Unconventional Computing Architectures with Reconfigurable Devices in the Cloud

Michaela Blott
Distinguished Engineer
Jan. 2019

Lucian Petrica, Giulio Gambardella, Alessandro Pappalardo,
Ken O'Brien, me, Nick Fraser, Yaman Umuroglu (from left to right)

Agenda

- Background
- Industry Context
- Unconventional Computing Architectures

Background

XILINX

Xilinx Research - Ireland

Ivo Bolsens
CTO

- Part of the worldwide CTO organization (9 out of 36)
- Including Xilinx University Program (Cathal, Katie)
- AI Lab expansion part-financed through IDA Ireland
- Mission: Application driven technology development

Kees Vissers
Fellow

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

NTNU
Norwegian University of Science and Technology

KIT
Karlsruher Institut für Technologie

Politecnico
Milano 1863

University of Cambridge

Technische Universität Kaiserslautern

ETH Zürich

Technische Universität Dresden

The University of Sydney

Imperial College London

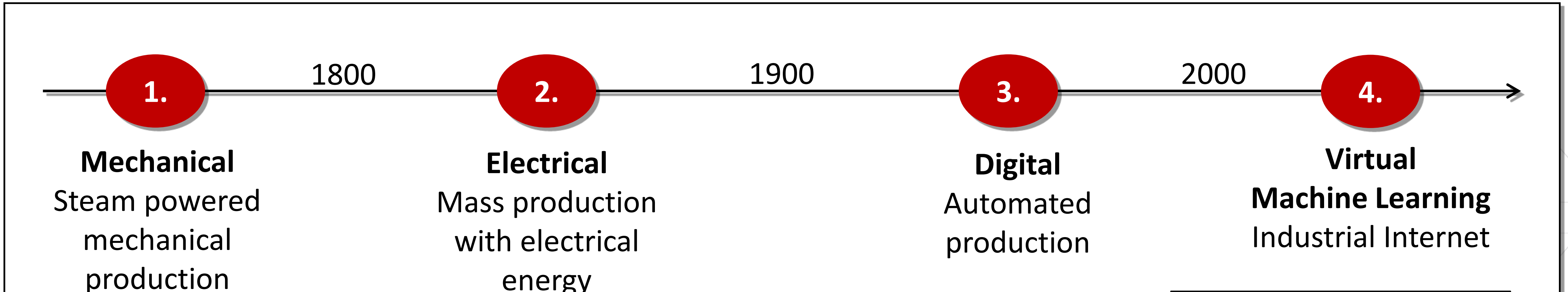
Industry Context

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
WATERSON

“Trends meeting Technological Reality”

XILINX®

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

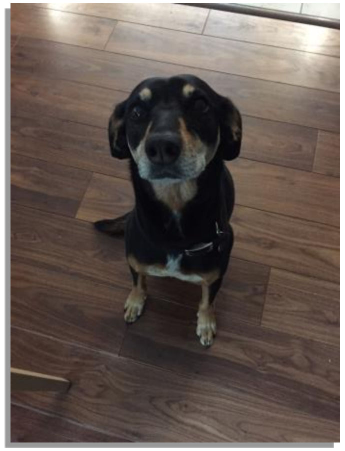
3. **Digital**
Automated production

4. **Virtual**
Machine Learning
Industrial Internet

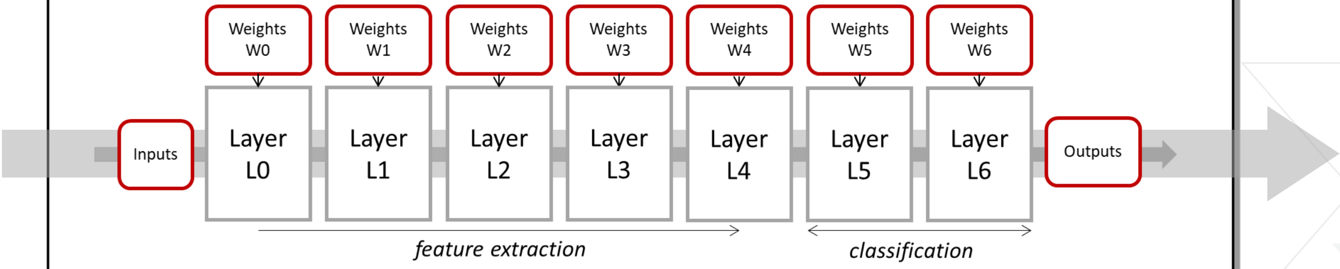
- **Potential to solve the unsolved problems**
  - Making solar energy economical, reverse engineering the brain (Jeff Dean, Google Brain 2017)

What's the Challenge?
Example: Convolutional Neural Networks
Forward Pass (Inference)

Input Image



Neural Network



Neural Network

Cat?

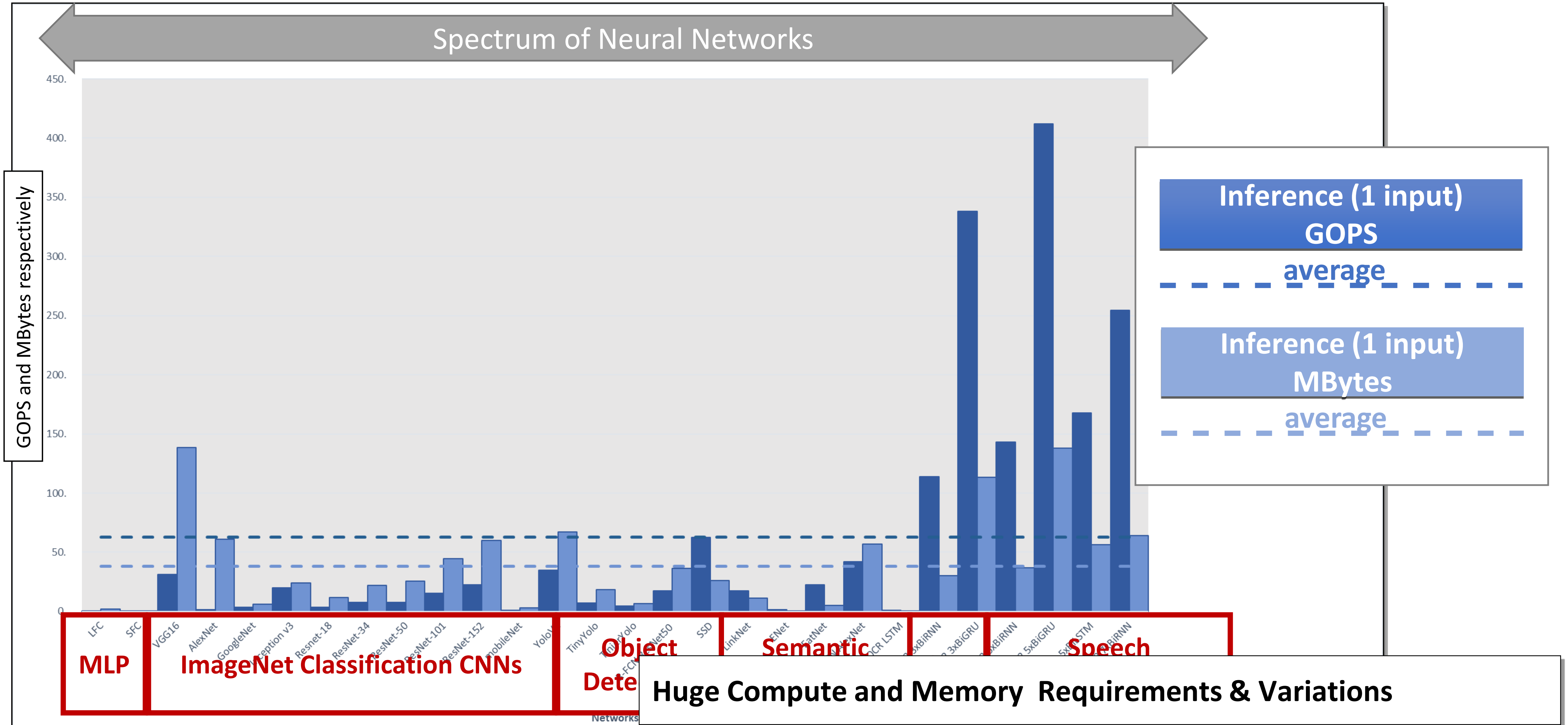
For ResNet50:

