Design Trade-offs for Machine Learning Solutions on Reconfigurable Devices

Michaela Blott
Principal Engineer, Xilinx Research
July 2018

Agenda

Background – Xilinx Research

Machine Learning

Research Efforts

Summary & Outlook

XILINX.

Agenda

Background – Xilinx Research

Machine Learning

Research Efforts

Summary & Outlook

XILINX

Xilinx Research - Ireland

- Since 13 years
- Part of the worldwide CTO organization (8 out of 36)
- AI Lab expansion part-financed through IDA Ireland
- Increasingly external funding (H2020))

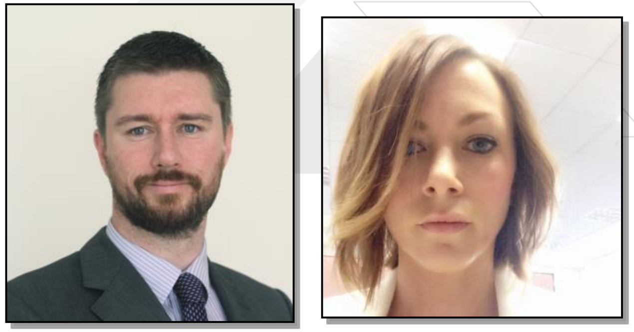
Ivo Bolsens
CTO

Kees Vissers
Fellow

Current Xlabs Dublin Team

- **Yaman Umuroglu, Ken O'Brien, Nick Fraser, Giulio Gambardella, Alessandro Pappalardo, Peter Ogden, Lucian Petrica, me (from left to right)**
  - More faces to be added soon

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



XILINX.

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
- Derisk emerging technologies
- In partnership with universities, customers, and partners
- Current Focus:

  Quantifying value proposition for FPGAs in Machine Learning
  - Prototyping, testdriving, benchmarking

XILINX

Agenda

Background – Xilinx Research

Machine Learning

Research Efforts

Summary & Outlook

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 BILLION-DOLLAR CRUSADE TO STOP THE A.I. APOCALYPSE

Elon Musk is famous for his futuristic gambles, but Silicon Valley's latest rush to embrace artificial intelligence scares him. And he thinks you should be frightened too. Inside his efforts to influence th[e] rapidly advancing field and its proponents, and to save humanity from machine-learning overlords.

BY MAUREEN DOWD
APRIL 2017

Convolutional Neural Networks (CNNs)

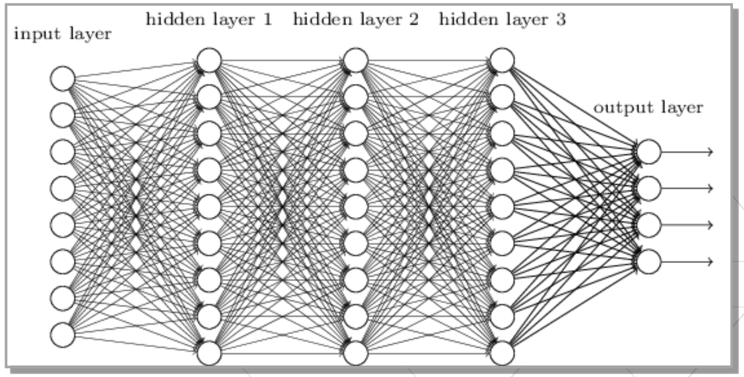
> **CNNs are the predominant ML algorithm used**
  >> Mimics the human brain
  >> Works very well for image classification, speech recognition

> **NNs are the "universal approximation function"**
  >> If you make it big enough and train it with enough data
  >> Can outperform humans on specific tasks

> **Requires zero domain expertise**

> **Will increasingly replace other algorithms**
  >> unless for example simple rules can describe the problem

> **and solve previously unsolved problems**



Machine Learning will help address the Grand Engineering Challenges of the 21st Century (NAE)

- Make solar energy economical
- Reverse-engineer the human brain
- Secure cyper space
- Restore & improve urban infrastructure
- Engineering better medicine
- Advance health informatics
- …

