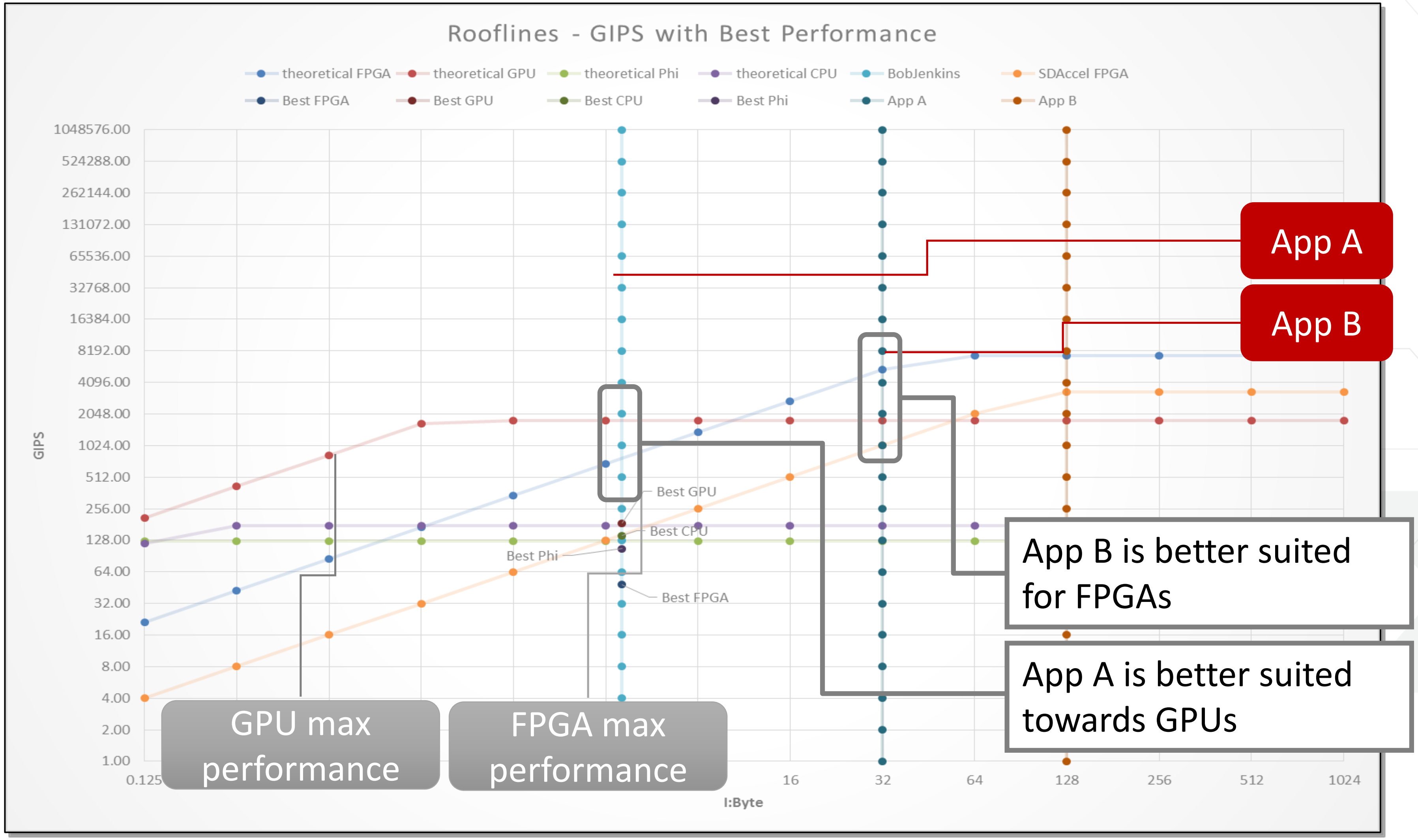
Increasing number of customized accelerators

Rooflines for Hardware Platforms

> **Peak performance as a function of operational intensity**
  - *P = min{ OI*BW; P}*