70 Layers
7.7 Billion operations
25.5 millions of weight

Basic arithmetic, incredible parallel but Huge Compute and Memory Requirements

Compute and Memory for Inference

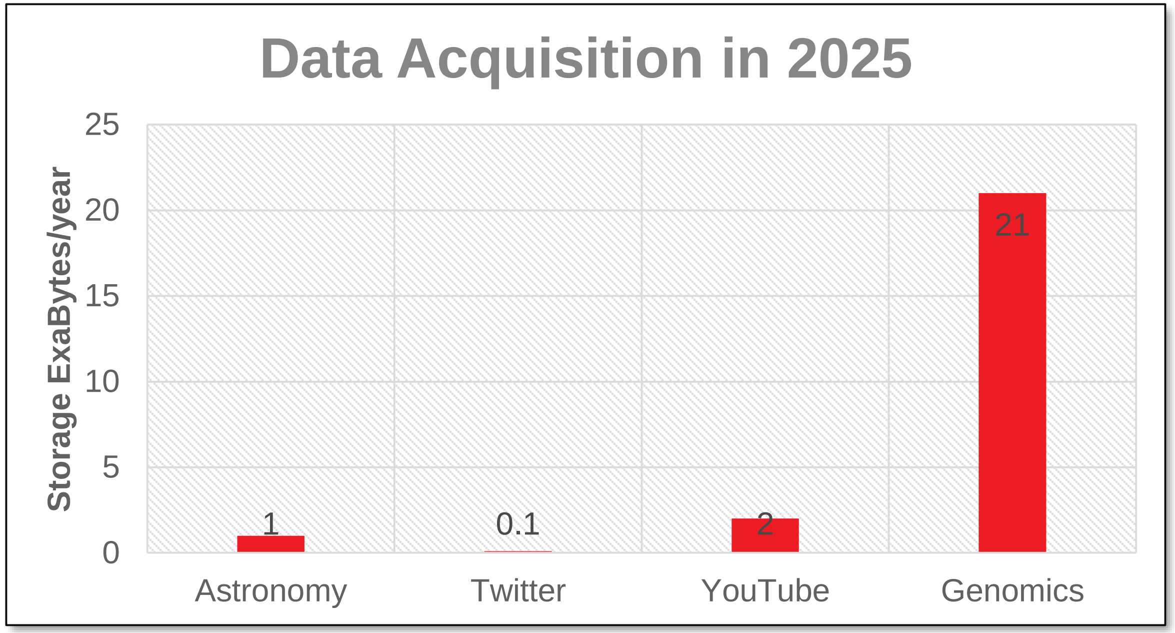
*architecture independent
**1 image forward
*** batch = 1
**** int8



Huge Compute and Memory Requirements & Variations

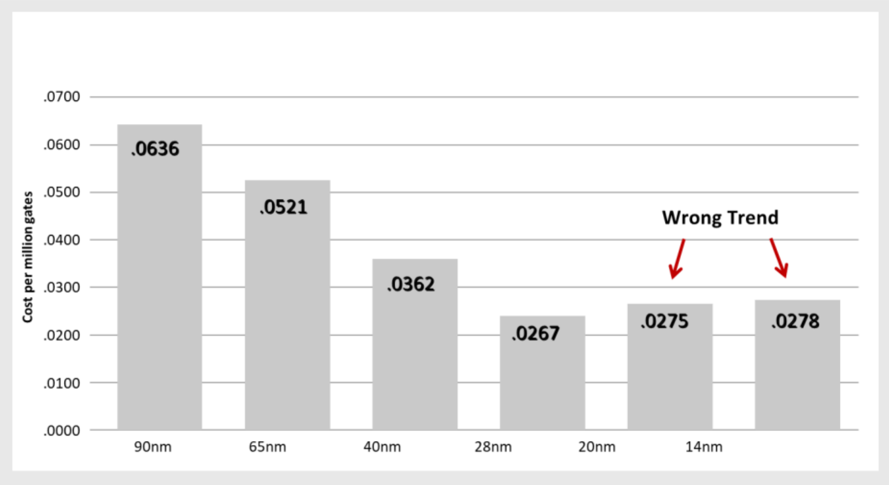
Mega-Trend:
Explosion of Data

- **Astronomically growing amounts of data**
  - More sensors
  - More users
  - More use cases: Genomics (DNA) **“Genomical”**

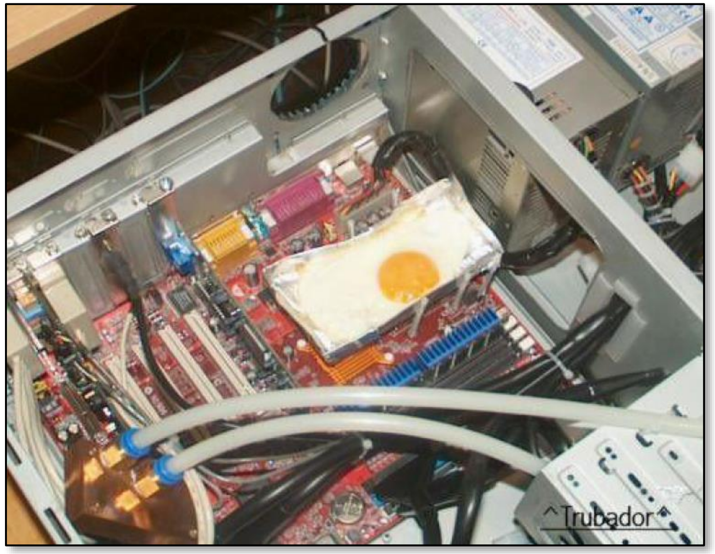


Stephens, Zachary D., et al.
"Big data: astronomical or genomical?."

