XILINX
ALL PROGRAMMABLE™

Machine Learning for Embedded Systems

Michaela Blott, Principal Engineer, Xilinx Research

- Background: Xilinx & Xilinx Research & Challenges in the Industry
- Machine Learning and its Challenges
- Xilinx Effort
- Summary & Questions

Introduction to Xilinx

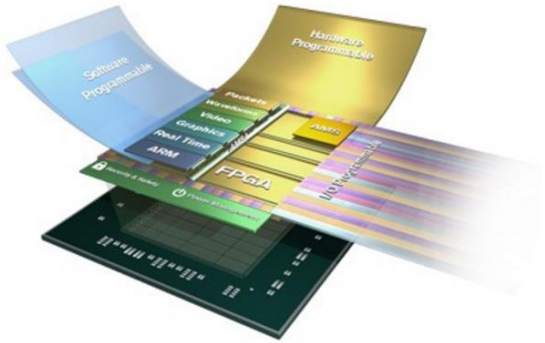
- **Fabless semiconductor company**
- **Founded in Silicon Valley in 1984**
- **Today: Approximately 3,500 employees and $2.25B revenue**
- **Invented the FPGA**

1st FPGA in 1985: XC2064

128 3-input LUTs

30 years

Ultrascale +: VU13P

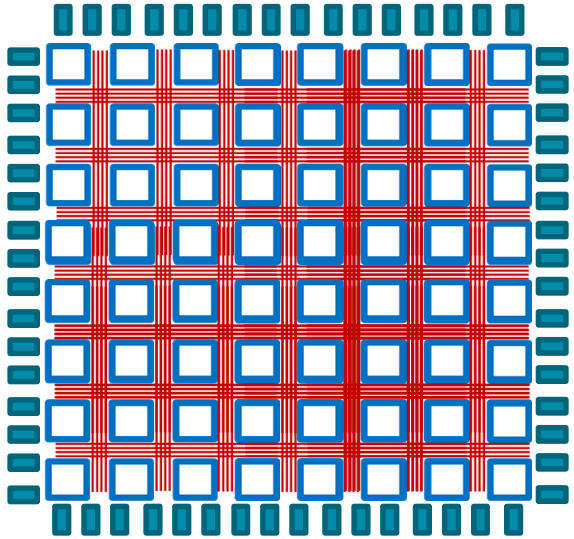


1.7M 6-input LUTs

What are FPGAs?

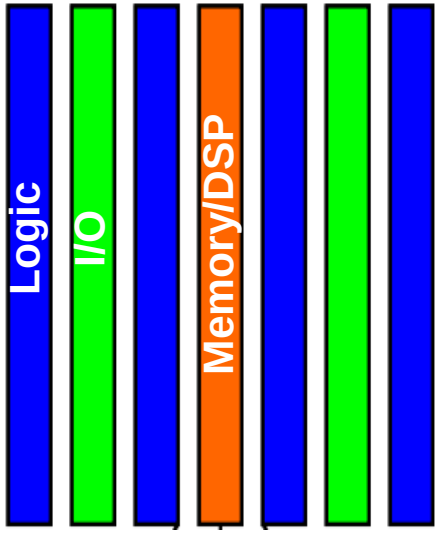
Customizable, Programmable Hardware Architectures

FPGA Technology over Time



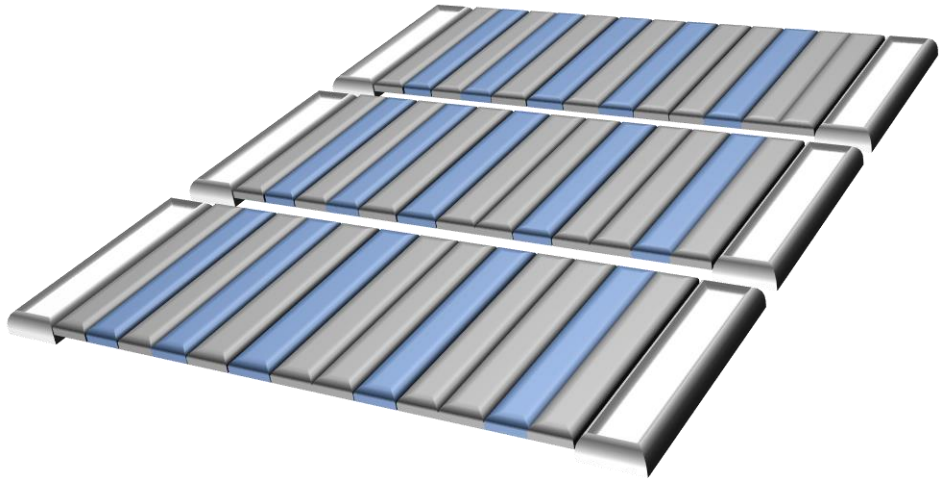
LUTs

Basic bit-oriented logic
4-input, 6-input Lookup table



Columns

Basic bit-oriented logic +
Word-oriented Multiply-accumulate
Word-oriented Memory



Die-stacked slices

Basic bit-oriented logic +
Word-oriented Multiply-accumulate
Word-oriented Memory
System integration, e.g. PCIe, DDR

Your program becomes a configuration that sets table values and switches via synthesis, Place and Route tools

Diversified Across Multiple Markets

Mars Rover

MRI scanners

ADAS

AR

3D televisions

Dish Washers

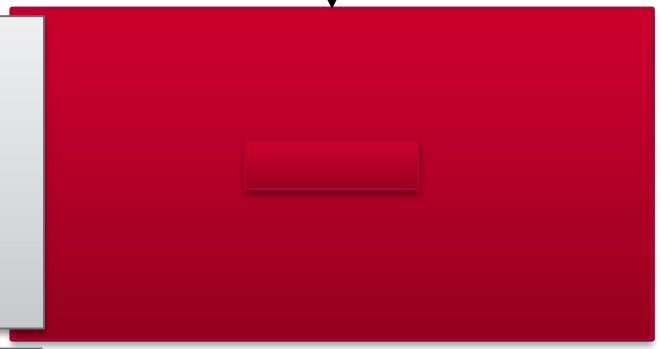
XILINX ALL PROGRAMMABLE™

The Potential: Customized Hardware Architectures

Task: X = (a + b) * 256;