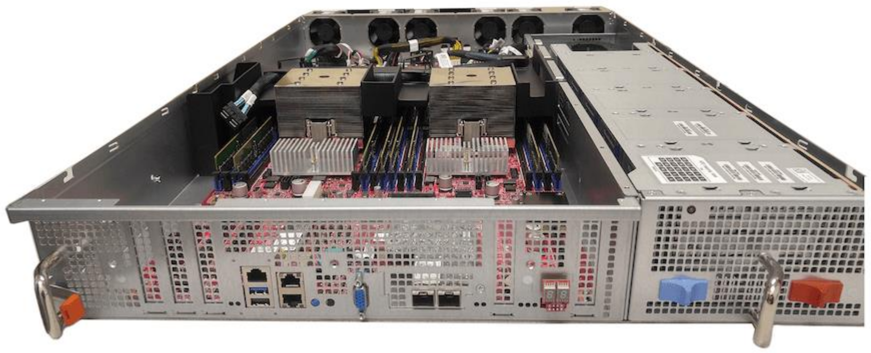
Horses for Courses



Increasingly Heterogeneous Hardware Platforms

System Level: Diversification with Accelerator Support

HP Moonshot

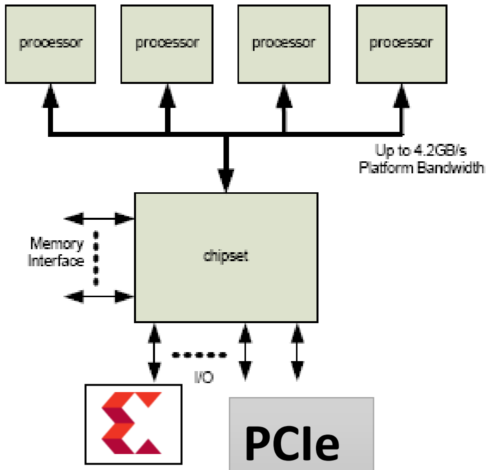


OpenPOWER™

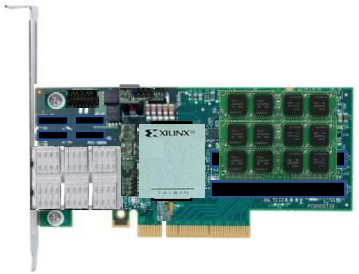
IBM's OpenPower

Accelerator Integration – Moving Closer to CPU

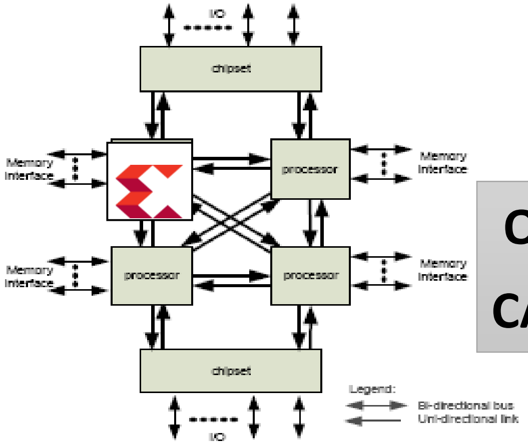
IO-Connected