Technology:
End of Moore’s Law & Dennard Scaling



Source: IBS

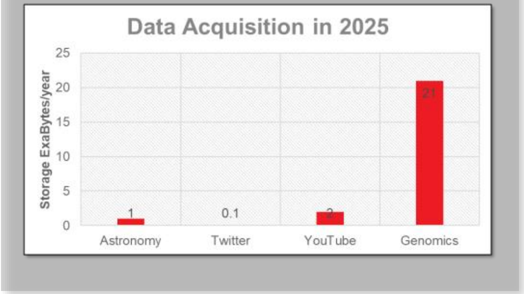


Economics become questionable

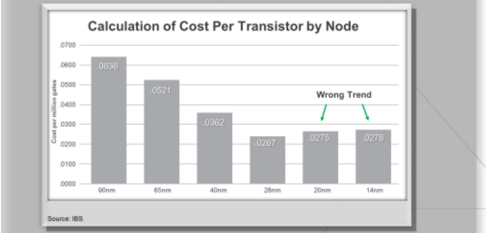
Power dissipation becomes problematic

Era of Heterogeneous Compute using Accelerators

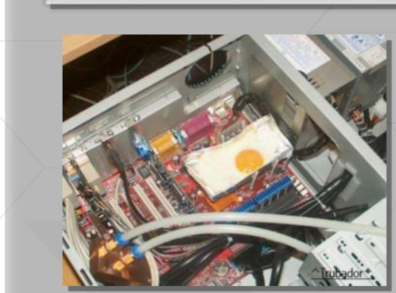
Trends



Technology

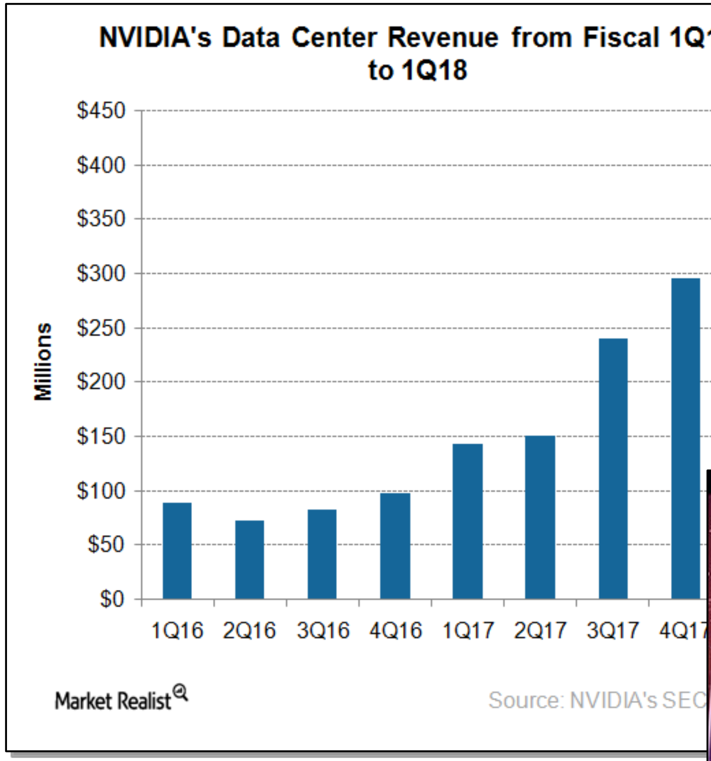


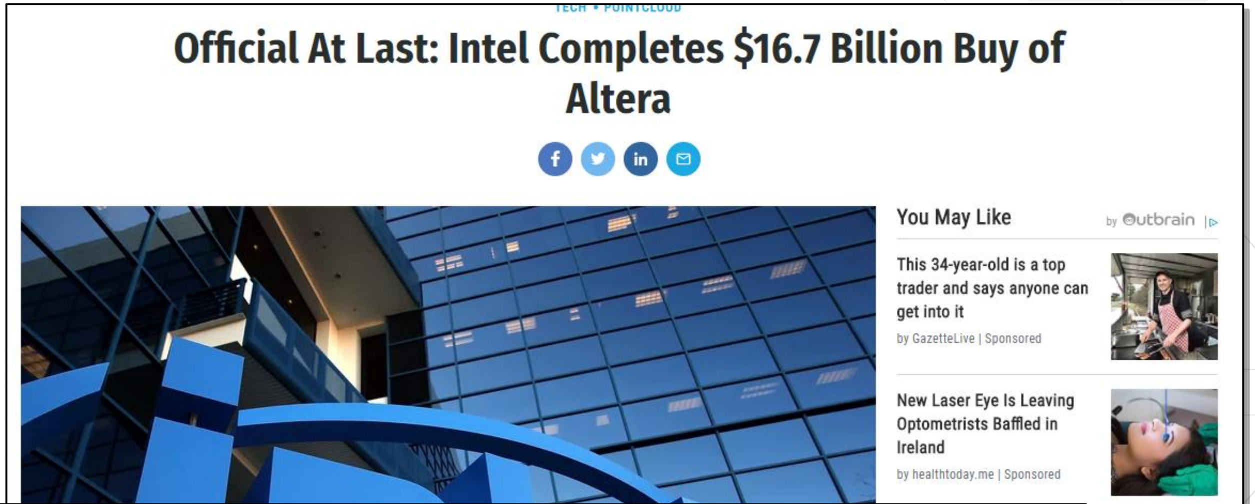
Source: IBS



- Diversification of increasingly heterogenous devices and system
  - Moving away from standard van Neumann architectures
- True Architectural innovation & Unconventional Computing Systems