No instruction fetch and decode, no data fetch

Customized arithmetic operations



1 cycle @ 400MHz

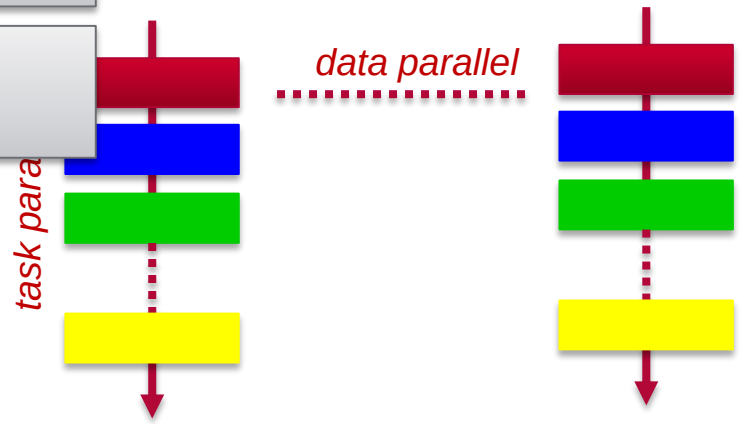
6 cycles @ 2GHz

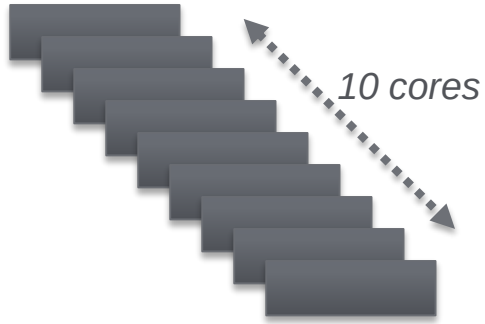
1 compute unit:
2 OP/2.5nsec
=0.8GOP/s

1 core:
2 OP/6*0.5nsec
=0.67GOP/S

10'000 compute units:
80TOP/s

28cores:
18.8GOP/s





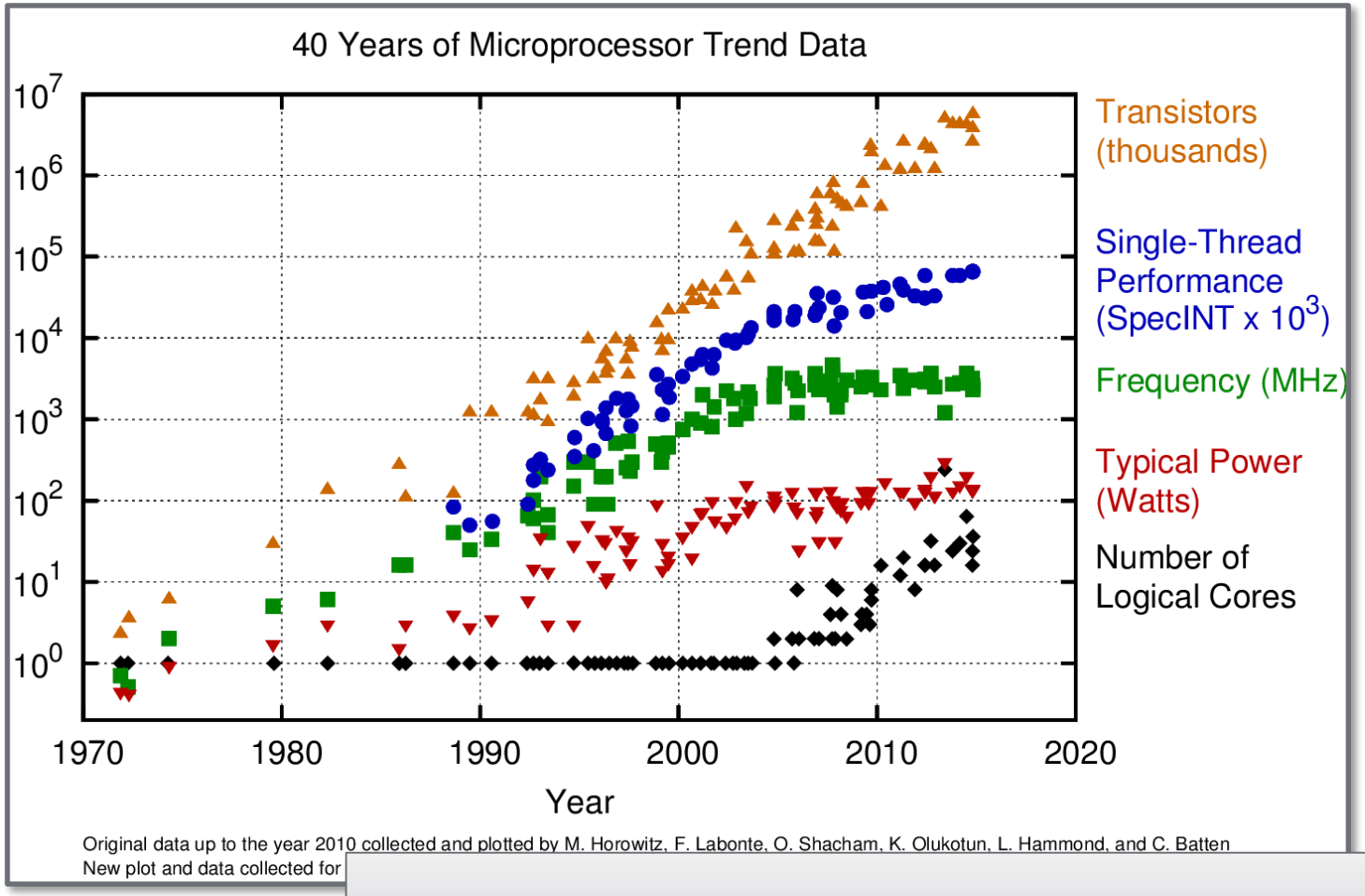
FPGA promise very high throughput, and power savings

Xilinx Research - Ireland

- **Part of the CTO organization**
  - 9 (out of 35 worldwide) researchers
- **With a very active internship program**
  - 6-10 students & visiting scholars
- **University Program**
- **IDA funding for build-out of AI lab**

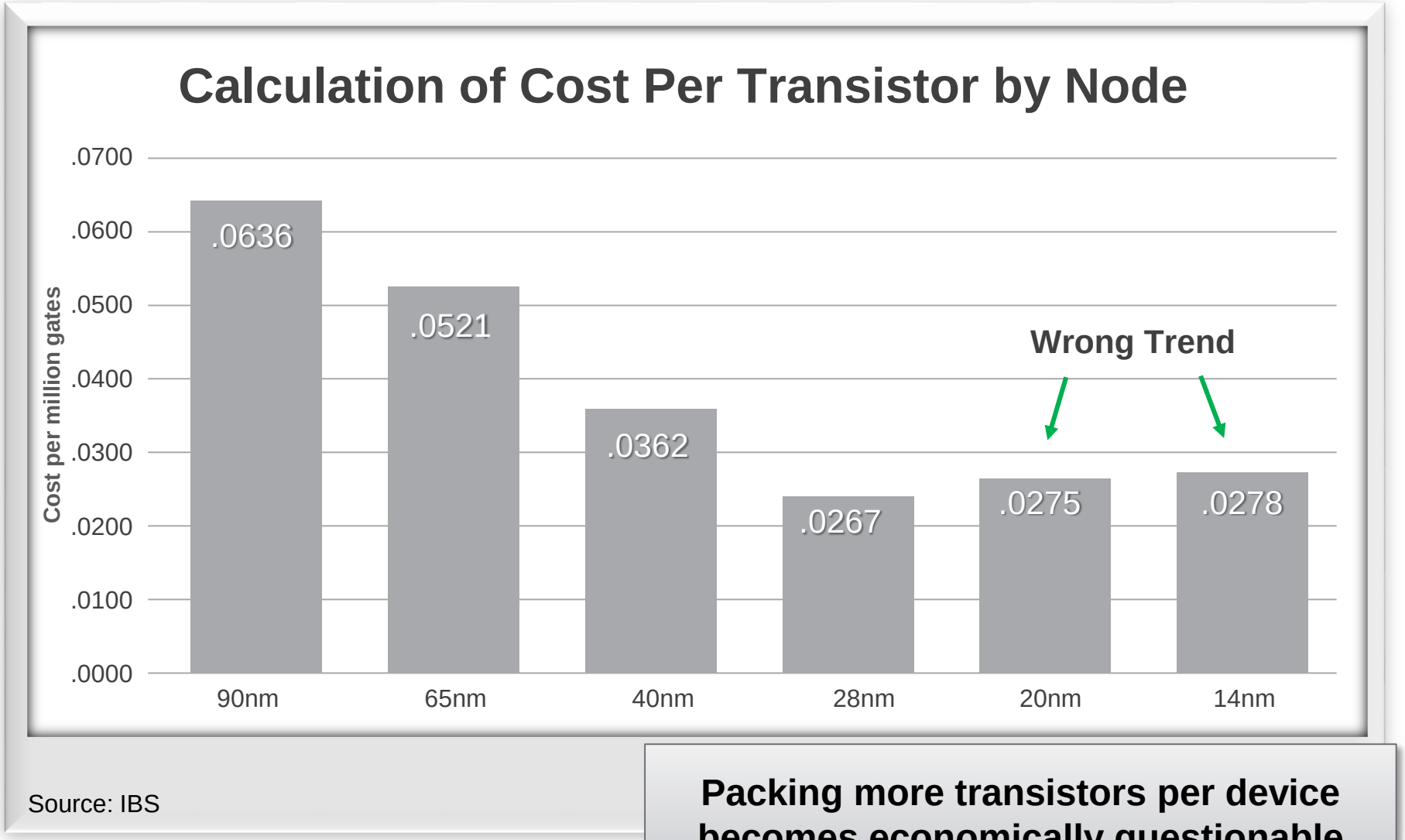
IDA Ireland

Clock scaling has slowed down since ~2005



Performance scalability is limited

End of Moore's Law: Transistor Cost Trend



Source: IBS

Packing more transistors per device becomes economically questionable

End of Dennard Scaling

Power dissipation with increased transistor density becomes problematic

Source: Intel

- Background: Xilinx & Xilinx Research
- Machine Learning and its Challenges
- Xilinx Effort
- Summary, Demo & Questions

New York Times: "The Great A.I. Awakening"

(Dec 2016)

Elon Musk's Billion-Dollar AI Plan Is About Far More Than Saving the World

The Race For AI: Google, Twitter, Intel, Apple In A Rush To Grab Artificial Intelligence Startups

World's Largest Hedge Fund to Replace Managers with an AI System

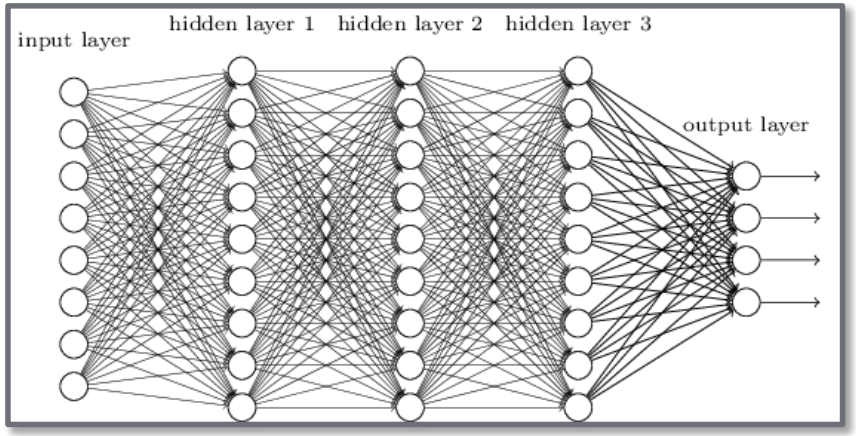
Drones Can Defeat Humans Using Artificial Intelligence

Elon Musk's leads 116 Experts on Open Letter to Ban Killer Robots

Demonstrated to work well for numerous use cases & is here to stay

Neural Networks

- **Based on simple models of the human brain (neurons and synapses)**
- **NNs have the theoretical property of being a "universal approximation function"**
  - Empirically outperforming other approximator functions
  - Increasing adoption for new use cases
- **Requires less expertise/specialization in the target domain**
- **NNs are the predominant algorithm used**
  - Outperforming humans and traditional CV algorithms for image recognition



Increasing adoption (replacing other solutions and for previously unsolved problems)