Jeff Dean, Google @ Strata Data Conference, 2018

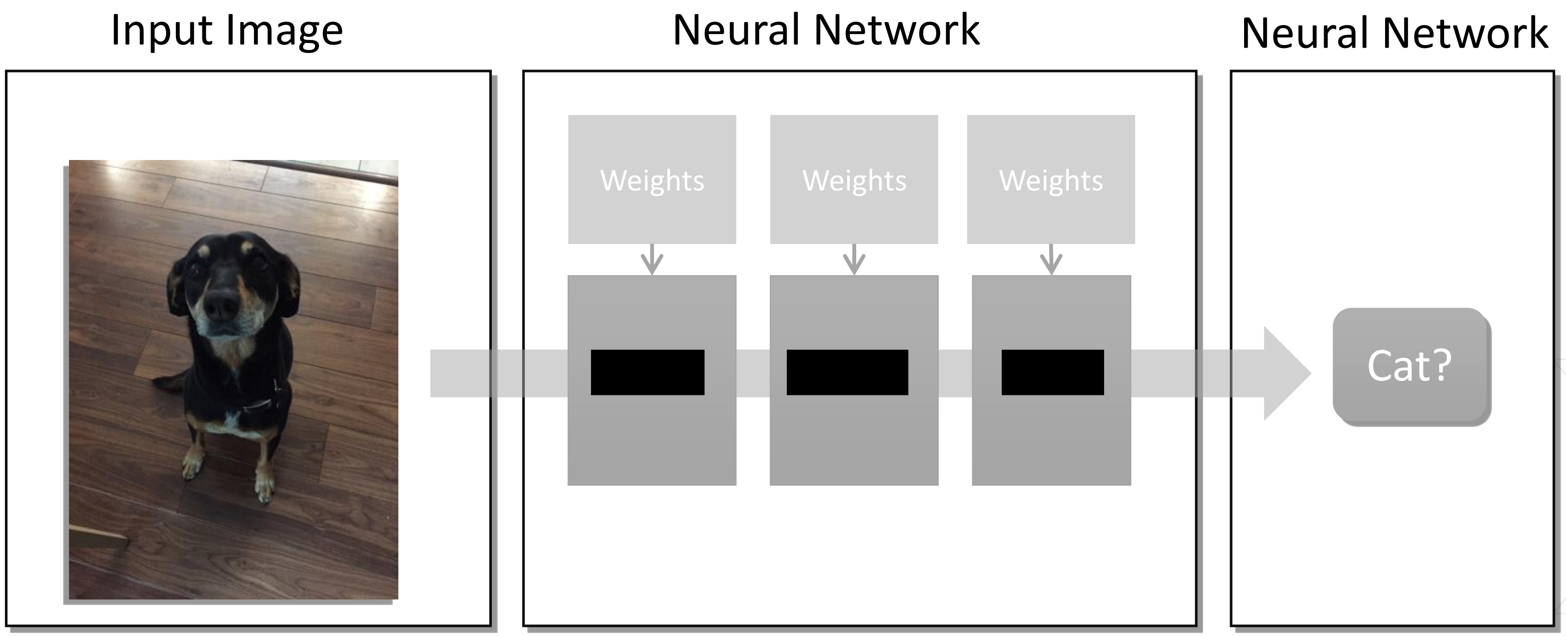
> "I actually think machine learning is going to help with all of these," the legendary computer scientist said. "I think there are actually going to be significant breakthroughs in some of these Grand Challenges that are at least in part fueled by the fact that we now have machine learning at scale with many of these techniques that can really push us forward in the areas of commuter vision, language understanding, speech recognition, and automating and solving engineering problems."

What is the Challenge?

Challenges

- **Challenge 1:**
  - Although predominant CNN computation is simple linear algebra
  - Huge amount of compute and memory is required