Evidence: Heterogenous Data Centers





TECH • POINTCLOUD

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

You May Like

by Outbrain

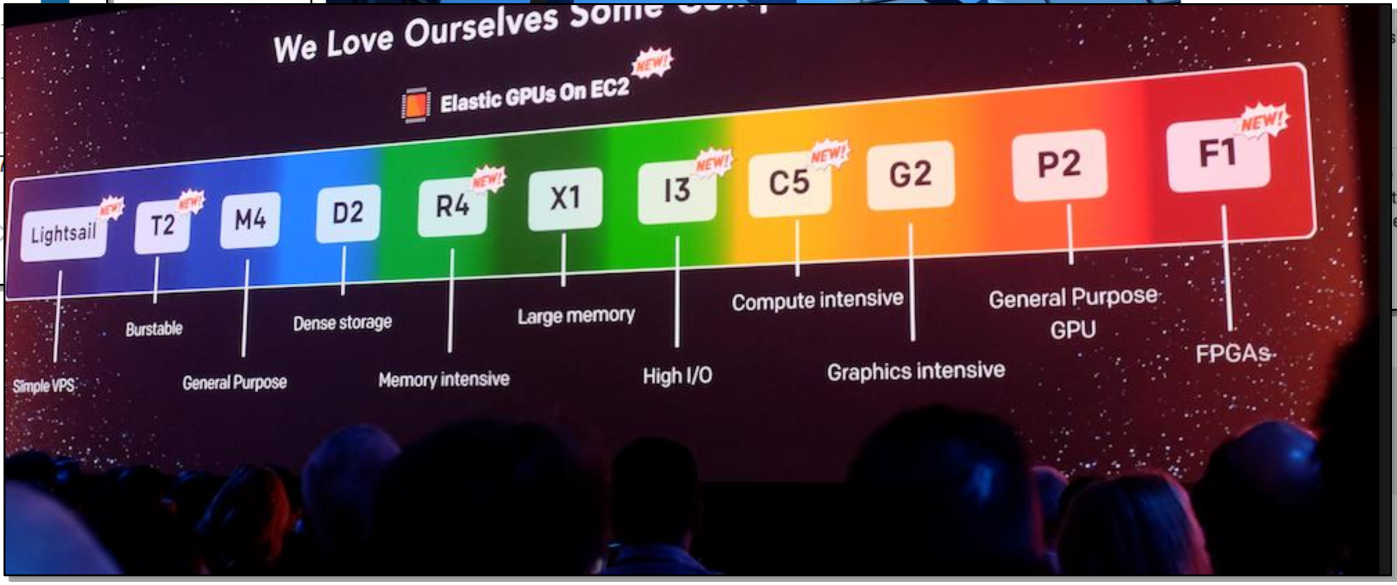
This 34-year-old is a top trader and says anyone can get into it

by GazetteLive | Sponsored

New Laser Eye Is Leaving Optometrists Baffled in Ireland

by healthtoday.me | Sponsored

Ideal Media



Insight 2016: AWS adding FPGA instances

Unconventional at System Level: Diversification with Accelerator Support

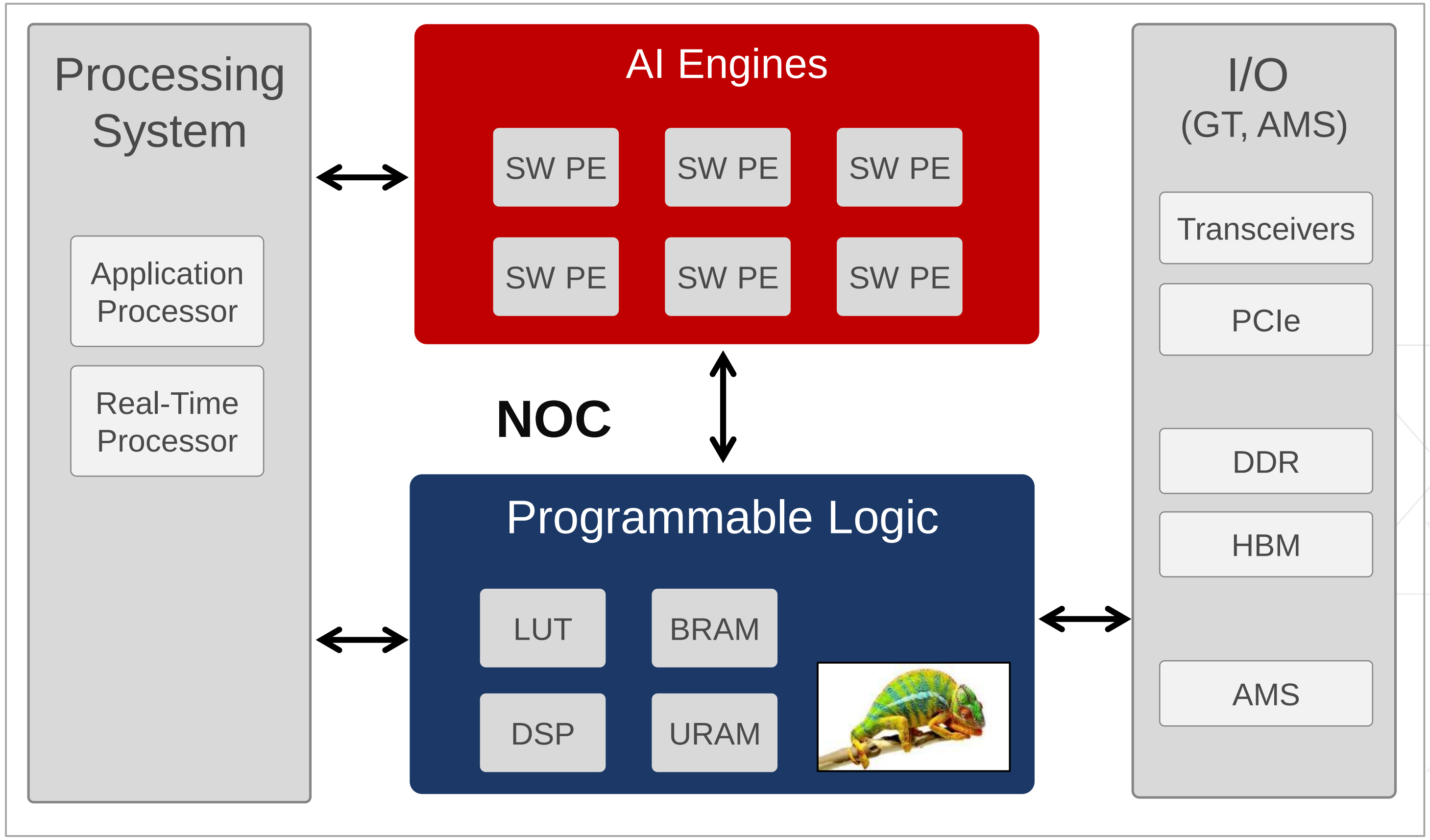
HP Moonshot

OpenPOWER™

IBM's OpenPower

> With accelerators moving closer to the CPU (OpenCAPI, CCIX, etc…)