Challenge 1: Diverse Applications

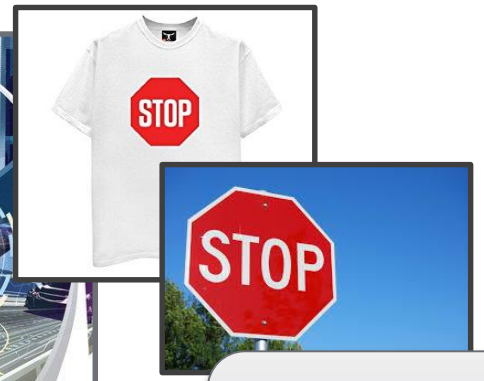
Translate & AlphaGo:
Huge networks

AlphaGo

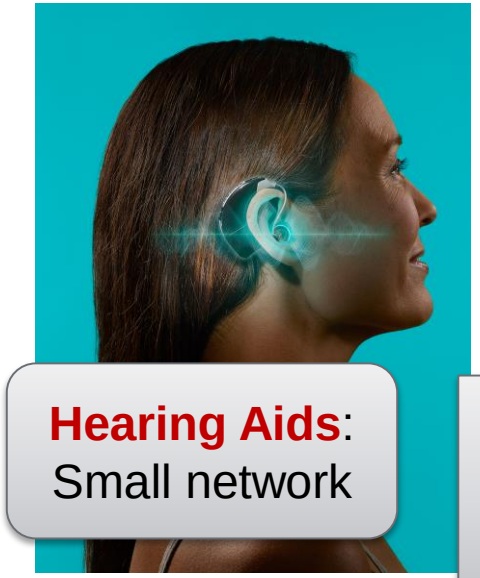
Medical Diagnosis:
Small networks

Robotics:
Real-time

Recommender Systems



ADAS



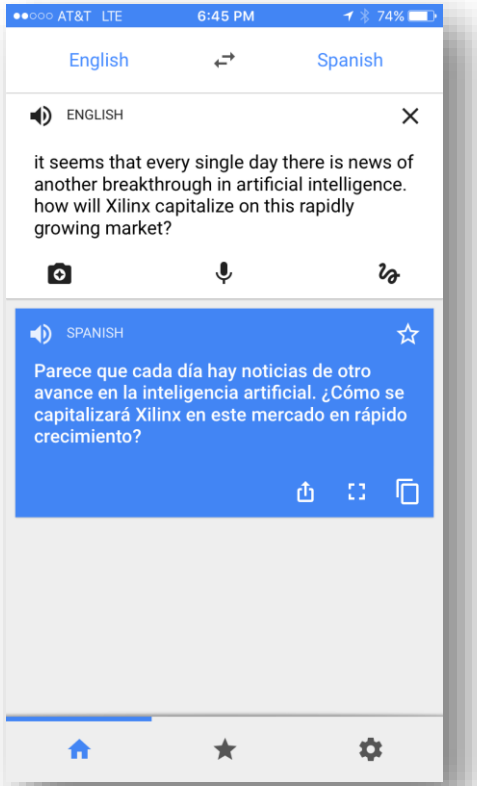
Hearing Aids:
Small network

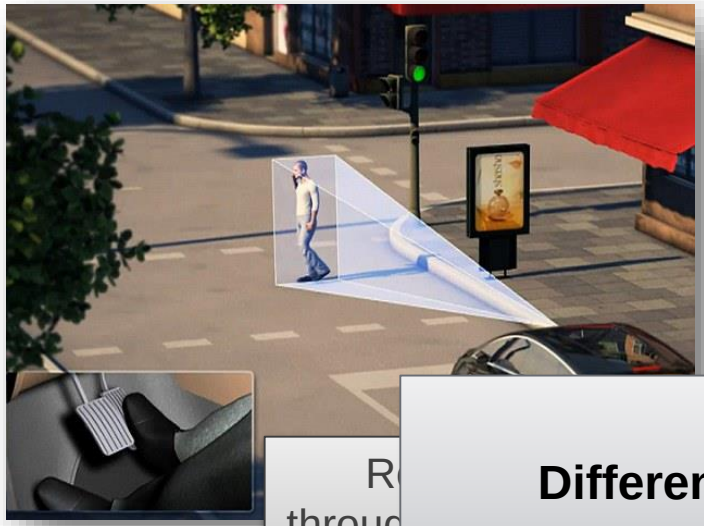
Challenge:
Different use cases require different networks

Challenge 2: Different Figures of Merits

Accuracy requirements vary with applications
Recommender systems, data analytics vs ADAS

Reduced latency: Results in a better user experience in cloud-based systems (Google defines 7ms) and vital for robotics





R
throug

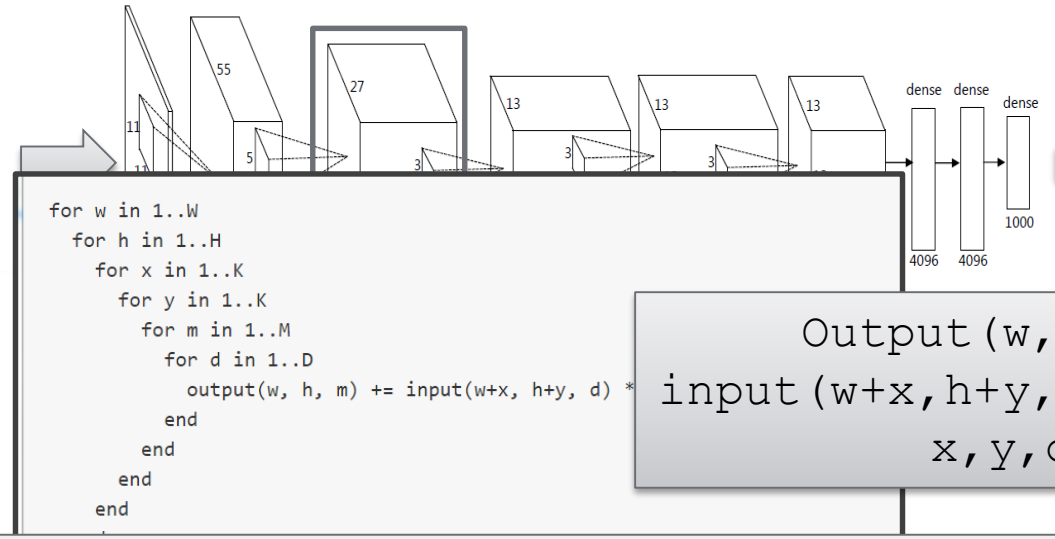
eavily power

Data centers: OPEX = f(energy)

Challenge:
Different figures of merits (power, throughput, latency, accuracy)

Challenge 3: Highly Compute and Memory Intensive

- **The predominant CNN computation is linear algebra**
  - Demands lots of (simple) computation and lots of parameters (memory)
    - AlexNet: 244MB & 1.5GOPS, VGG16: 552MB & 30.8GOPS; GoogleNet: 41.9MB & 3.0GOPS for ImageNet



«cat»

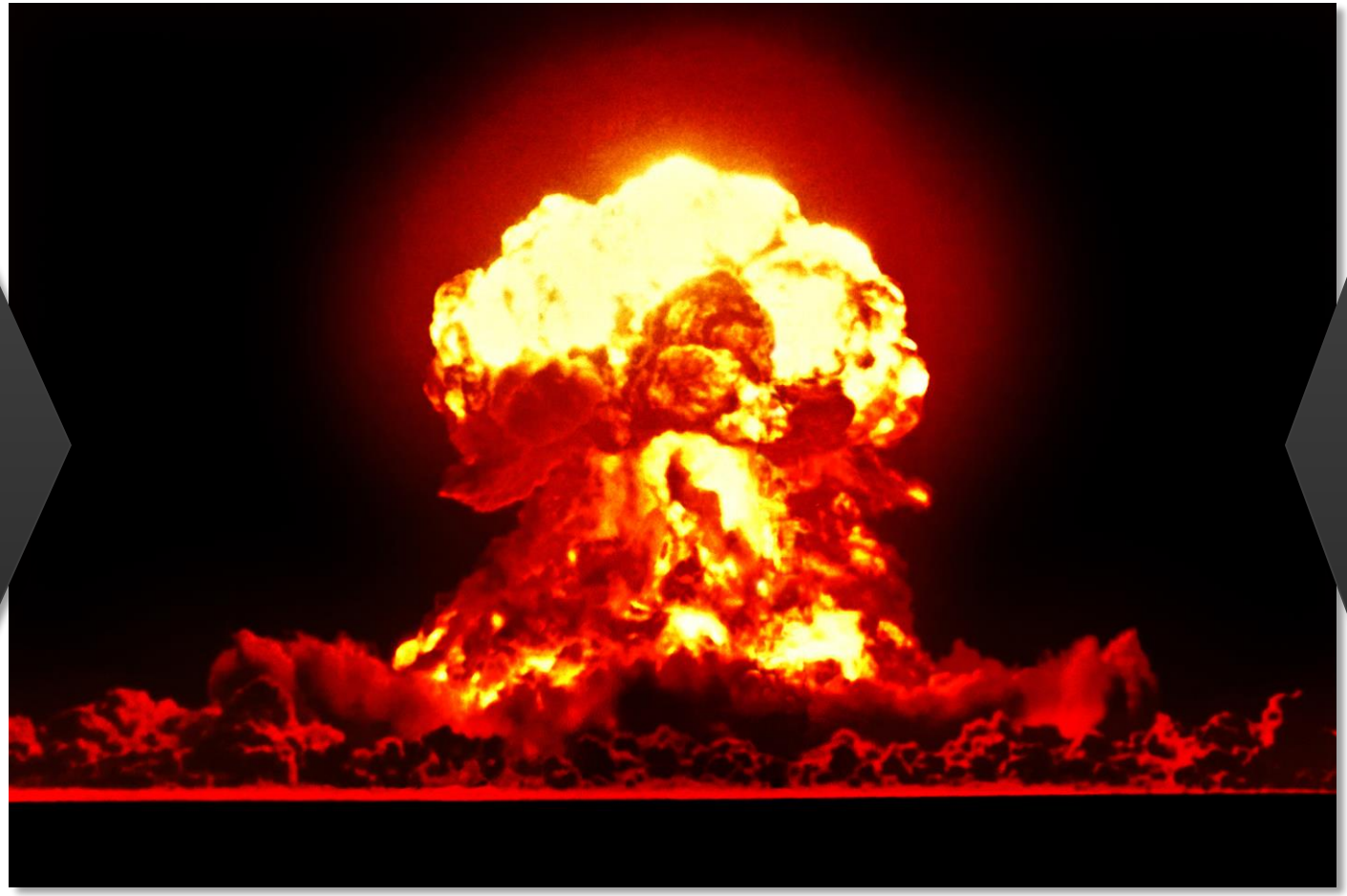
```
for w in 1..W
  for h in 1..H
    for x in 1..K
      for y in 1..K
        for m in 1..M
          for d in 1..D
            output(w, h, m) += input(w+x, h+y, d) *
          end
        end
      end
    end
```

```
Output(w,h,m) +=
input(w+x,h+y,d)*filter(m,
x,y,d);
```

Challenge:
billions of multiply-accumulate ops & tens of megabytes of parameter data