Example Inference

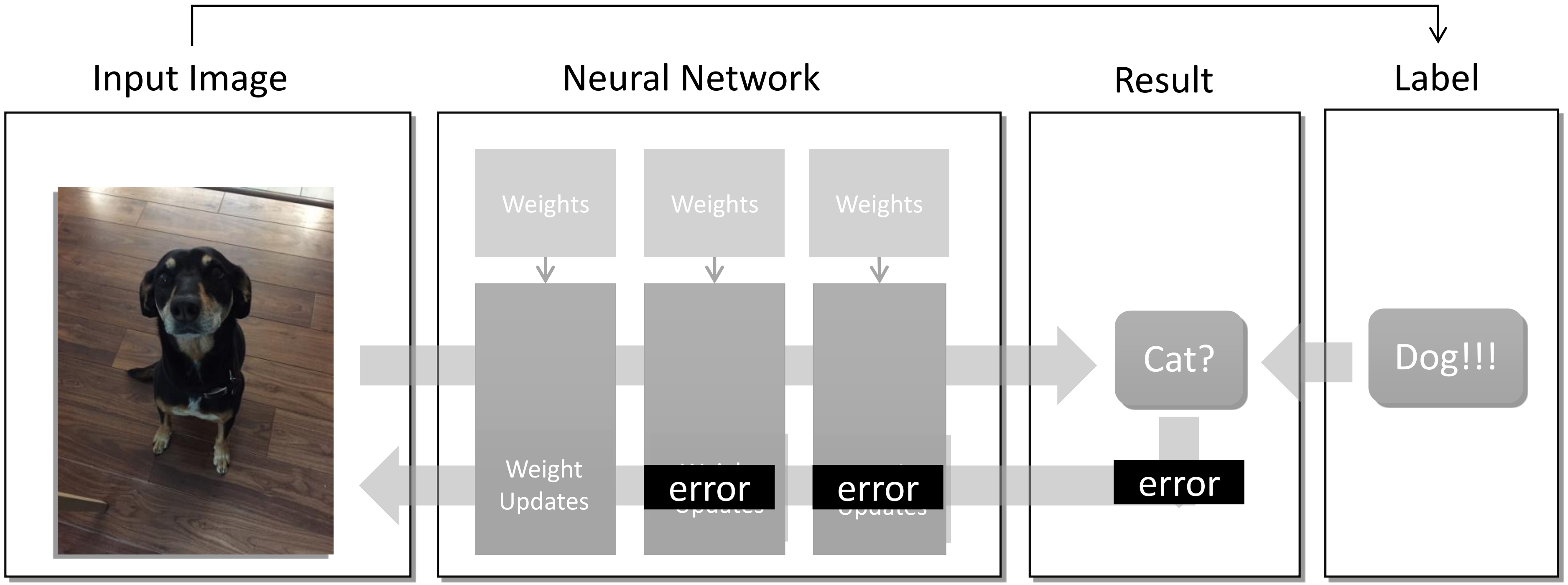


For ResNet50:

- 70 layers
- 7.7 billion operations
- 25.5 Mbytes of weight storage*
- 10.1 Mbytes for tensors*

*Assuming int8

Training – 1 Image



For ResNet50:

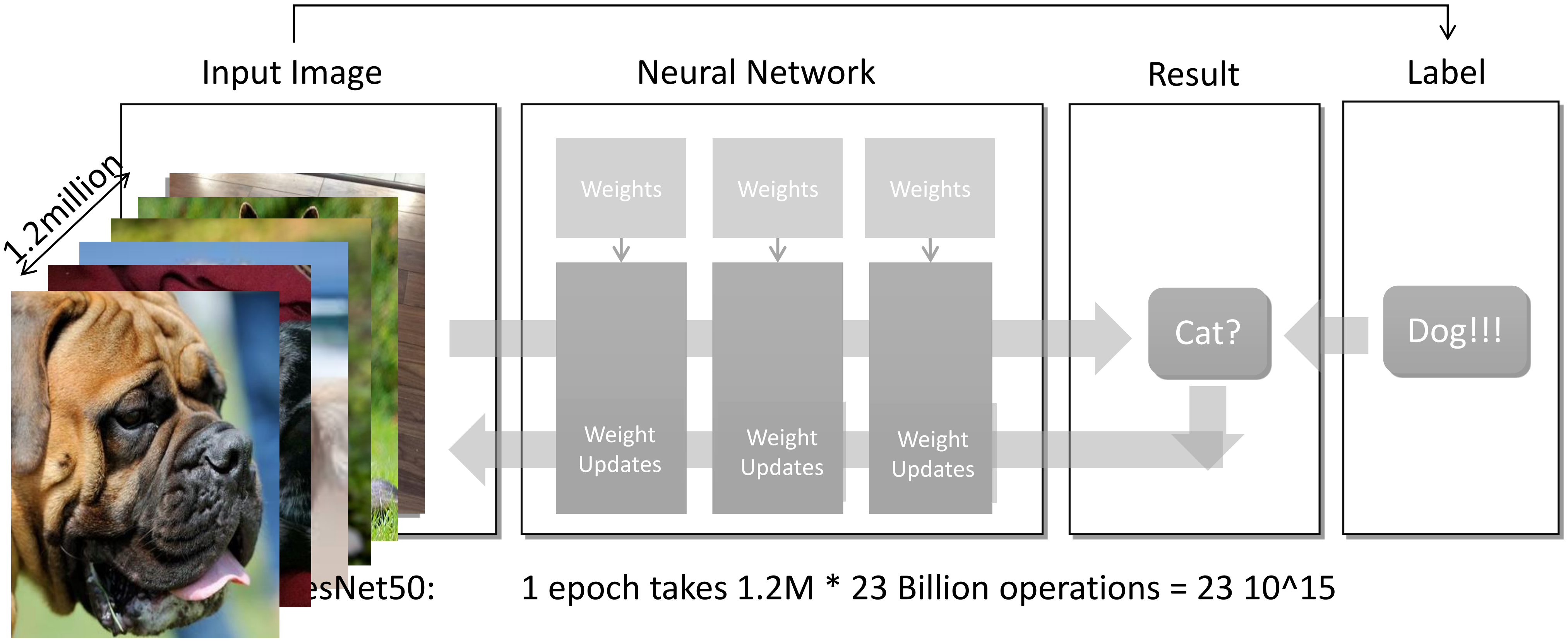
23 billion operations

weights, weight gradients, updates: 303Mbytes of storage (3-5x)