Evidence: Heterogeneous Devices



> From the Xilinx World: Evolution of FPGAs to **ACAPs**

With reconfigurable computing, we can go even more unconventional:

some examples

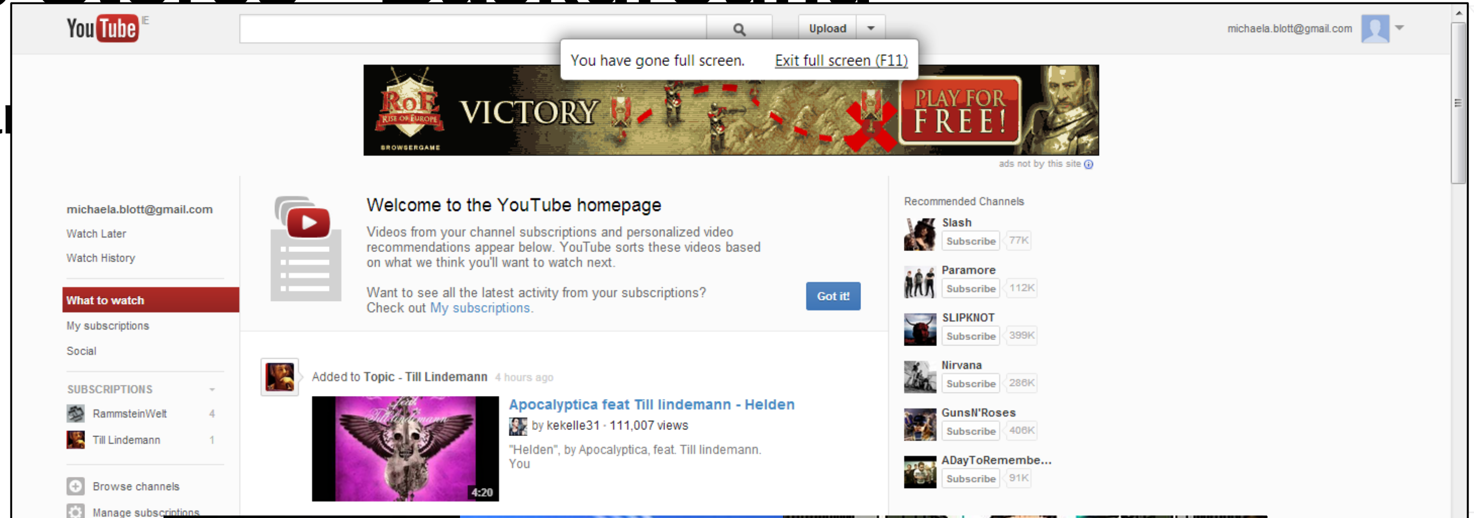
XILINX

Key-Value Stores
- customized data paths
- customized memory subsystem

XILINX

Key Value Stores - Background

- Many popular



You Tube IE

Upload

michaela.blott@gmail.com

You have gone full screen. Exit full screen (F11)

michaela.blott@gmail.com
Watch Later
Watch History
What to watch
My subscriptions
Social

SUBSCRIPTIONS

RammsteinWelt	4
Till Lindemann	1

Browse channels
Manage subscriptions

Welcome to the YouTube homepage

Videos from your channel subscriptions and personalized video recommendations appear below. YouTube sorts these videos based on what we think you'll want to watch next.

Want to see all the latest activity from your subscriptions? Check out My subscriptions.

Got it!

Added to Topic - Till Lindemann 4 hours ago

Apocalyptica feat Till lindemann - Helden
by kekelle31 · 111,007 views
"Helden", by Apocalyptica, feat. Till lindemann. You

4:20

Recommended Channels

Slash	Subscribe	77K
Paramore	Subscribe	112K
SLIPKNOT	Subscribe	399K
Nirvana	Subscribe	286K
GunsN'Roses	Subscribe	406K
ADayToRemembe...	Subscribe	91K

Send feedback

FROM THE DIRECTOR OF TERMINATOR 2 AND TITANIC
AVATAR
007
GANGNAM STYLE
INDIANA JONES and the RAIDERS of the LOST ARK
STAR WARS

Only sto
recent
records

ool of x86-
vers with
M running

Up to 30% o

WIKIPEDIA
The Free Encyclopedia

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,…)

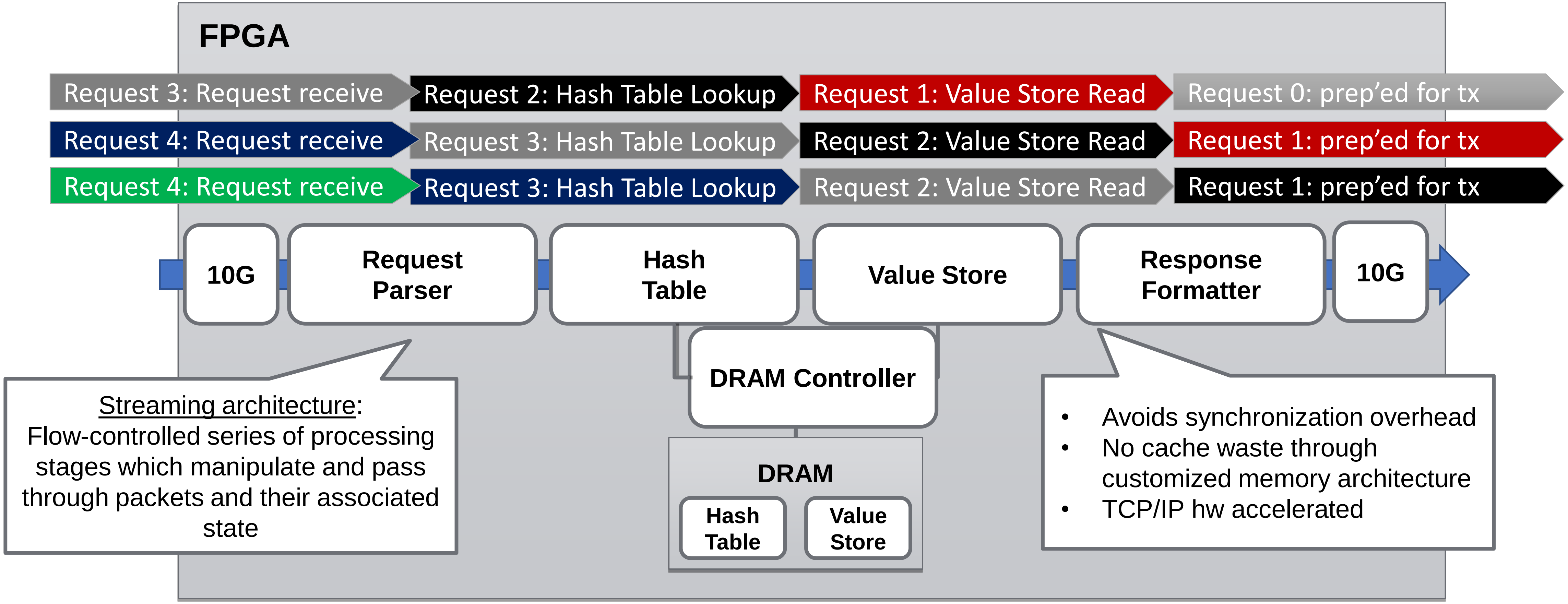
- **Bottlenecked by:**
  - Synchronization overhead
    - Threads stall on memory locks, serializing execution for x86s
  - TCP/IP is CPU intensive, interrupt intensive, too large to fit into instruction cache
  - Last level cache ineffective due to random-access nature of the application (miss rate 60% - 95% on x86)