ML Challenges meet Semiconductor Industry

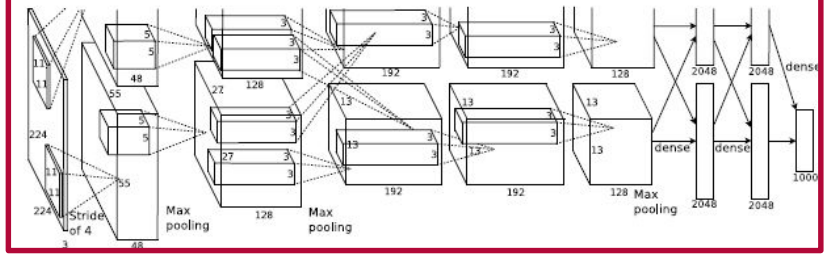
End Technology & Energy Scaling



ML compute & memory requirements

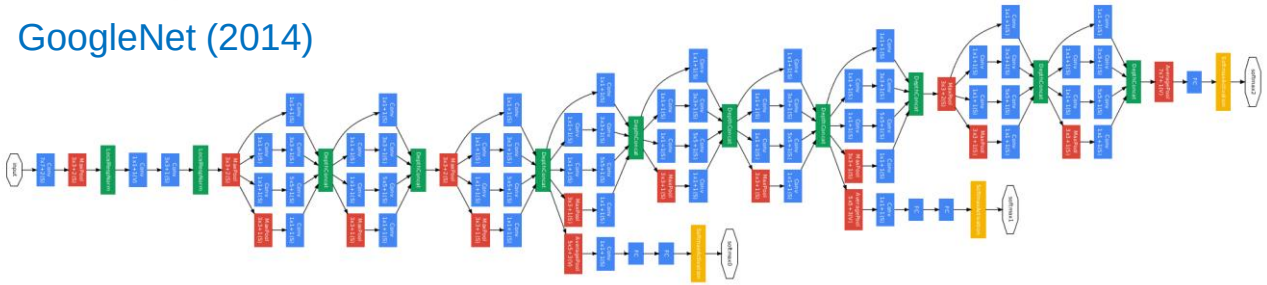
Challenge 4: Neural Networks Will Continue to Change

AlexNet (2012)

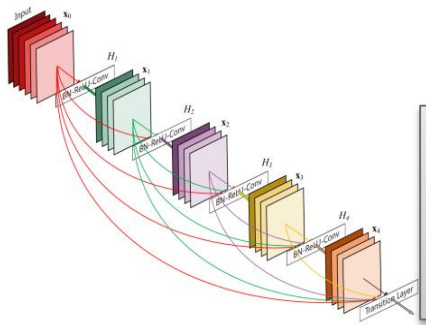


Number and **types** of layers are changing

GoogleNet (2014)



DenseNet (2016)



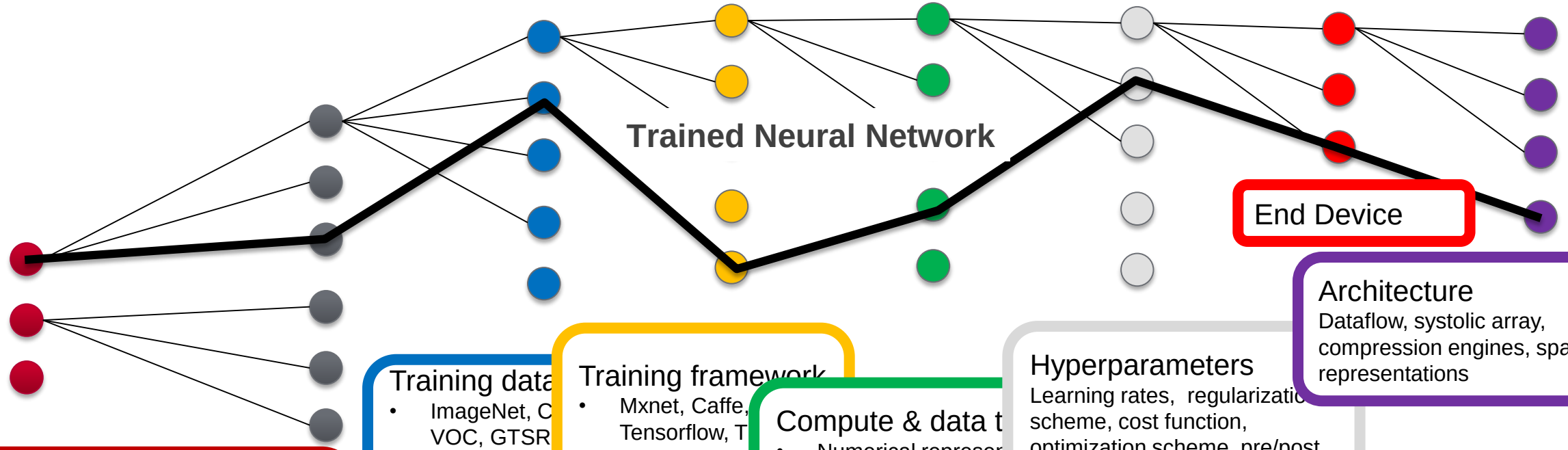
Data representations and **quantization methods** are changing

Graph Connectivity is changing

Challenge

Continuous stream of new algorithms

Challenge 5: Multidimensional Design Space



End Device

Architecture
Dataflow, systolic array, compression engines, sparse representations

Hyperparameters
Learning rates, regularizatio scheme, cost function, optimization scheme, pre/post processing

Compute & data t
- Numerical represer precision, quantizat functions, activation functions, non-linearity

Training framework
- Mxnet, Caffe, Tensorflow, T Darknet

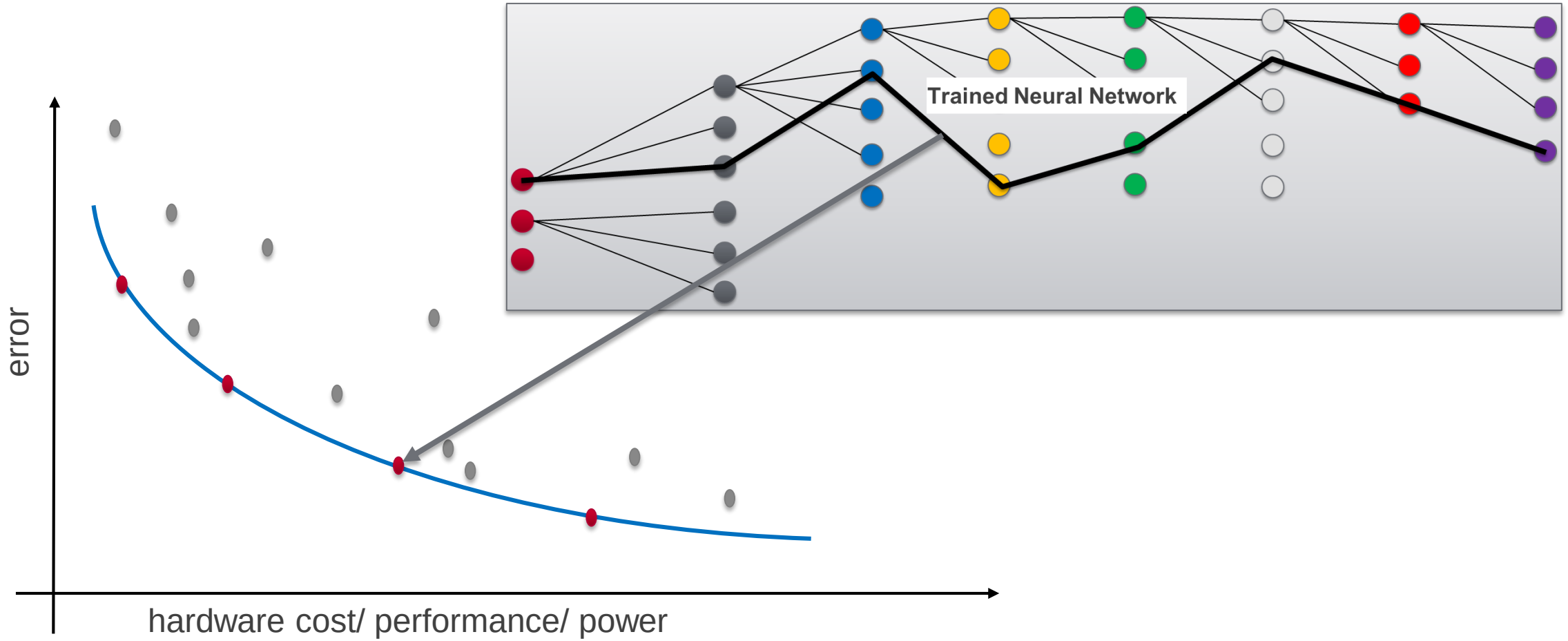
Training data
- ImageNet, C VOC, GTSR MNIST, CIFA GSVHN, City KITTY

ML task (growing list)
- Vision: image classification, recognition, sem segmentation, S
- Audio: voice con
- Gaming strategy
- Recommender s

Neural Network topology
- AlexNet, GoogleNet, ResNet-X, Enet, Yolo

Each Trained & Implemented NN
Each combination yields to a different point in the design space: error, cost, throughput, latency, power

Each Trained Neural Network Implementation yields a different Trade-Off



Research Questions at Hand:

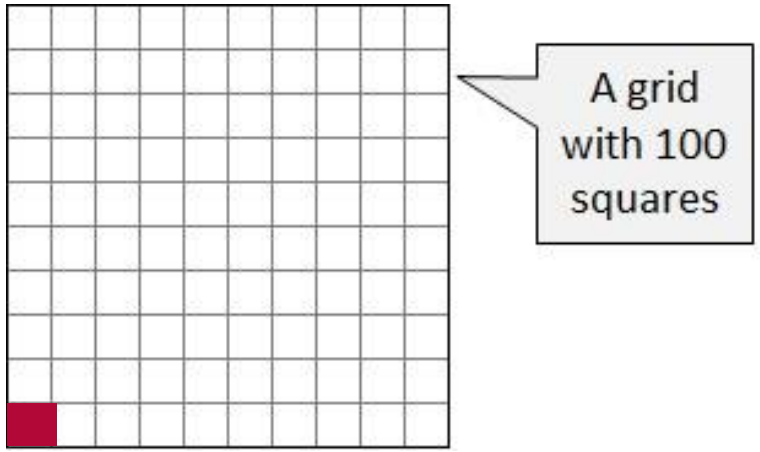
- **Given a specific ML problem, what is the best implementation that we can achieve with existing and future devices**
  - Customizing hardware architecture
- **Complimentary: What network would give me the highest performant implementation?**
  - Transforming the algorithm