PCIe

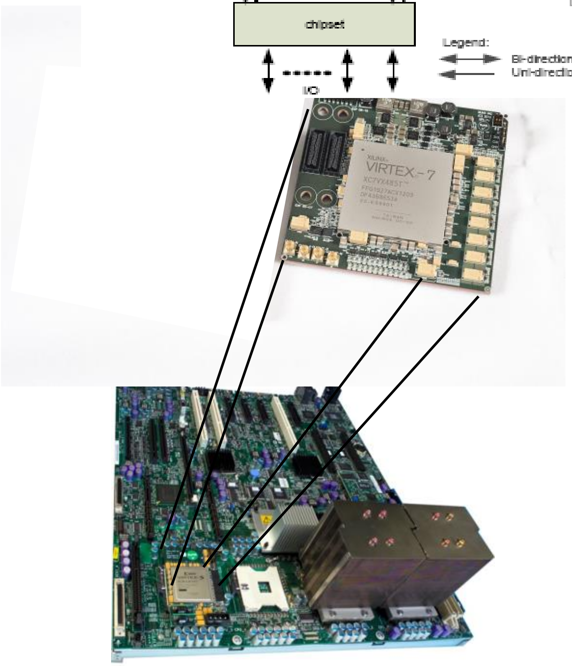


Coherent

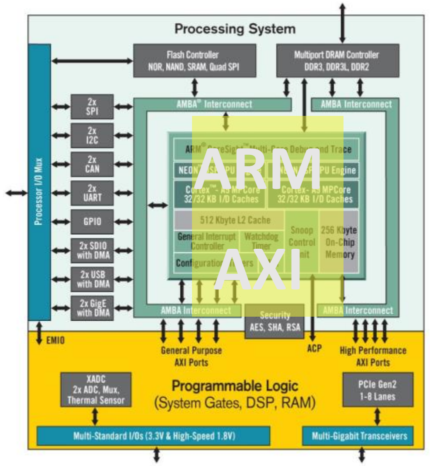


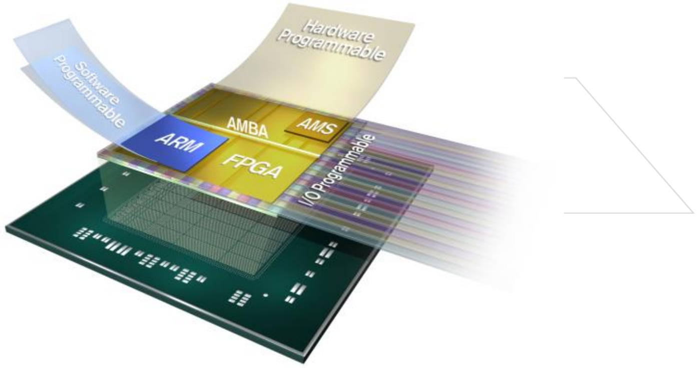
QPI

CAPI



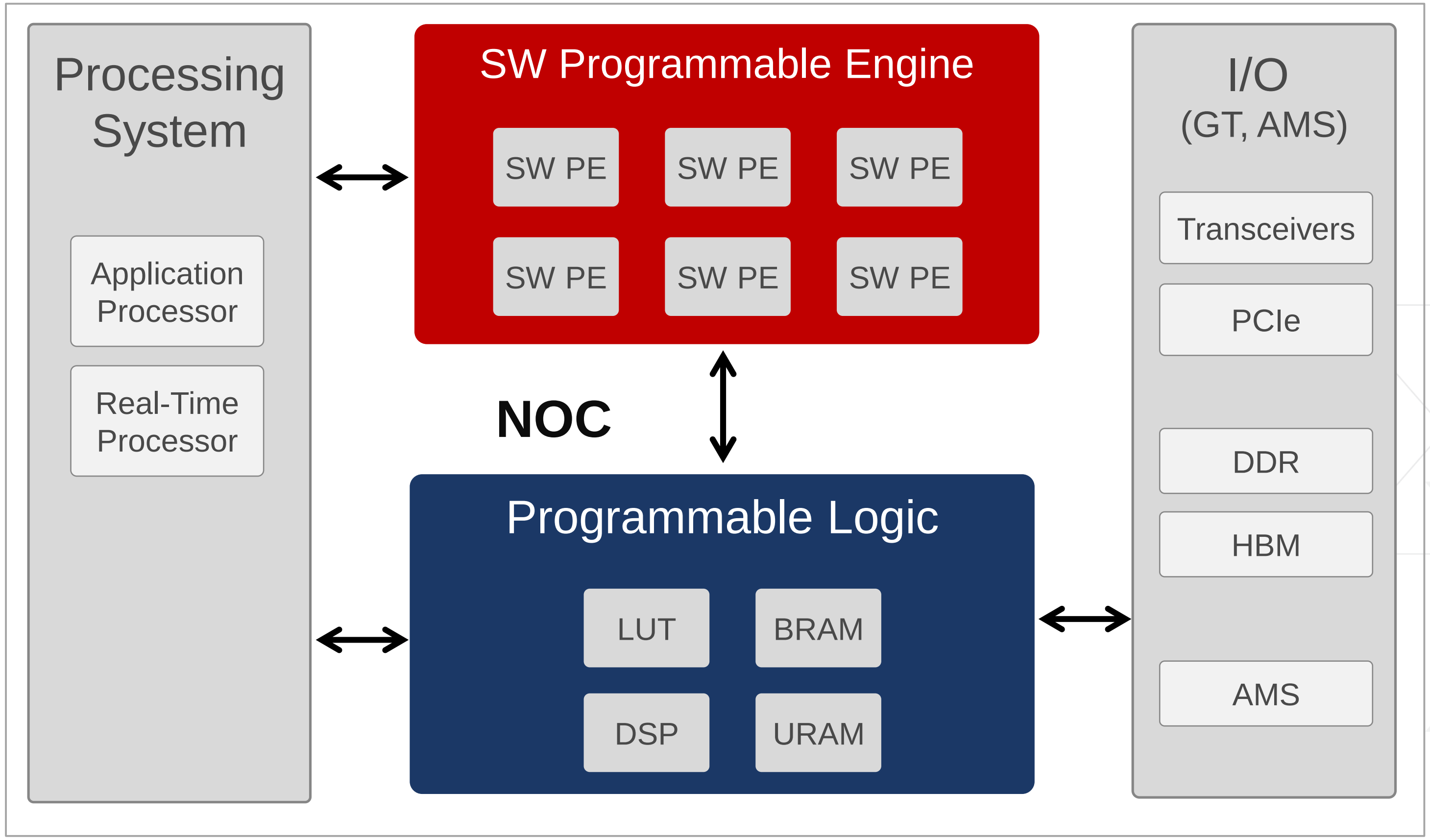
Integrated SOC





XILINX

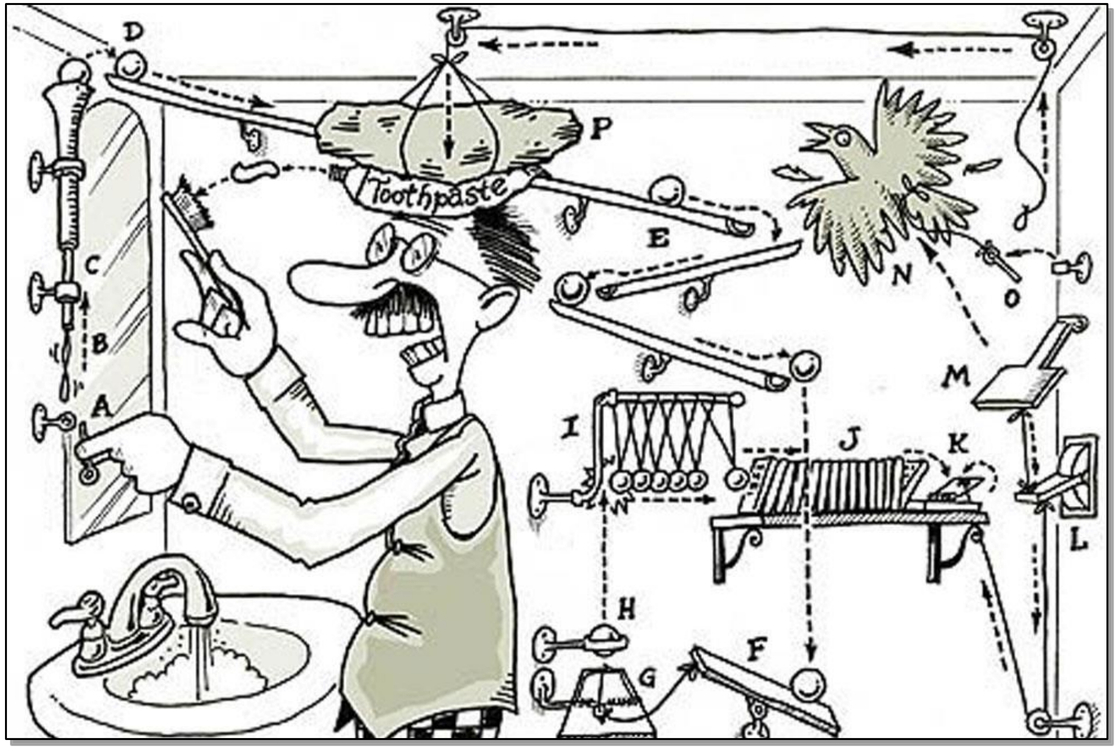
Device-Level:
Evolution of FPGAs to ACAPs

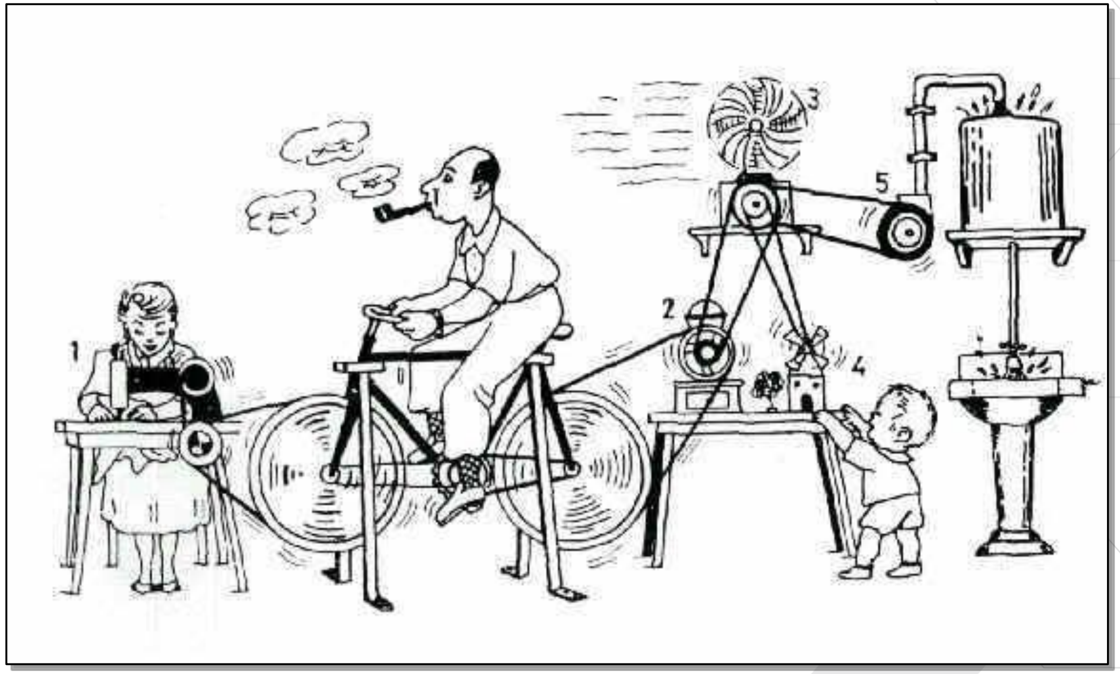


Exciting Times in Computer Architecture Research!