tensors, gradients: 80 Mbytes for tensors

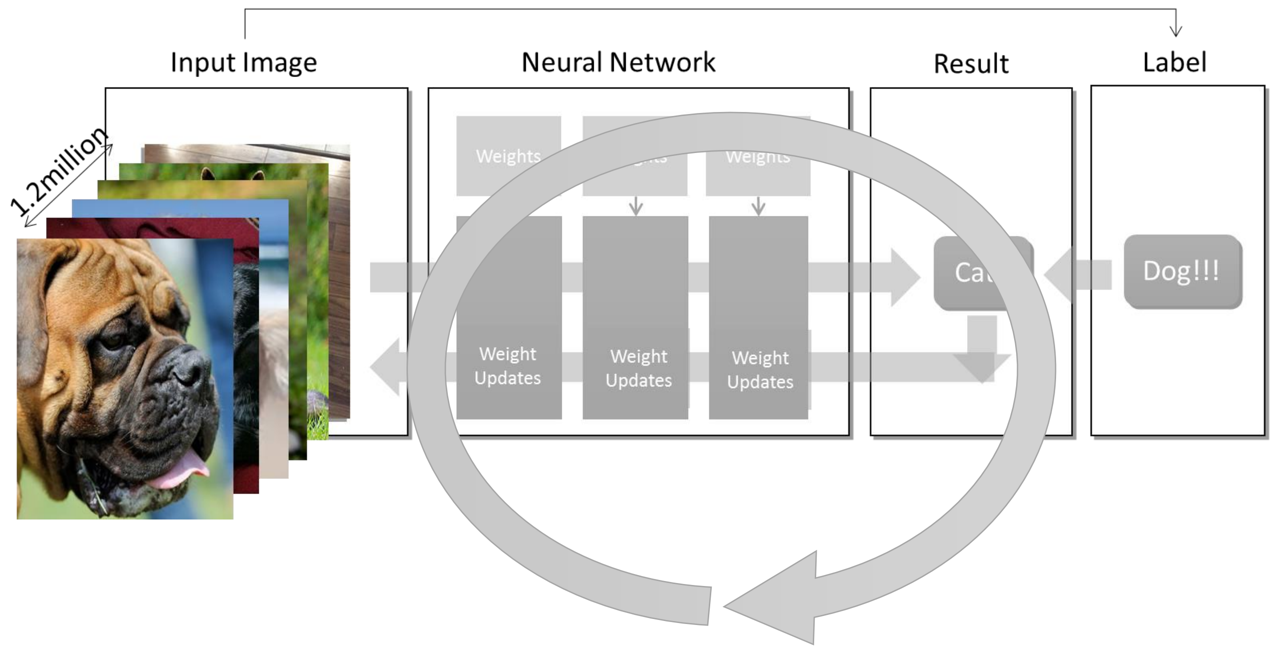
Assuming 32b SP

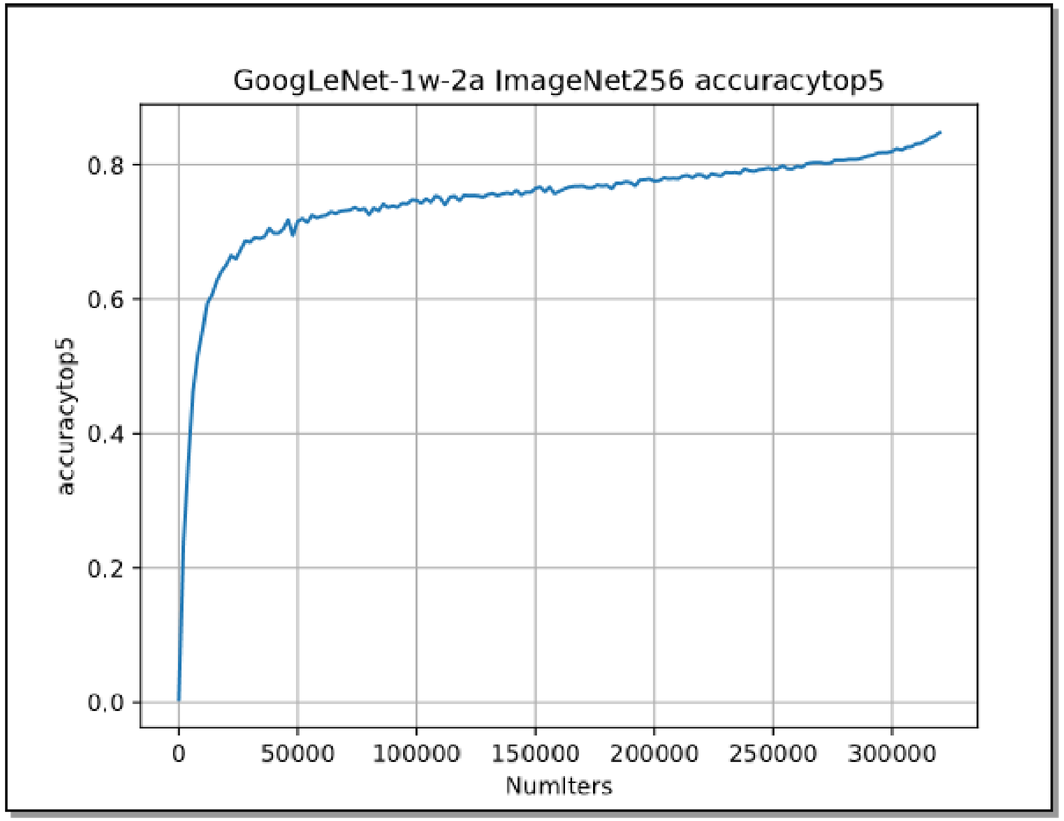
Training – 1.2 Million Images for 1 epoch



esNet50: 1 epoch takes 1.2M * 23 Billion operations = 23 10^15

Training – 100 Epochs