We're doing a bit of both:
using reduced precision neural networks
to build end-systems for embedded compute environments

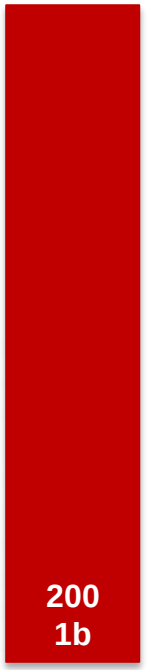
- Background: Xilinx & Xilinx Research
- Machine Learning and its Challenges
- Xilinx Effort
- Summary & Questions

Reduced Precision Neural Networks

Potential for Reducing Precision



- **Reducing precision from 32b to 1b shrinks LUT cost**
  - Instantiate 100x more compute within the same fabric
- **Potential to scale CNN performance to above 200TOPS on todays devices**
- **Potential to reduce memory footprint by 32x**
  - NN model can stay on-chip => no memory bottlenecks

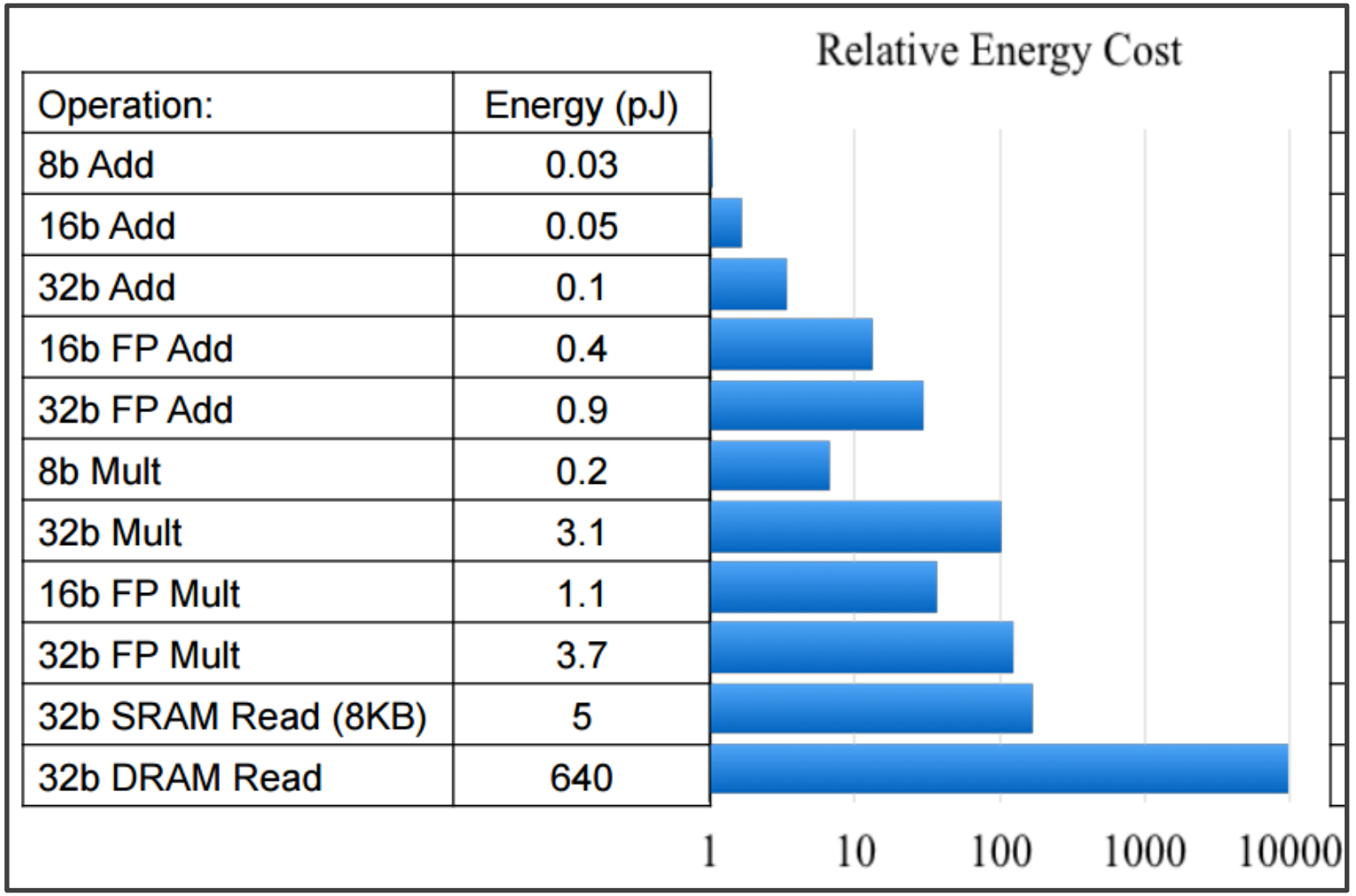


16nm

Reducing precision provides the ability to scale above 200TOPS on 16nm

Quantizing and Fixed Point 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

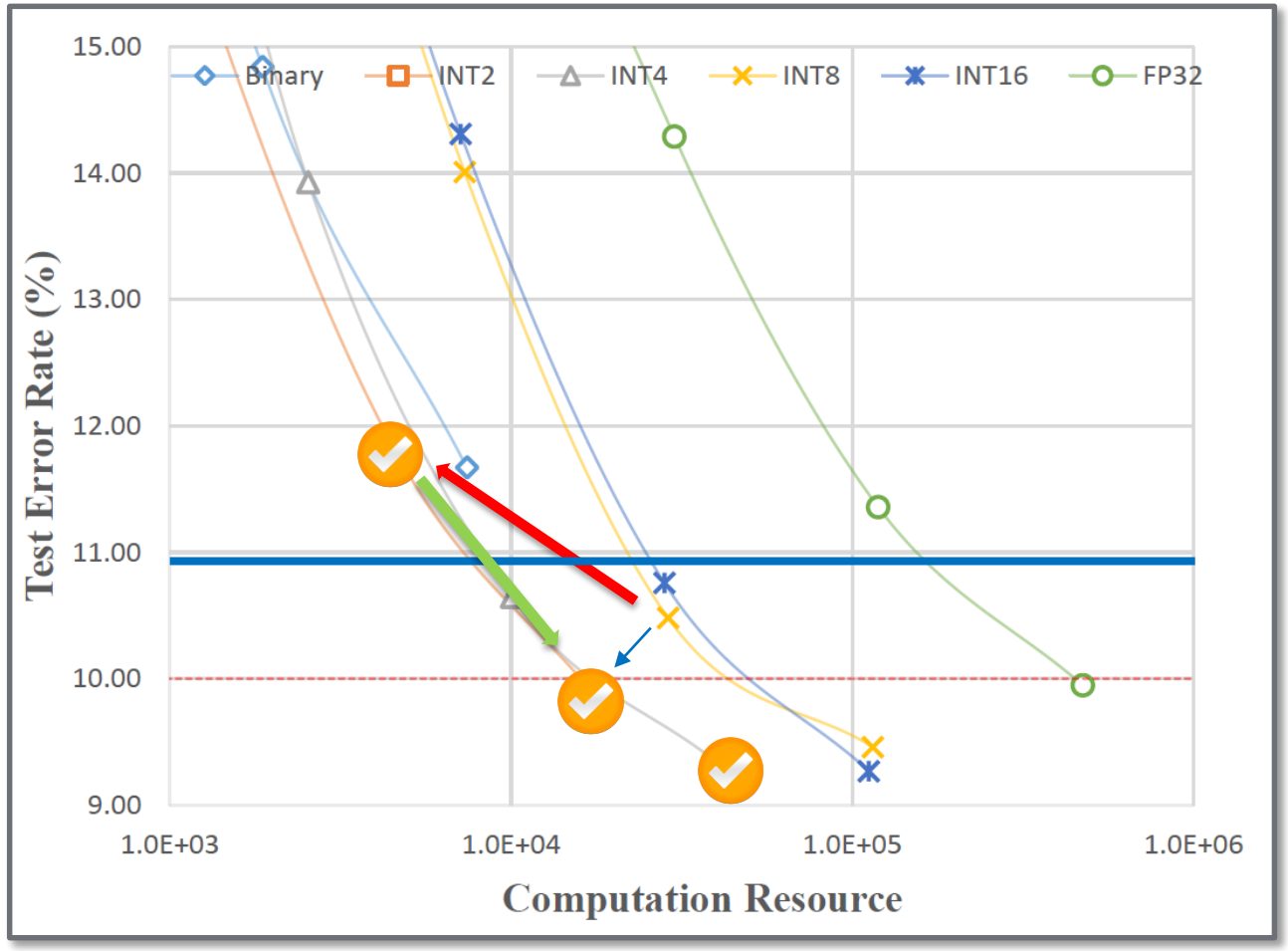


Source: Bill Dally (Stanford), Cadence ***Embedded Neural Network Summit****, February 1, 2017*

Do we loose Accuracy?