XILINX

Unconventional Architectures Emerge





> On system and device level…

Unconventional Examples

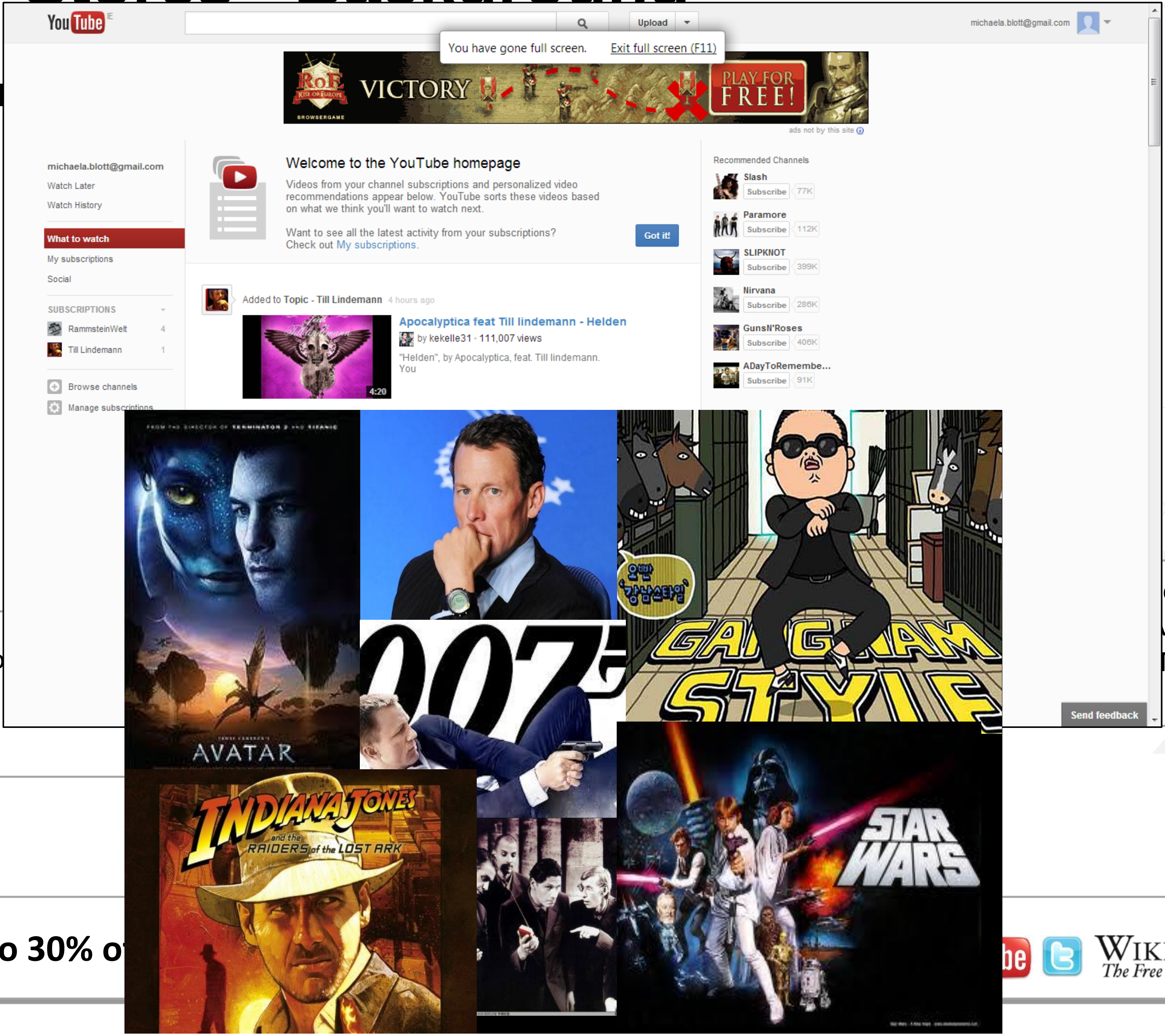
XILINX

Key-Value Stores

XILINX

Key Value Stores - Background

> Many popular



Only sto
recent
records

ool of x86-
vers with
M running

Up to 30% o

Current Implementations

- **Multithreaded implementation (pthreads)**
  - Each request is a connection
  - All threads execute drive_machine(), processes connections from one state to next, and switches over connection state
  - Shared data structures (hash tables, value store,…)