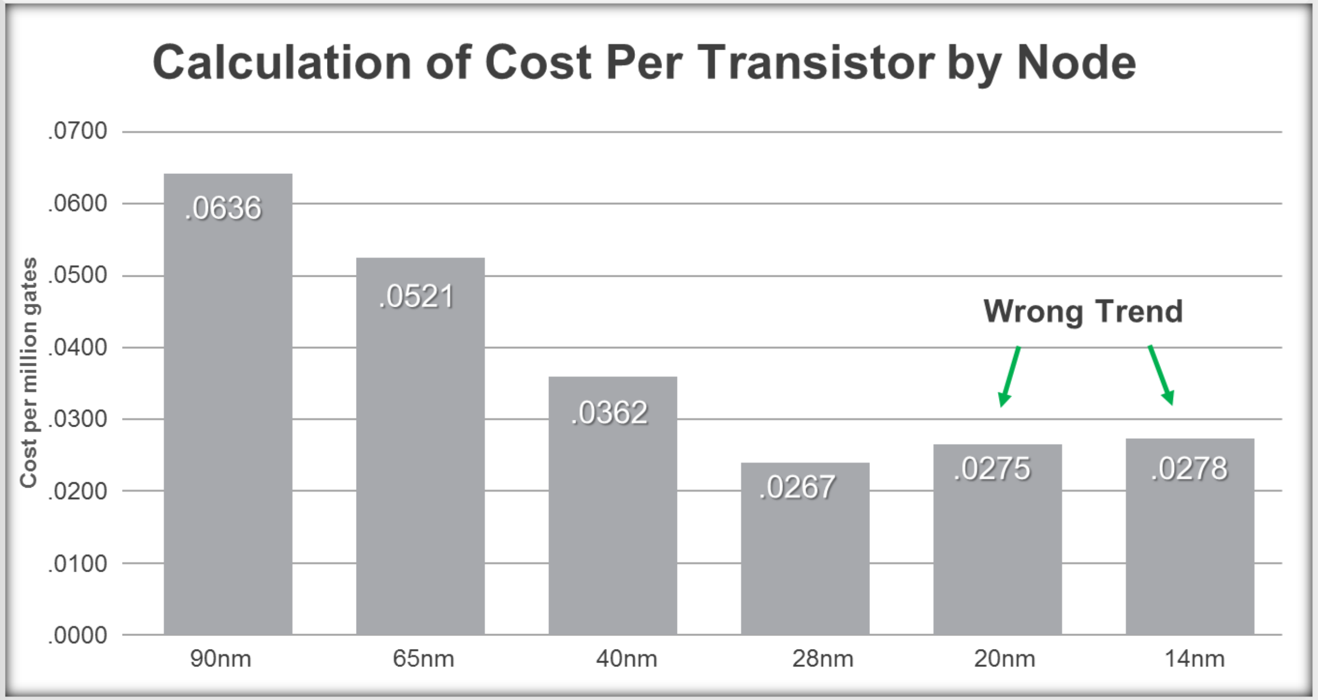


For ResNet50: 100 * 23 10^15 = 2.3 10^18

P40 GPU (12TFLOPS): 11days @ 100%, usually ~2 weeks

For inference: Billions of operations, and 10s of megabytes
For training: Quintillions of operations, and 100s of megabytes