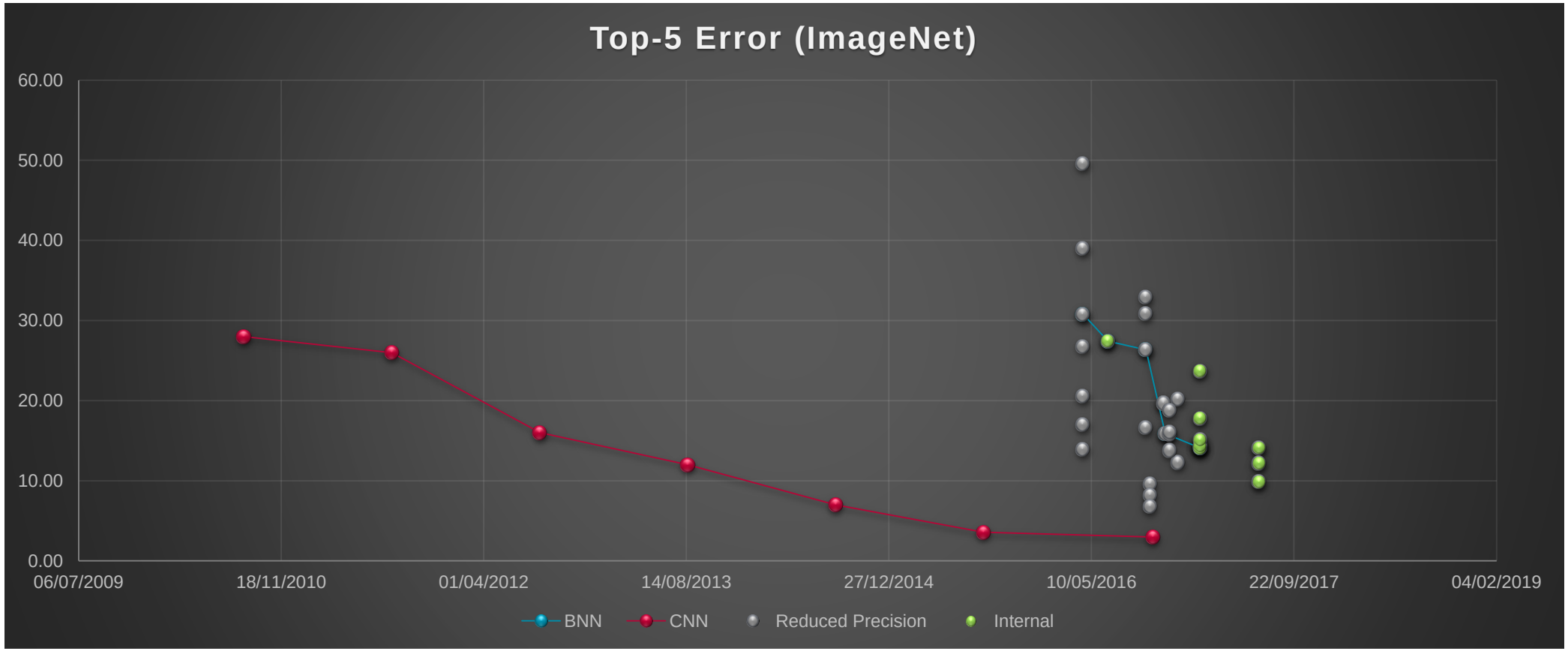
Compensating Quantization with Network Complexity

- Just reducing precision, reduce hardware cost & increases error
- Recuperate accuracy by retraining & increasing network size
- 1b, 2b and 4b provide pareto optimal solutions



Accuracy of Quantized Neural Networks (QNNs) Improving

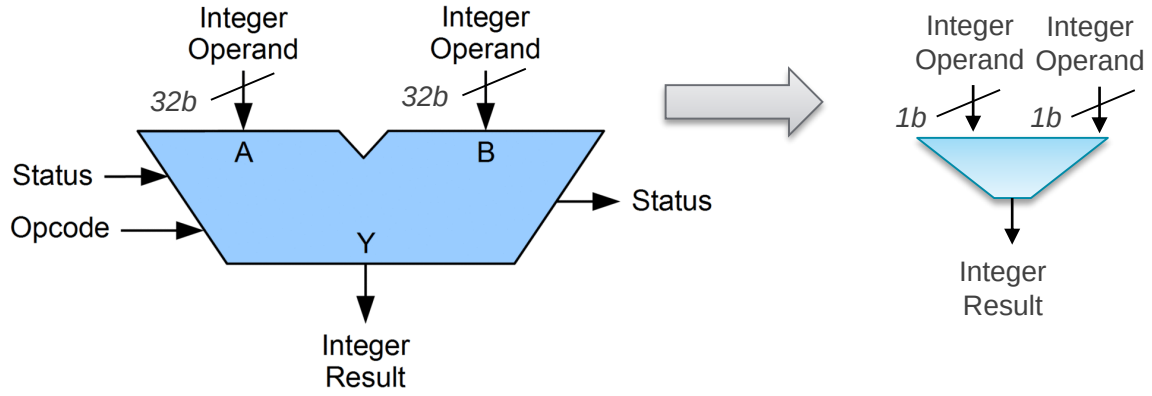
Published Results for FP CNNs, QNNs and binarized NNs (BNNs)



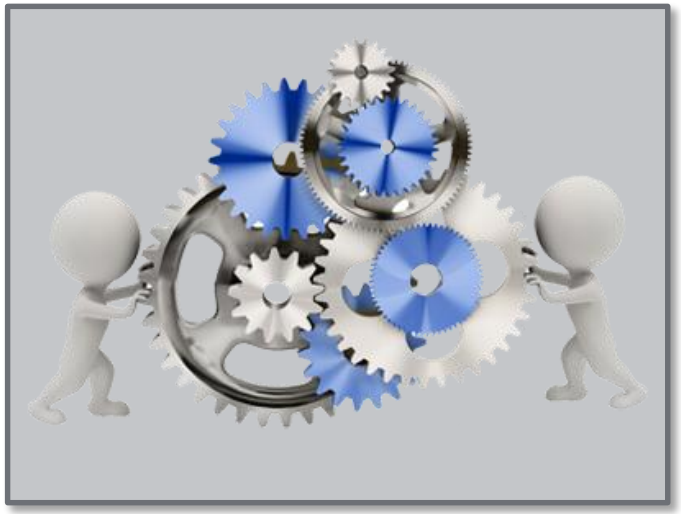
Accuracy results are improving rapidly through for example new training techniques, topological changes and other methods

Customized Hardware Architecture

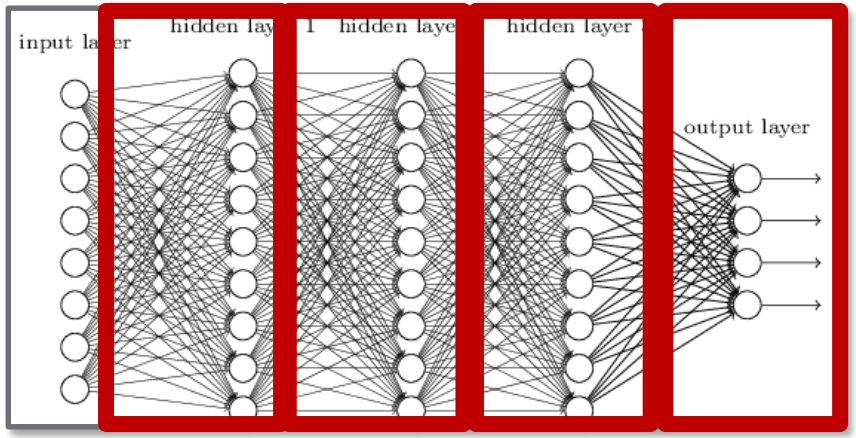
- **1. Customizing the arithmetic operations to specific reduced precision operations**