- **Performance significantly below 10Gbps line rate**
    - Intel Xeon (8cores): 1.34MRps, 200-300usec, 7KRPS/Watt

```
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:
      …
```

Dataflow Architectures to Scale Performance



Request 3: Request receive	Request 2: Hash Table Lookup	Request 1: Value Store Read	Request 0: prep'ed for tx
Request 4: Request receive	Request 3: Hash Table Lookup	Request 2: Value Store Read	Request 1: prep'ed for tx
Request 4: Request receive	Request 3: Hash Table Lookup	Request 2: Value Store Read	Request 1: prep'ed for tx

<u>Streaming architecture</u>: Flow-controlled series of processing stages which manipulate and pass through packets and their associated state

- Avoids synchronization overhead
- No cache waste through customized memory architecture
- TCP/IP hw accelerated

> Order of magnitude improvement in latency and best in class for jitter

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

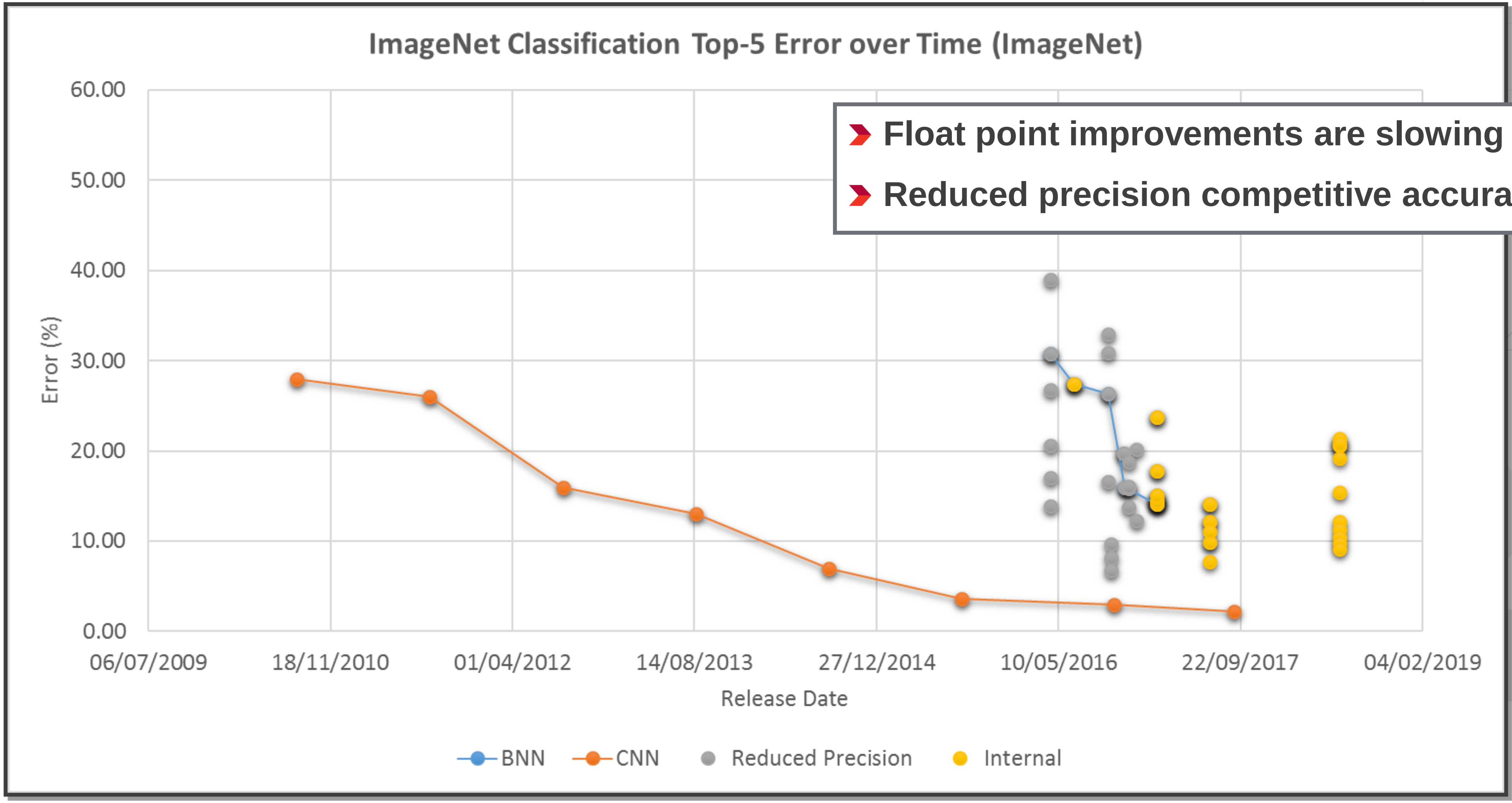
Source: [4] Blott et al: Achieving 10Gbps line-rate key-value stores with FPGAs; HotCloud 2013