On Crash course with End of Moore's Law



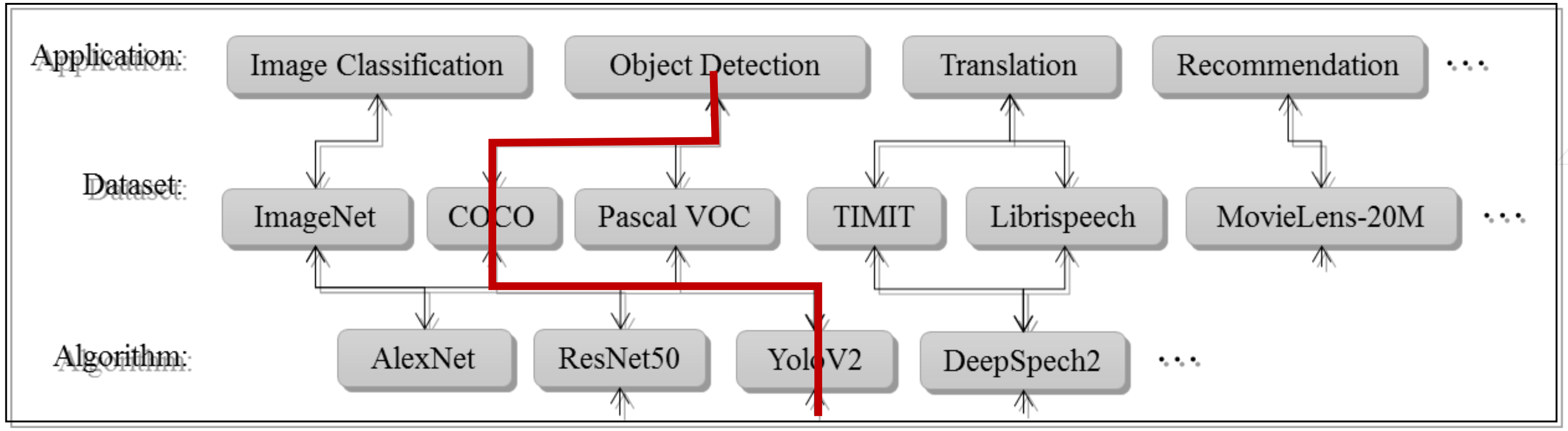
Source: IBS

> **Compute performance is no longer scaling and becomes more expensive**

Challenges

- **Challenge 1:**
  - Challenging compute and memory requirements

- **Challenge 2:**
  - Complicated design space
  - Huge variation in applications, requirements and design targets



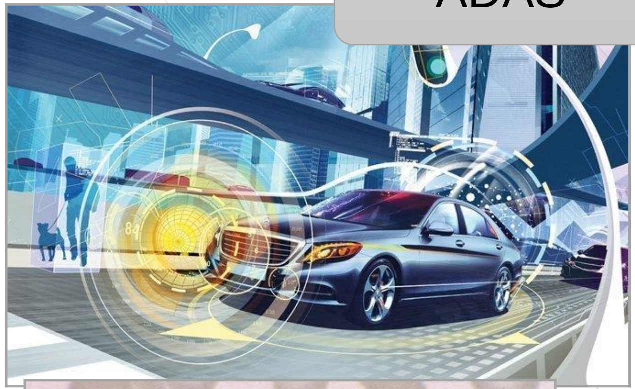
C2: Many Applications Require Different Networks