- **2. Building a dataflow implementation of the neural network**

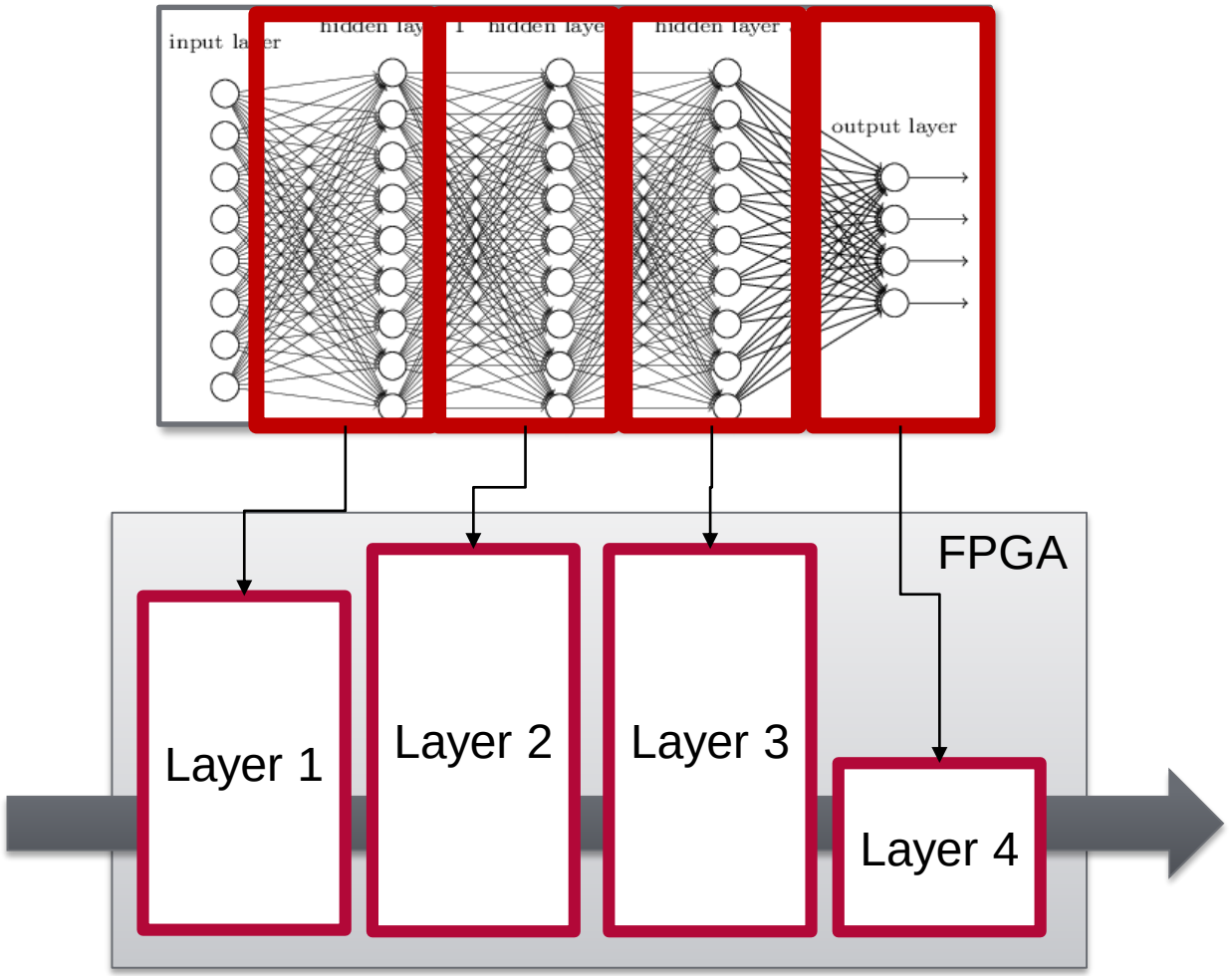


GPU and ASIC Architectures

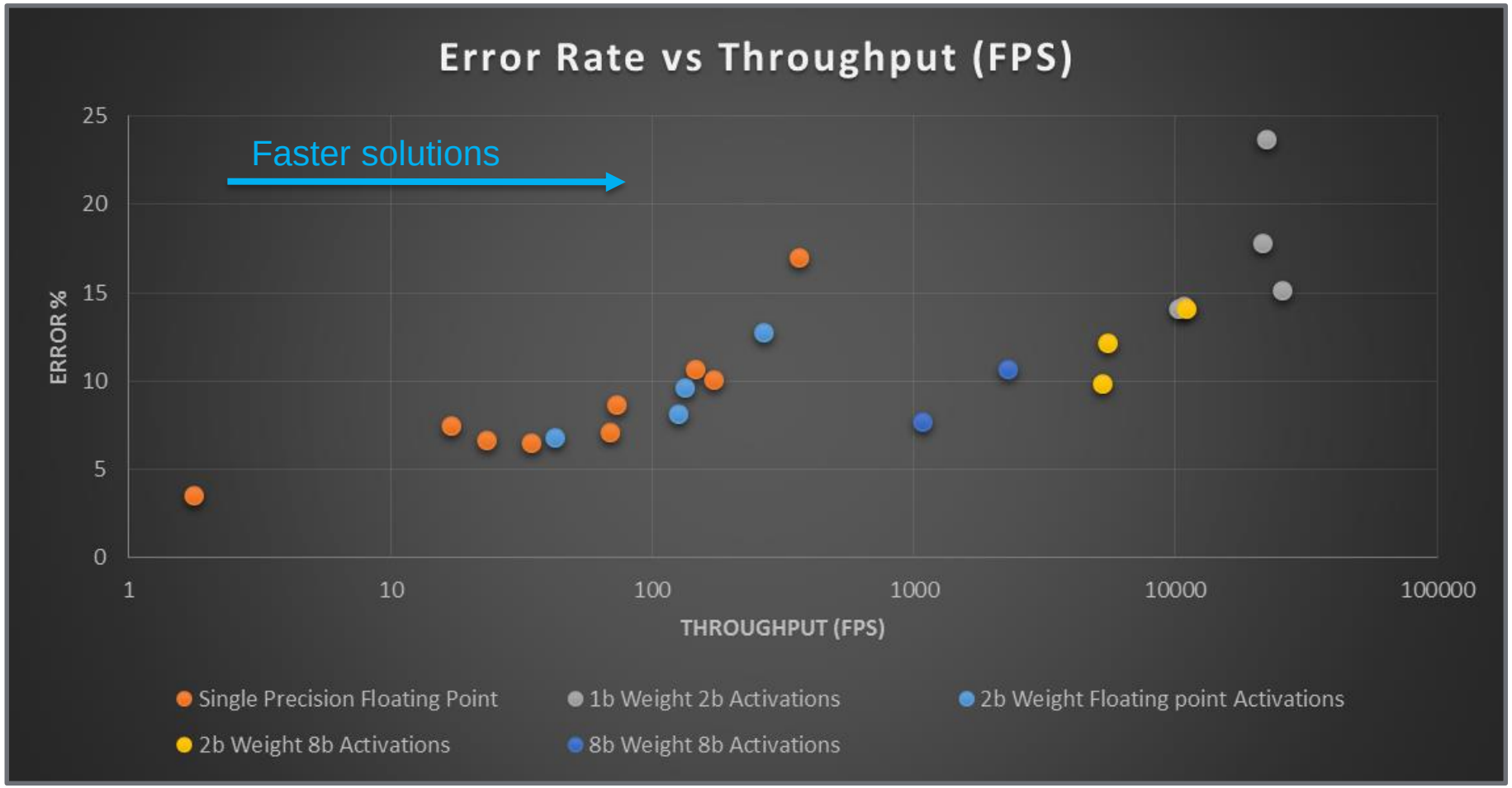


- Compute 1st layer
- Compute 2nd layer
- Compute 3rd layer
- Compute 4th layer

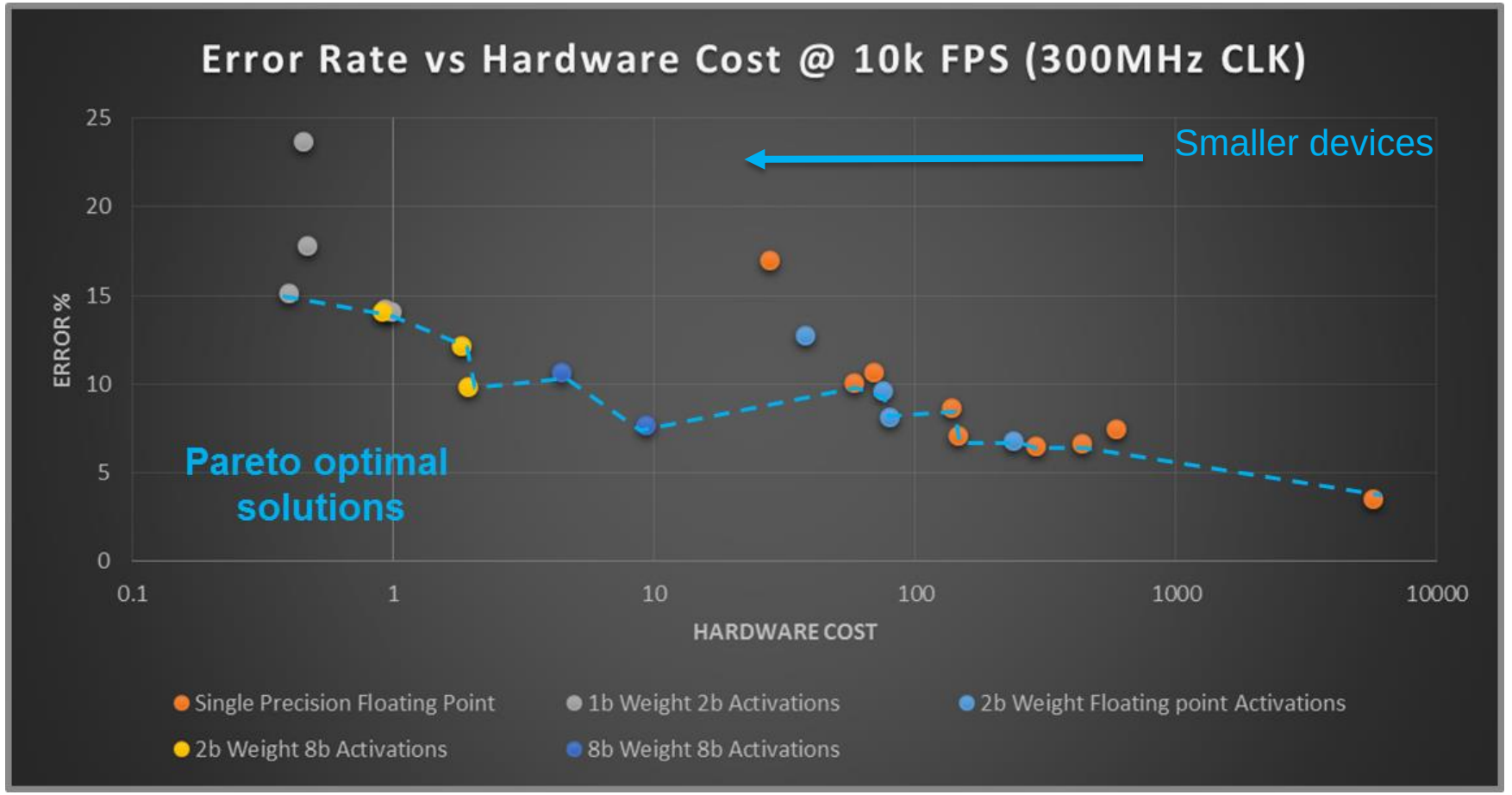
Custom-Tailored Hardware



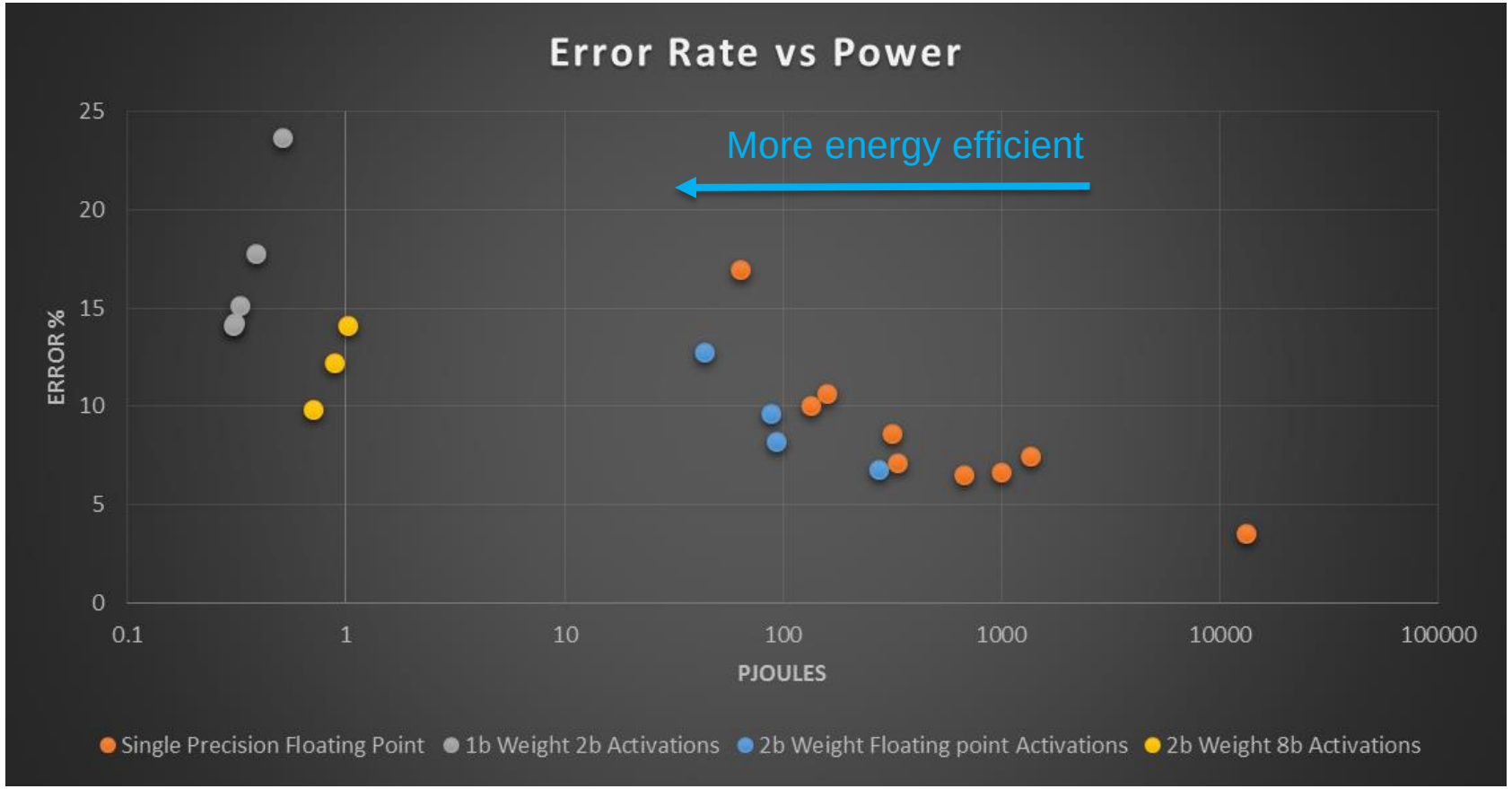
Results: Design of neural network accelerators ***that are faster***



Results: Design of neural network accelerators ***that are cheaper***



Results: Design of neural network accelerators ***that save energy***



*Pjoules excluding memory subsystem, only calculated & extrapolated on the basis on pjoule/operation (Horowitz)

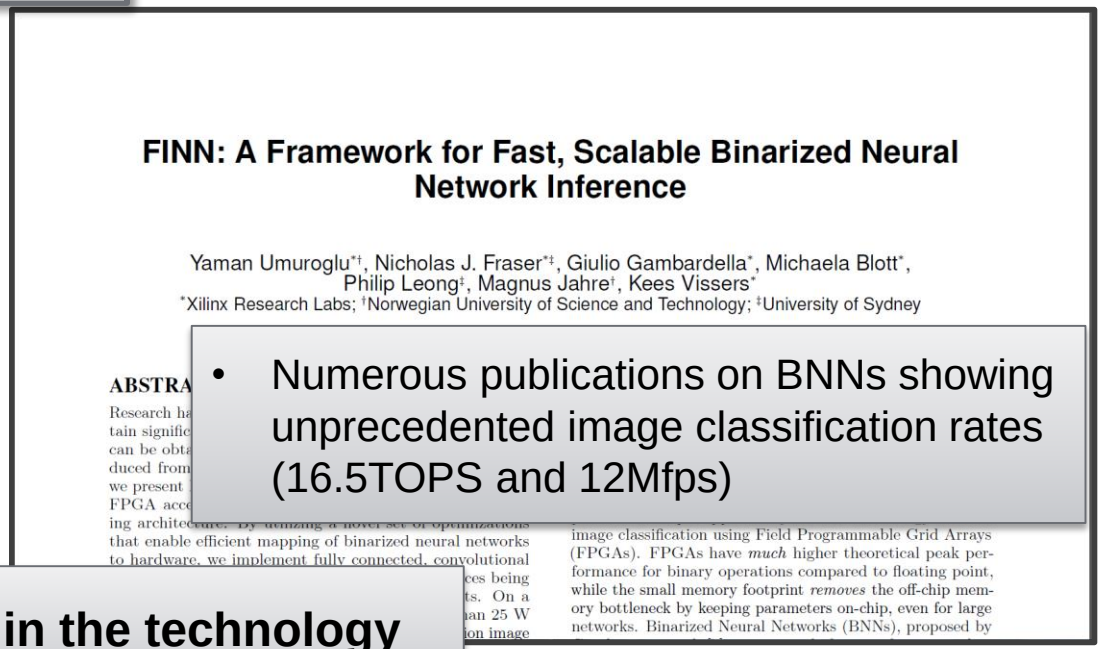
Status & Highlights

- Solitaire demo & Yolo demo powered by at numerous conferences (EmbeddedWorld, CVPR 2017, NIPS 2017, FPL 2017)

PYNQ™

- Open source release showing 1000x over Raspberry Pi3
- 140*s on github

- 1 Patent filed & 2nd in preparation


FINN: A Framework for Fast, Scalable Binarized Neural Network Inference

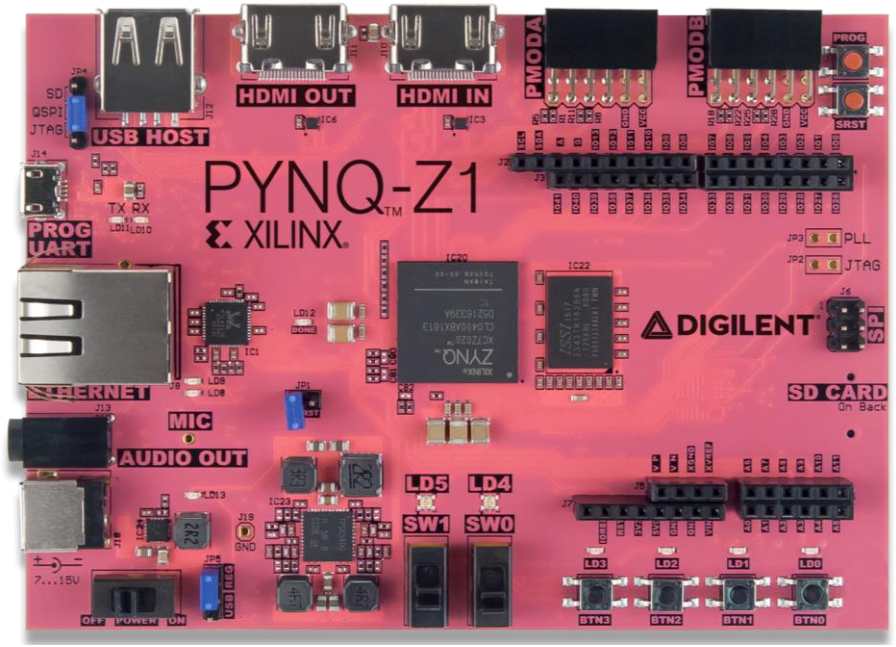
Yaman Umuroglu*†, Nicholas J. Fraser*‡, Giulio Gambardella*, Michaela Blott*, Philip Leong‡, Magnus Jahre†, Kees Vissers*
*Xilinx Research Labs; †Norwegian University of Science and Technology; ‡University of Sydney