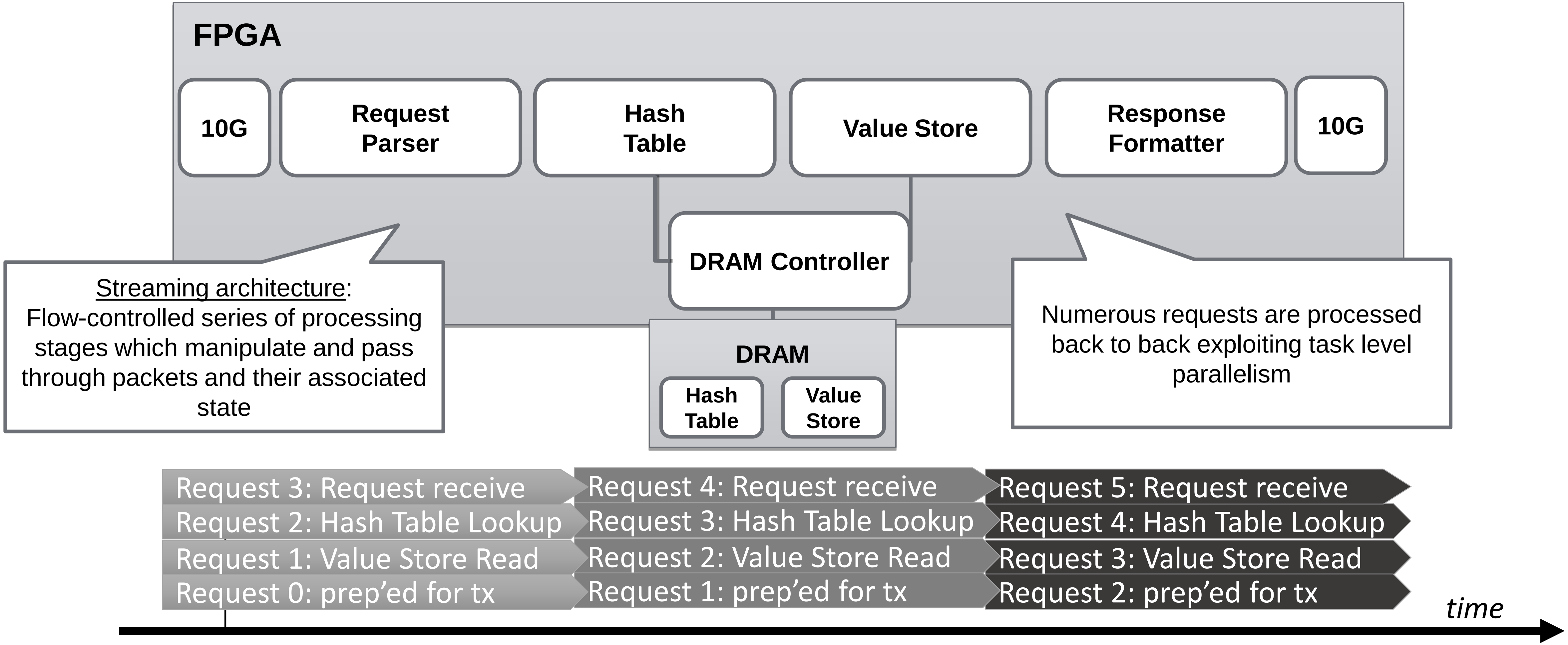
```
Receive & parse
Hash lookup
Value store access
Format & transmit
```

```
drive_machine():
while (!stop) {
   switch(c->state) {
      case connection_waiting:
      case connection_closing:
       …
      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
  - TCP/IP is CPU intensive, interrupt intensive, too large to fit into instruction cache (up to 160 MPKI)
  - 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, 7KRPS/Watt

Dataflow Architectures to Scale Performance



> **10Gbps demonstrated with a 64b data path @ 156MHz using 3% of FPGA resources**

> **80Gbps can be achieved by using a 512b @ 156MHz pipeline for example**

Source: [4] Blott et al: Achieving 10Gbps line-rate key-value stores with FPGAs; HotCloud 2013

Deep Learning

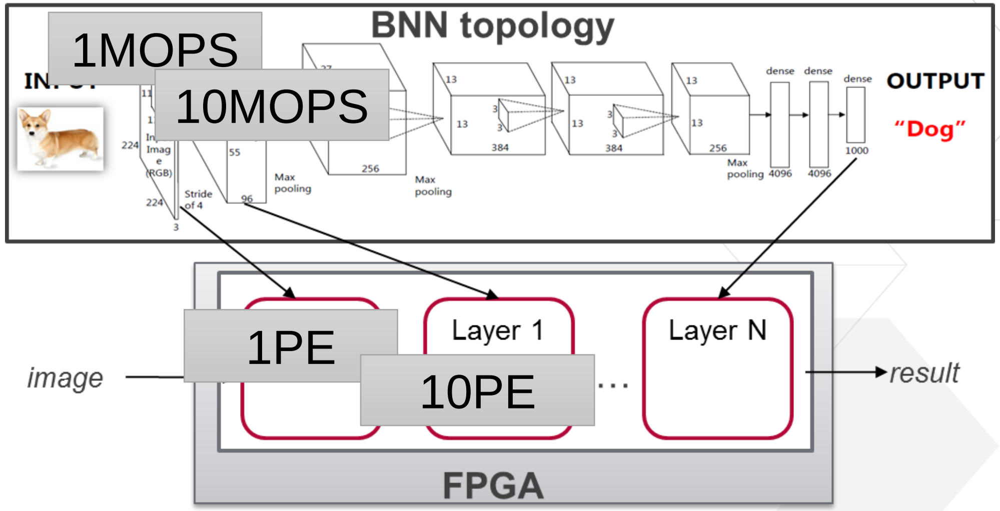
Custom-Tailored Hardware Architectures (Macro-Level)

Hardware Architecture Mimics the NN Topology

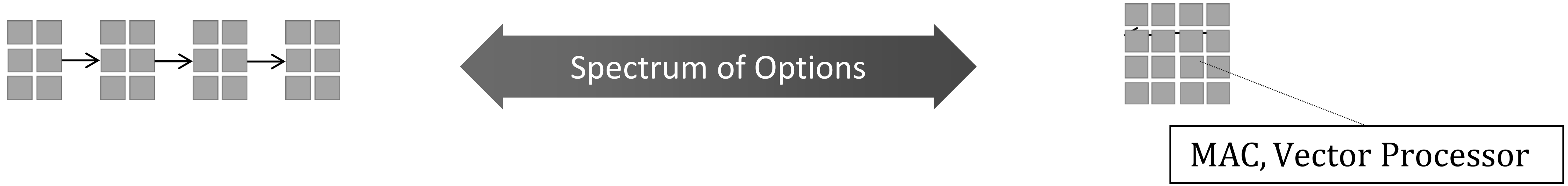
> **Customized feed-forward dataflow architecture to match network topology**
  - >> Only FPGAs can do this

> **Customized to meet design requirements**
  - >> Scaling towards resource and performance targets

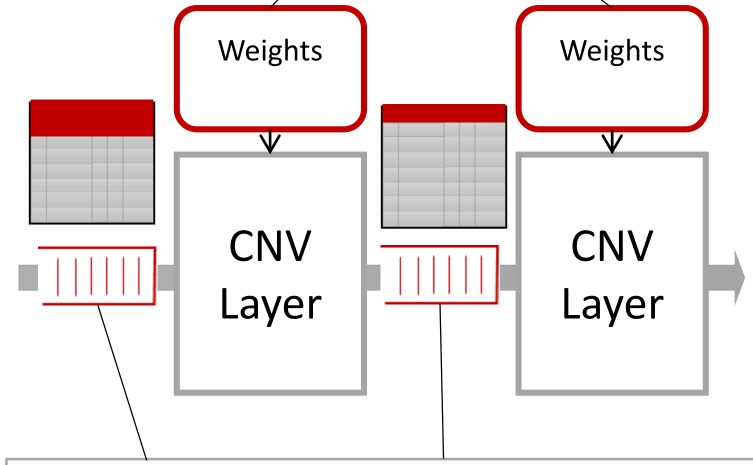
> **Customized micro-architecture**



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



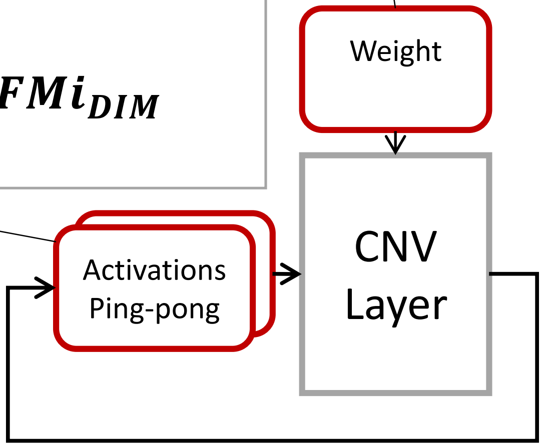
$$WC_{memory} = SUM(Wi)$$



$$AT_{memory} = Sum(ATi)$$
$$ATi = batch * FMi_{DIM} * CHi * (Ki + Si)$$

$$WC_{memory} = MAX(Wi)$$

$$AT_{memory} = 2 * batch * Max(FMi_{DIM} * FMi_{DIM} * \#CHi)$$



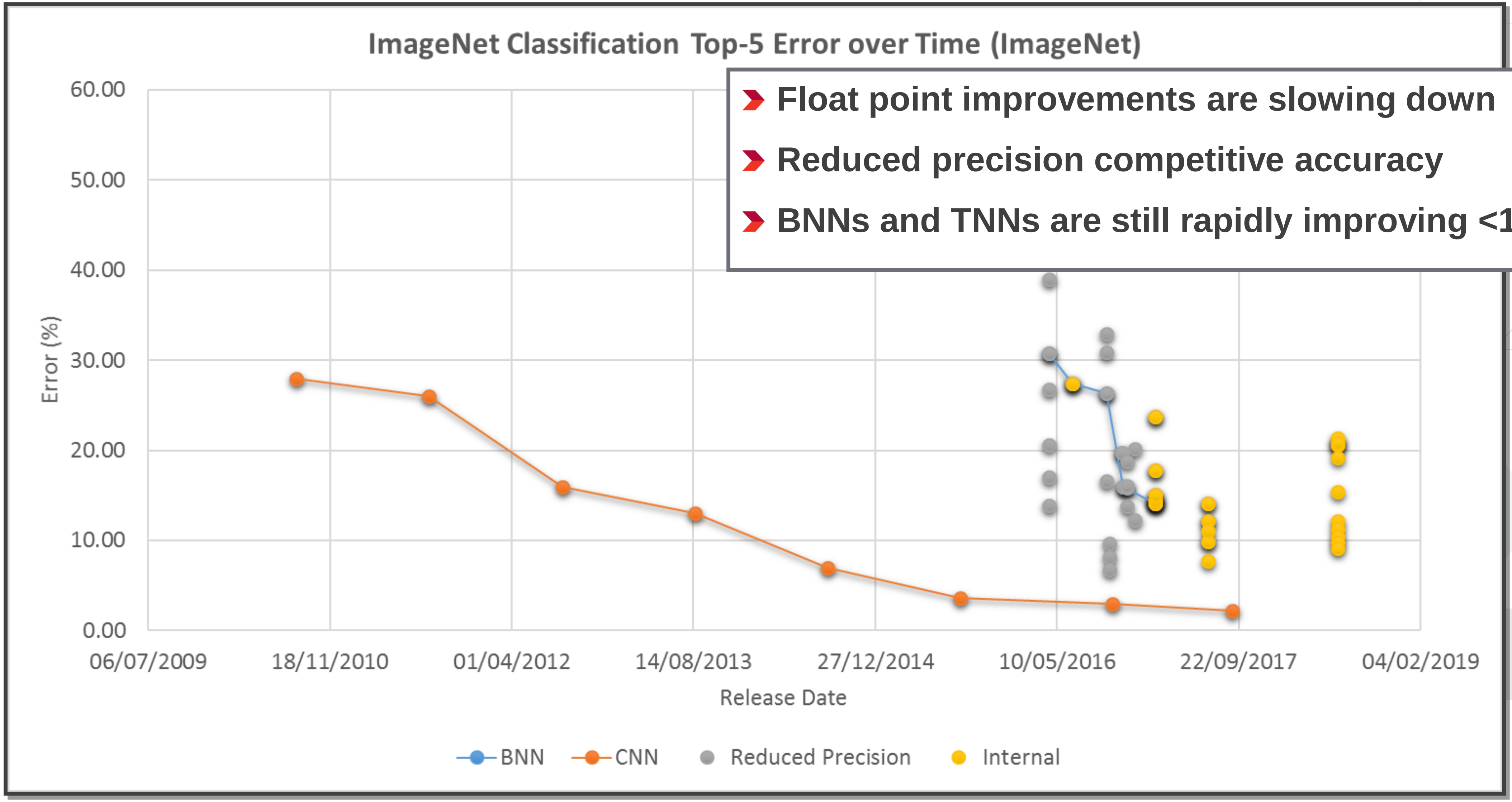
Lin, X., Yin, S., Tu, F., Liu, L., Li, X. and Wei, S. LCP: a layer clusters paralleling mapping method for accelerating inception and residual networks on FPGA. DAC'2016