Translation Service

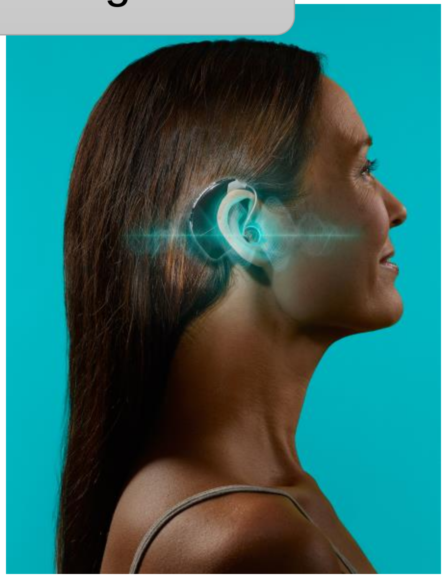
Gaming strategy

3D reconstruction from drone images

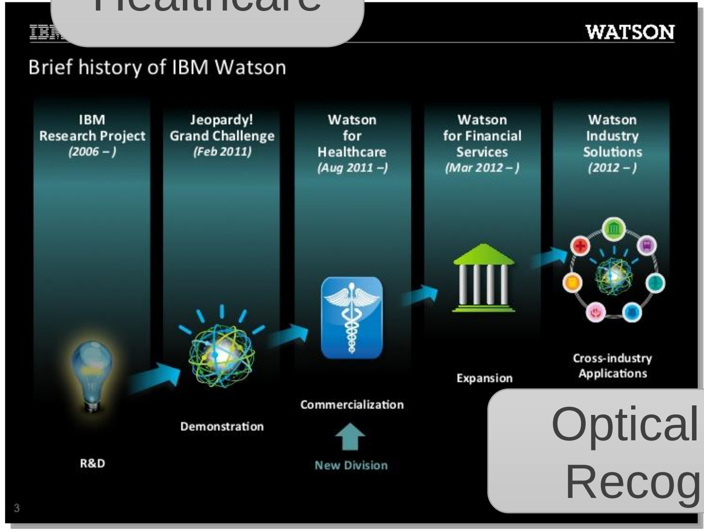
ADAS



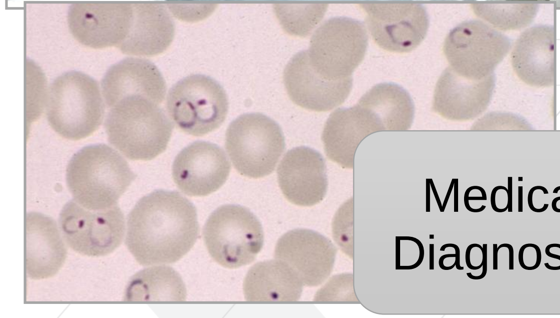
Hearing Aids



Data Analysis for Healthcare

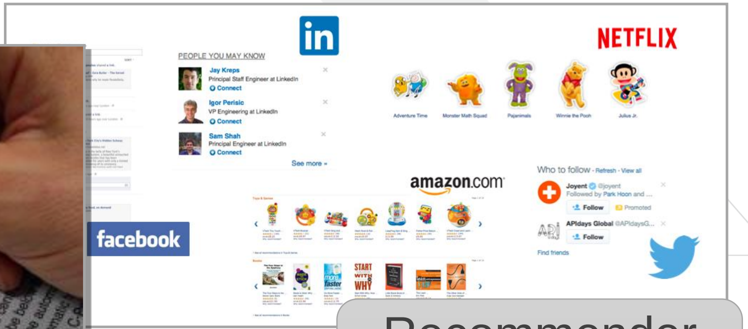


Real-time, sensor-based-control



Medical Diagnoses

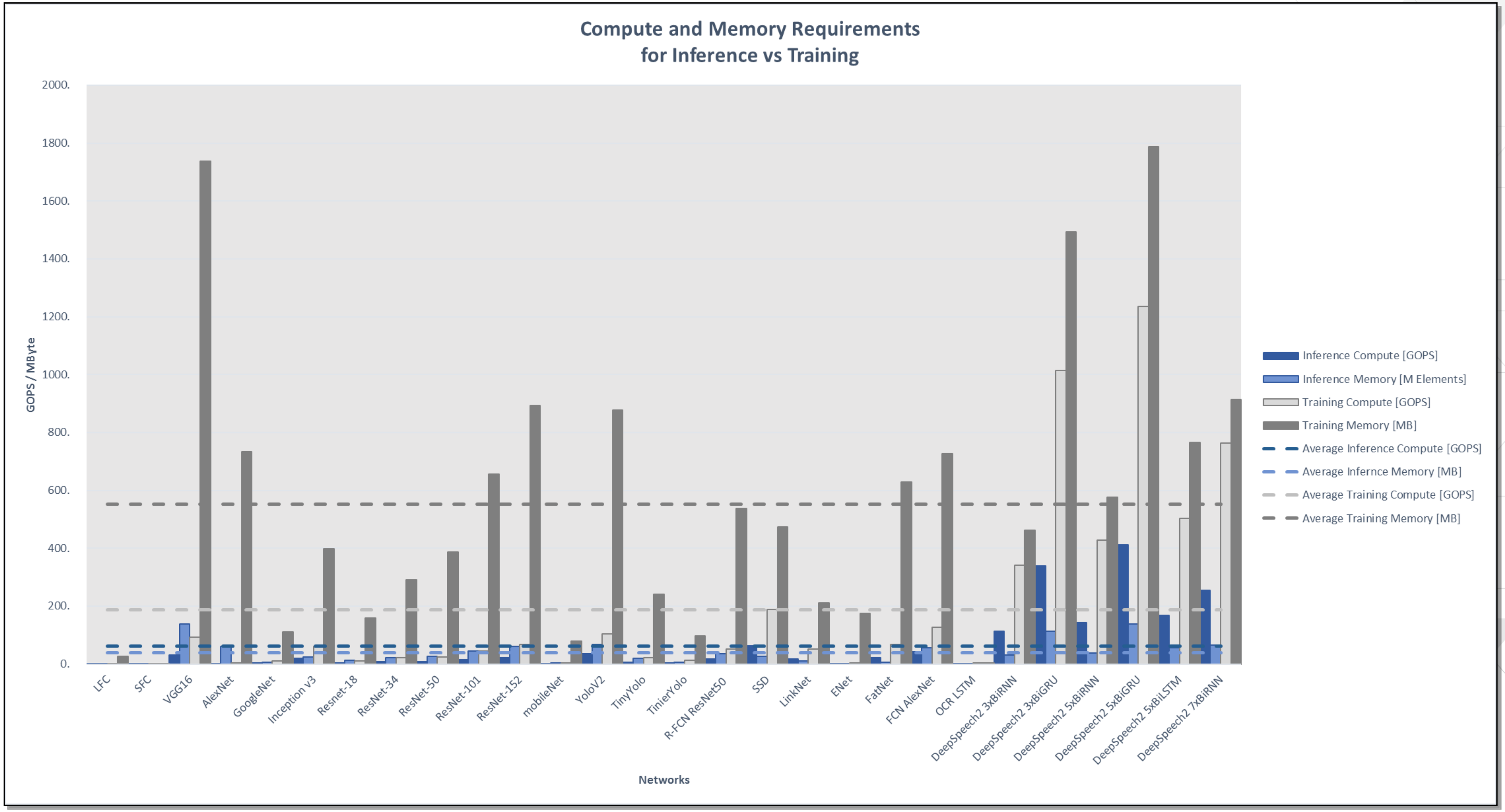
Optical Char. Recognition



Recommender Systems

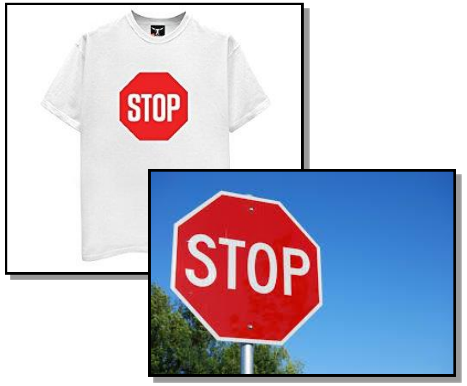