Algorithms in Logic ALGO-LOGIC® http://algo-logic.com

Deep Learning
- customized precision arithmetic

XILINX

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



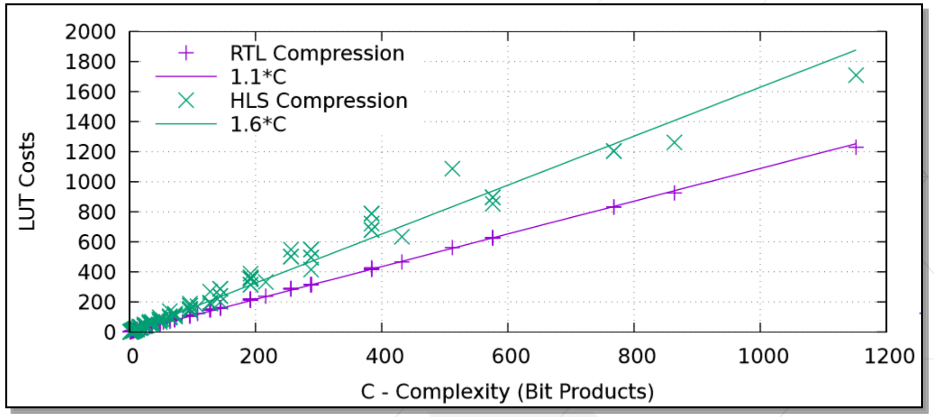
- Float point improvements are slowing down
- Reduced precision competitive accuracy

Reducing Precision

Scales Performance & Reduces Memory

- **Reducing precision shrinks LUT cost**
  - Instantiate **100x** more compute within the same fabric
- **Potential to reduce memory footprint**
  - NN model can stay on-chip => no memory bottlenecks

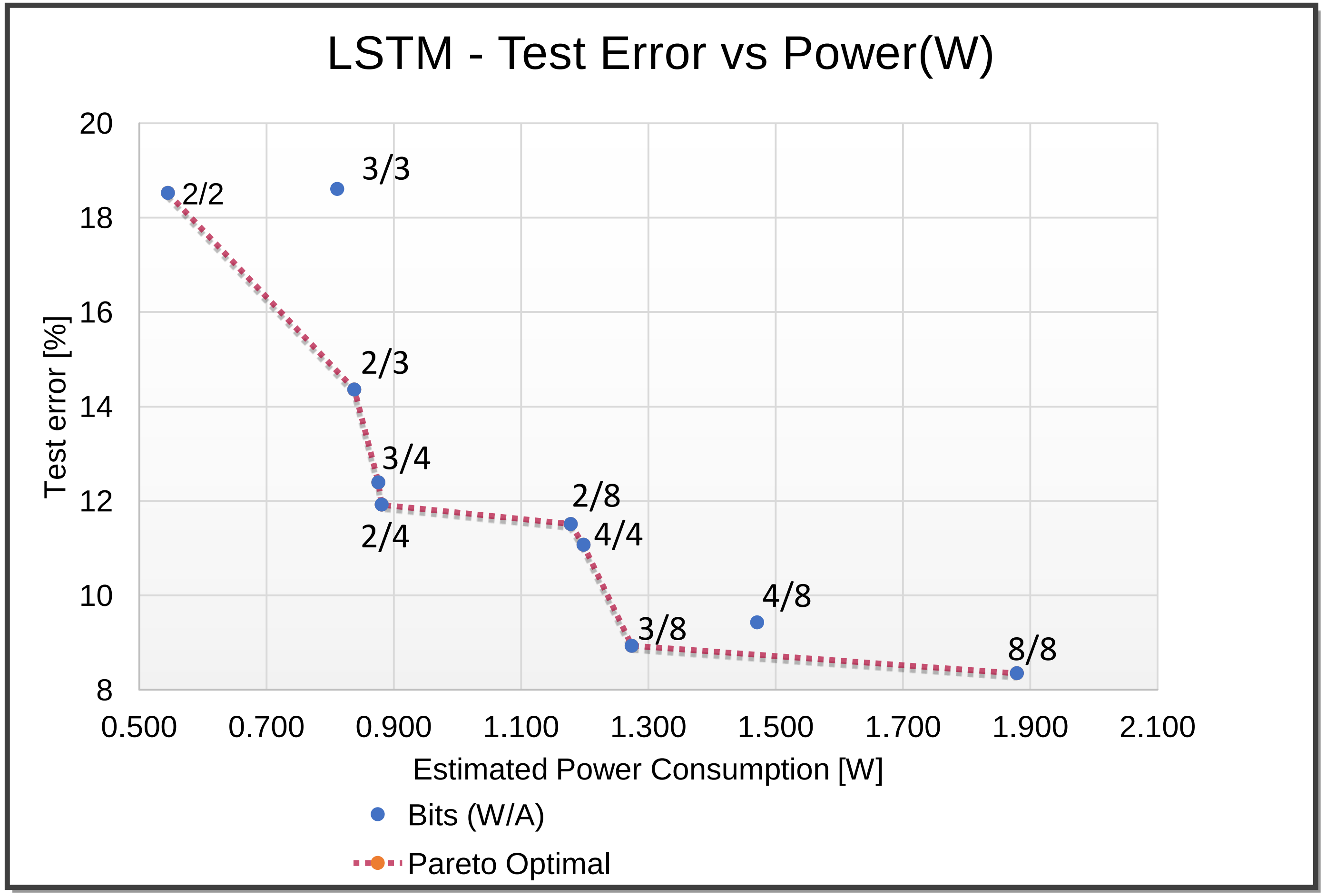
Precision	Modelsize [MB] (ResNet50)
1b	3.2
8b	25.5
32b	102.5



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

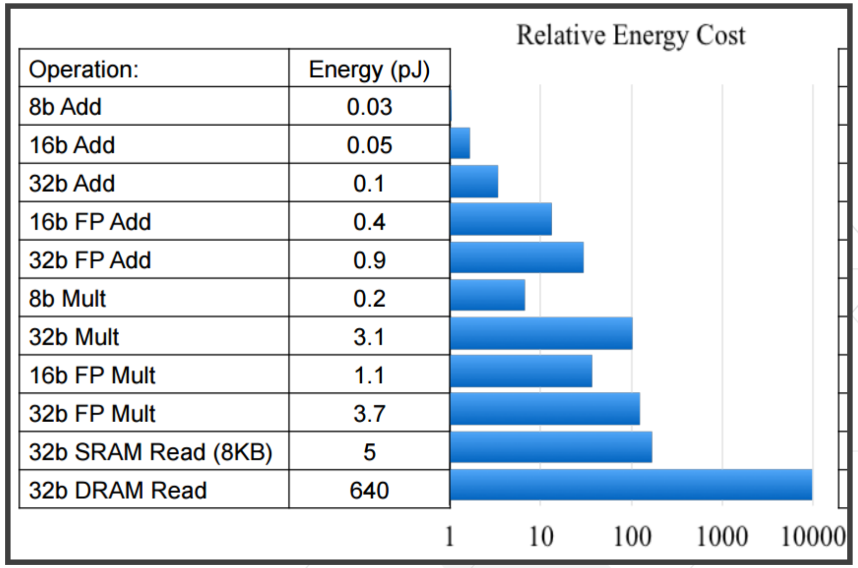
Reducing Precision Inherently Saves Power