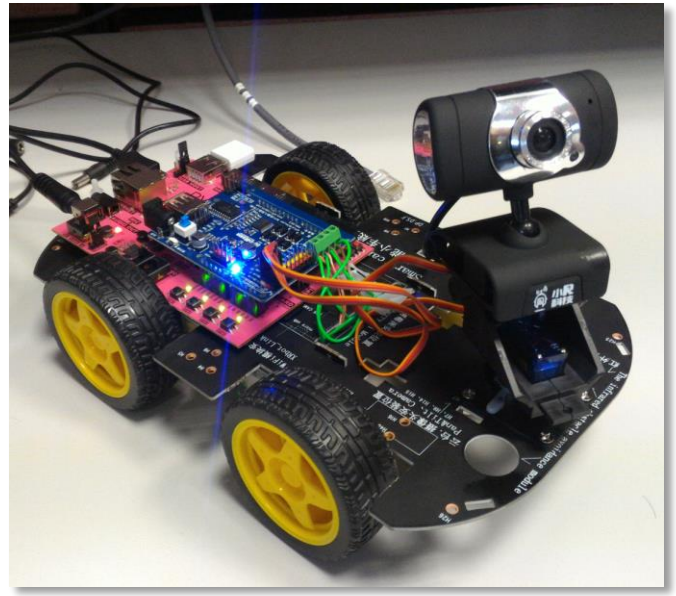
- Numerous publications on BNNs showing unprecedented image classification rates (16.5TOPS and 12Mfps)

Demonstrating the potential in the technology

PYNQ Example Designs

- **USB powered**
- **Jupyter notebooks**
- **Python APIs to talk to the FPGA fabric**
- **Open source release on PYNQ**
  - https://github.com/Xilinx/BNN-PYNQ
  - For binarized MLPs and CNNs
- **Can work of webcams and integrated into a robotics kit**
- **and much more to come…**





- Background: Xilinx & Xilinx Research
- Machine Learning and its Challenges
- Xilinx Effort
- **Summary & Questions**

Summary

- **Deep Learning creates challenges for the semiconductor industry**
- **Solutions: custom architectures on FPGAs & reduced precision can enable real-time processing on embedded systems at low power consumption**
- **Demonstrating customized algorithms and architectures on FPGAs**
  - Proofing significant value on reduced precision CNNs with dataflow architectures
  - Top-in class in regards to compute efficiency & real-time response
- **Many open research questions within this multi-dimensional design space remain**
- **Typical example of computer engineering challenge that will increasingly emerge**

Thank You.

- mblott@xilinx.com
- And remember: we do have internships on a broad base of subjects ☺