Alwani, M., Chen, H., Ferdman, M. and Milder, P. Fused-layer CNN accelerators. MICRO 2016.

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

- Requires less activation buffering, more weights
- Higher compute and memory efficiency due to custom-tailored hardware design
- Less flexibility
- Less latency (reduced activation buffering)
- No control flow (static schedule)

- Requires less on-chip weight memory, but more activation buffers
- Efficiency of memory for weights and activations depends on how well balanced the topology is
- Flexible hardware, which can scale to arbitrary large networks
- Higher latency
- Compute efficiency is a scheduling problem

Customizing the Micro-Architecture:
Reduced Precision Neural Networks



- Float point improvements are slowing down
- Reduced precision competitive accuracy
- BNNs and TNNs are still rapidly improving <10% top5

Reducing Precision

Scales Performance & Reduces Memory

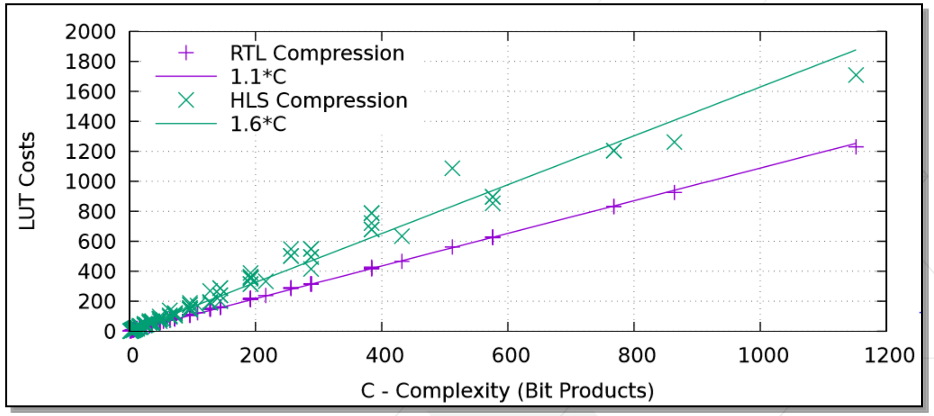
> **Reducing precision shrinks LUT cost**

>> 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



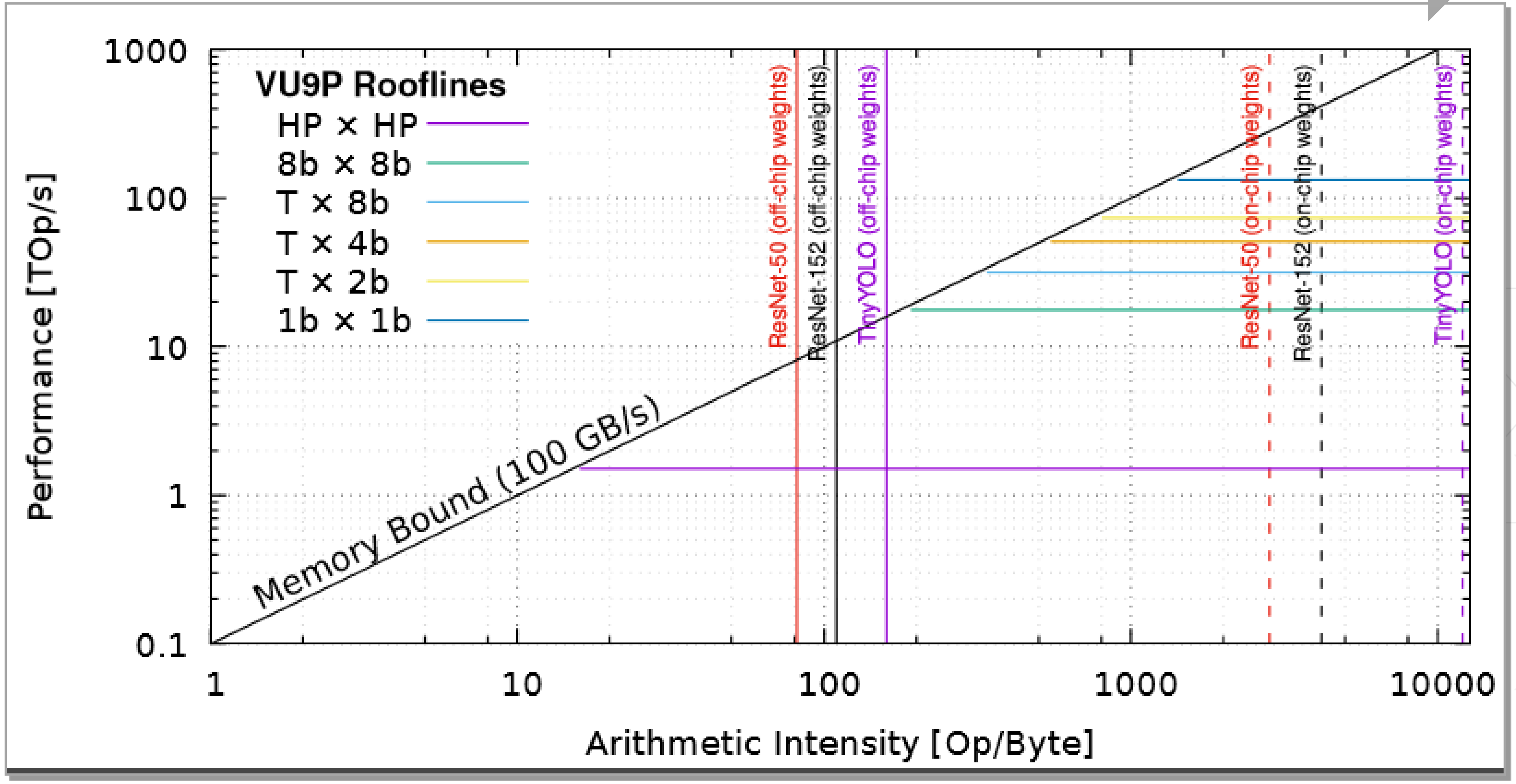
C= size of accumulator *
size of weight *
size of activation

Reducing Precision provides Performance Scalability

Example: ResNet50, ResNet152 and TinyYolo

RP reduces model size=> to stay on-chip

Up to 100x



RP scales compute performance

Theoretical Peak Performance for a VU9P with different Precision Operations

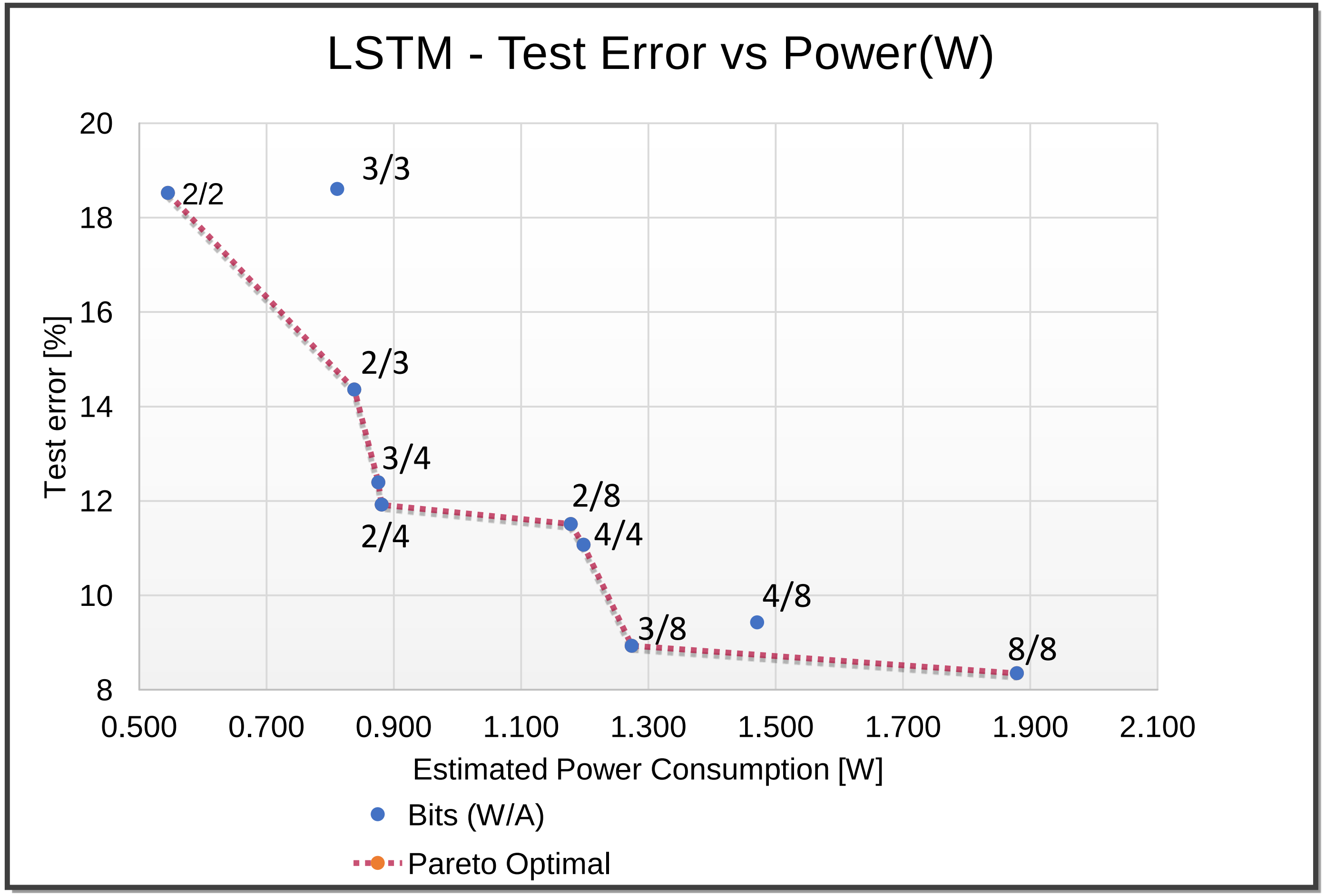
Assumptions: Application can fill device to 70% (fully parallelizable) 300MHZ

HLS overhea

> **Visualizing the benefits of customized compute**

Reducing Precision Inherently Saves Power

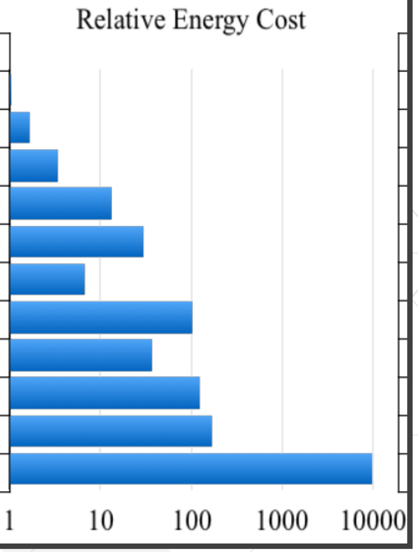
FPGA:



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