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:


Relative Energy Cost
1
10
100
1000
10000

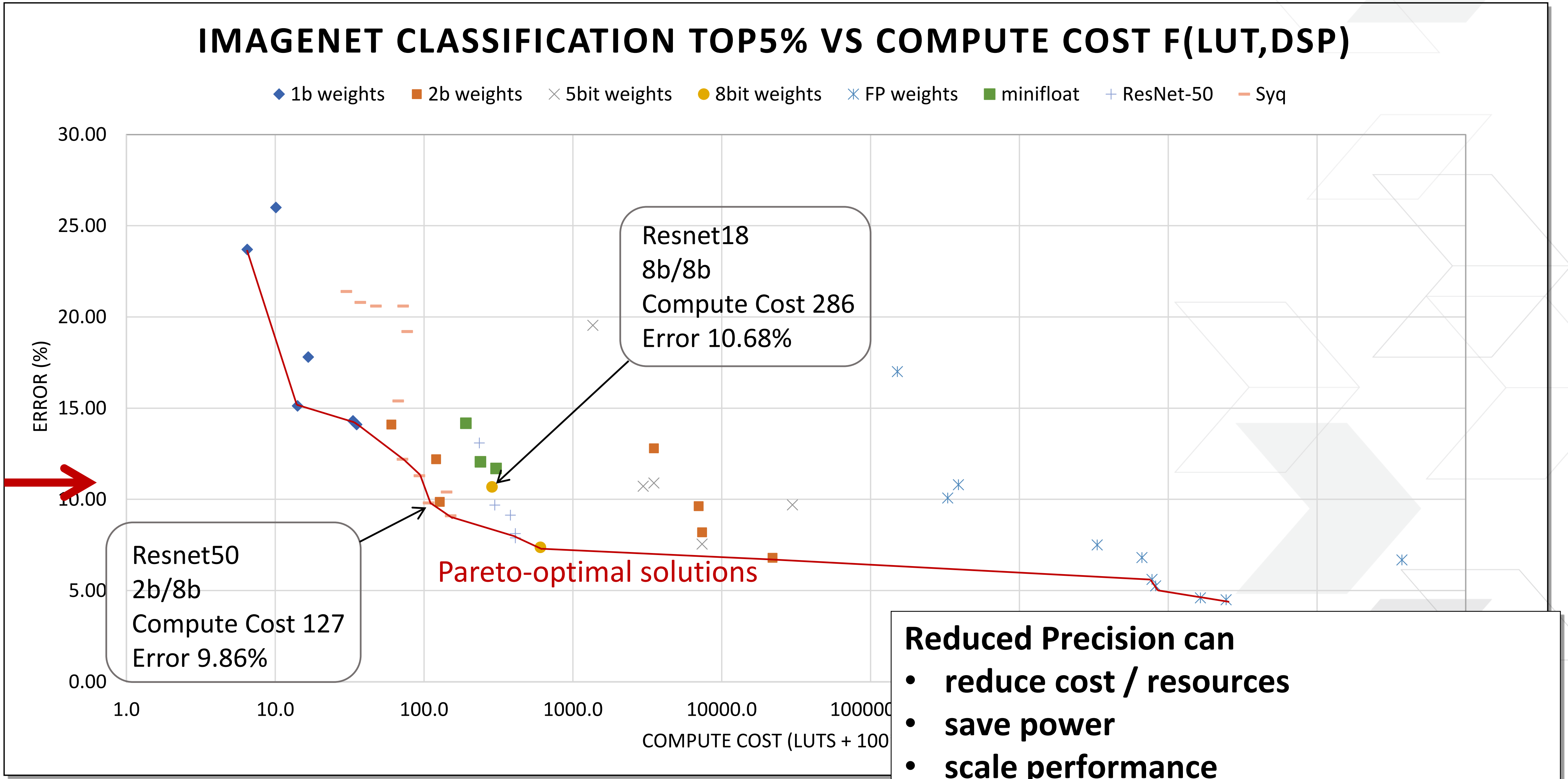
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

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."

Design Space Trade-Offs



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

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

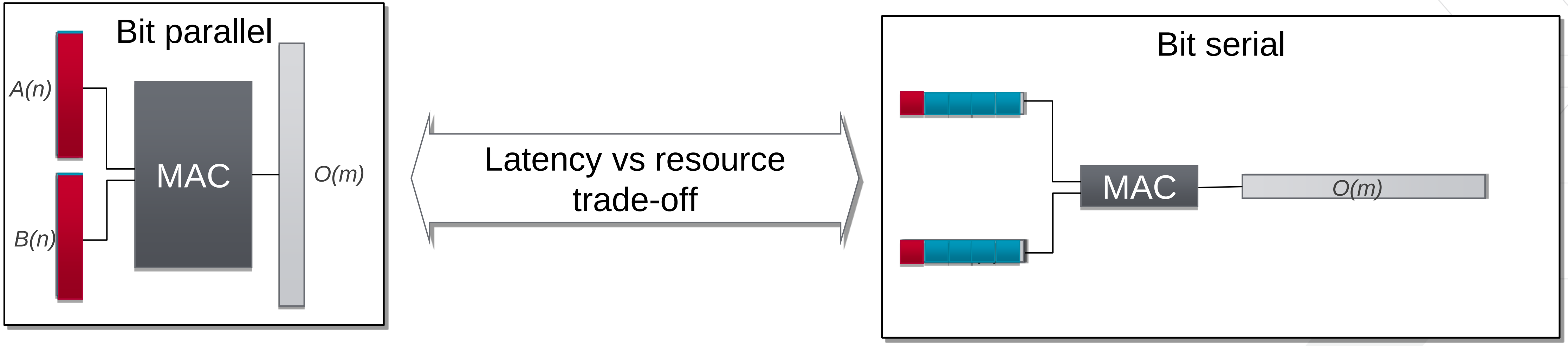
Pareto-optimal solutions

Reduced Precision can

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

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

Summary

Summary

- **Unconventional computing architectures emerge at data center, system and device level**
- **With reconfigurable computing we can go even more unconventional**
- **Leveraging customized dataflow architectures and memory subsystems, custom precisions**
  - To provide dramatic performance scaling and energy efficiency benefits
  - To enable new exciting trade-offs within the design space

Challenges in Futures

- **Programming unconventional systems**
- **Benchmarking heterogeneous systems for specific applications**
  - That are fundamentally differently programmed
  - That exploit different points within the design space
- **How can you apply some of these concepts to other applications?**

THANK YOU!

Adaptable.
Intelligent.

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

XILINX