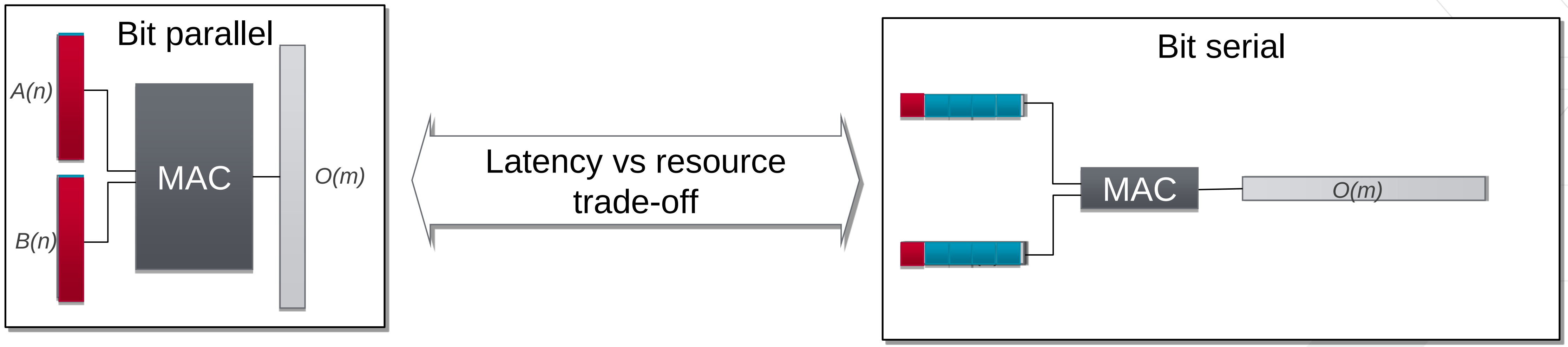
TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

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.

Even More Unconventional:
Bit-Parallel vs Bit-Serial

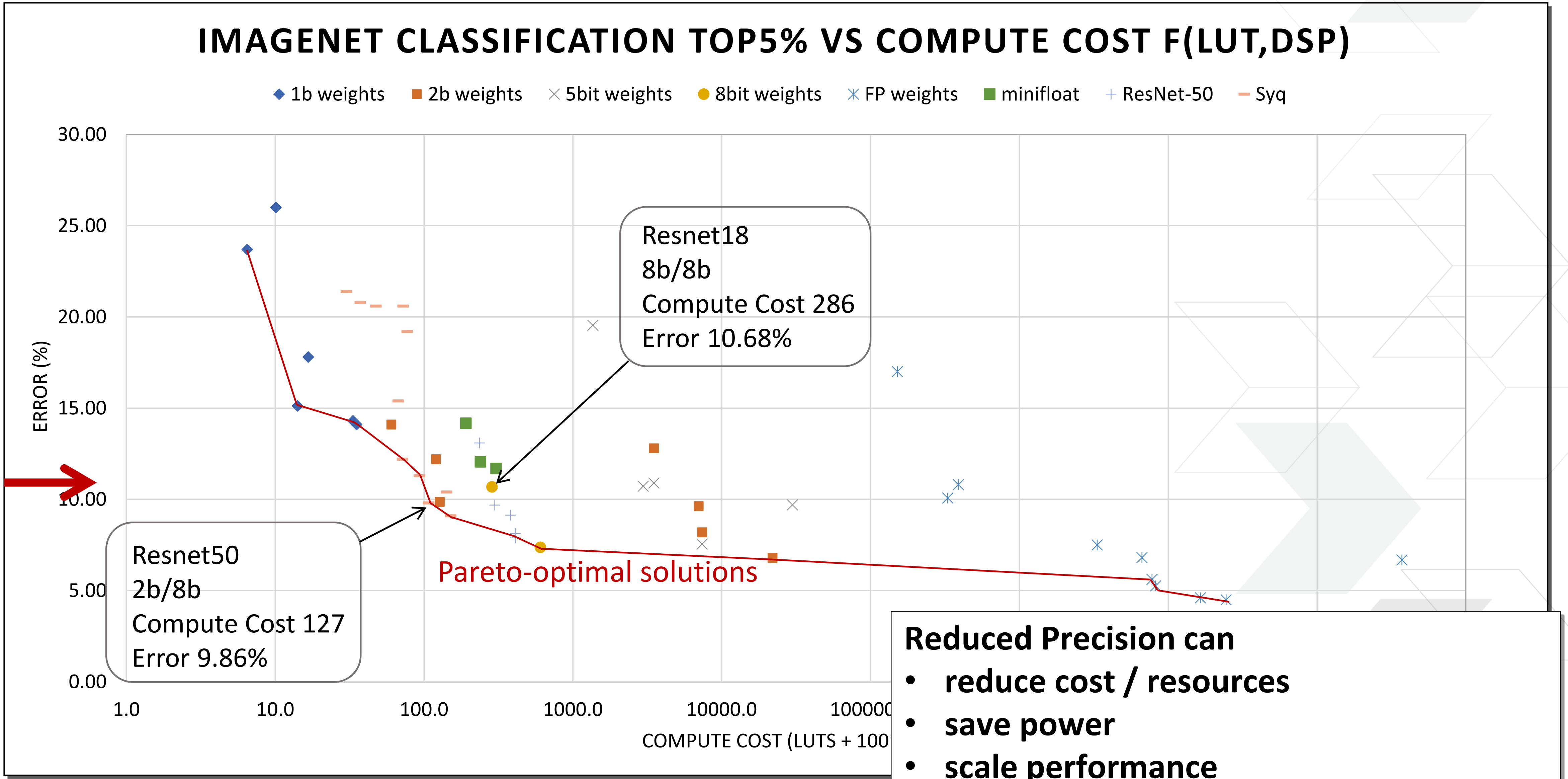
> **Furthermore, with bit-serial can provide run-time programmable precision with a fixed architecture**



> **FPGA: 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

Design Space Trade-Offs



Reduced Precision can

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

XILINX.

CNN	Platform	Clock (MHz)	BRAM18	LUTs	Perf.(predicted) (GOp/s)(%)	(Power) (W)	Efficiency (GOp/s/W)	Precision (%)
FINN-*R* Results								
FINN-*R* MLP-4	AWS F1 (DF)	205.3	1,612	325,722	50TOPS	-	-	W^1A^1
FINN-*R* MLP-4	ZZSoC (DF)	300	417	38,205	5,110 (75%)	11.8	433	W^1A^1
FINN-*R* MLP-4	PYNQ-Z1 (DF)	100	224	30,249	974 (99%)	2.5	390	W^1A^1
FINN-*R* CNV-6	AWS F1 (DF)	234	2,380	345,557	12,021 (95%)	-	-	W^1A^1
FINN-*R* CNV-6	ZZSoC (DF)	300	283	41,733	2,318 (99%)	10.7	217	W^1A^1
FINN-*R* CNV-6	PYNQ-Z1 (DF)	100	280	30,605	341 (99%)	2.25	152	W^1A^1
FINN-*R* Tincy-Yolo	AWS F1 (DF)	190.7	2,712	205,537	4,023 (93%)	-	-	W^1A^3
FINN-*R* Tincy-Yolo	ZZSoC (MO)	220	316	40,808	146.2 (42%)	9.7	15	W^1A^3
FINN-*R* Tincy-Yolo	PYNQ-Z1 (MO)	100	280	46,507	60.1 (36%)	2.5	24	W^1A^3
FINN-*R* DoReFa-Net/PF	AWS F1 (DF)	109.6	2,160	421,255	8,540 (92%)	-	-	W^1A^2
FINN-*R* DoReFa-Net/PF	ZZSoC (MO)	220	432	36,249	75.68 (88%)	10.2	4	W^1A^2
FINN-*R* DoReFa-Net/PF	PYNQ-Z1 (MO)	100	278	35,657	33.2 (73%)	2.5	8	W^1A^2

ACM TRETS Special Edition on DL: FINN-R

From Embedded to Cloud

Summary

XILINX®

Summary

- **Trend towards big data and the need for HPC and ML faces technology challenges.**
- **This spurns a diversification of increasingly heterogeneous compute architectures**
  - Fueled by cloud economics
- **Needs to be addressed through architectural innovation**
- **Unconventional computing architectures emerge, in particular exploiting FPGAs**
- **These can bring performance scaling and energy efficiency**

Challenges

- **Programming complex systems**
- **Integrating diversity of DSAs within the cluster context**
- **Benchmarking heterogeneous systems for specific applications**
  - That are fundamentally differently programmed
  - That exploit different points within the design space

THANK YOU!

Repositories:
https://github.com/Xilinx/BNN-PYNQ
https://github.com/Xilinx/QNN-MO-PYNQ
https://github.com/Xilinx/FINN
https://github.com/Xilinx/LSTM-PYNQ

Adaptable.
Intelligent.

Publications:

FPGA 2017: FINN: A Framework for Fast, Scalable Binarized Neural Network Inference
https://arxiv.org/abs/1612.07119

PARMA-DITAM 2017: Scaling Binarized Neural Networks on Reconfigurable Logic
https://arxiv.org/abs/1701.03400

ICCD 2017: Scaling Neural Network Performance through Customized Hardware Architectures on Reconfigurable Logic
https://ieeexplore.ieee.org/abstract/document/8119246/

H2RC 2016: A C++ Library for Rapid Exploration of Binary Neural Networks on Reconfigurable Logic
https://h2rc.cse.sc.edu/2016/papers/paper_25.pdf

ICONIP’2017: Compressing Low Precision Deep Neural Networks Using Sparsity-Induced Regularization in Ternary Networks
https://arxiv.org/abs/1709.06262

CVPR’2018: SYQ: Learning Symmetric Quantization For Efficient Deep Neural Networks

DATE 2018: Inference of quantized neural networks on heterogeneous all-programmable devices
https://ieeexplore.ieee.org/abstract/document/8342121/

ARC’2018: Accuracy Throughput Tradeoffs for Reduced Precision Neural Networks

XILINX