XILINX

C2: Huge Variation in Memory and Compute



C2: Different Use Cases, Different Design Targets

Accuracy, speed, power, latency, cost



> **ADAS:**
>> Accuracy
>> High throughput

> **Hearing aids:**
>> Low power
>> Very low latency
>> Low throughput

> **AR**
>> High throughput
>> Low latency
>> Low power

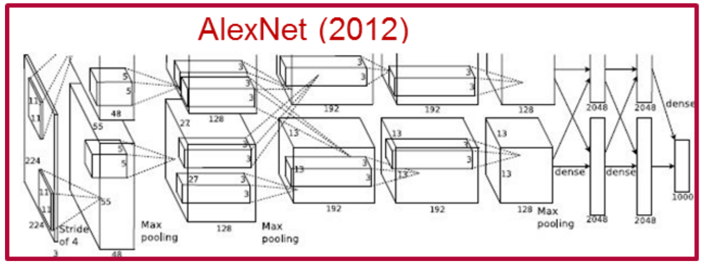
> **3D reconstruction of HR images**
>> High throughput
>> Offline

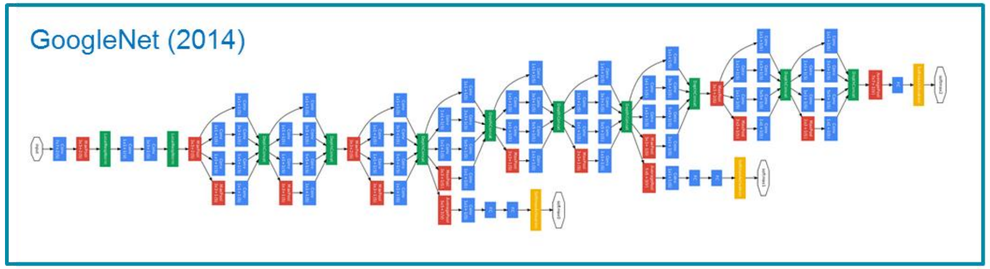
Challenges

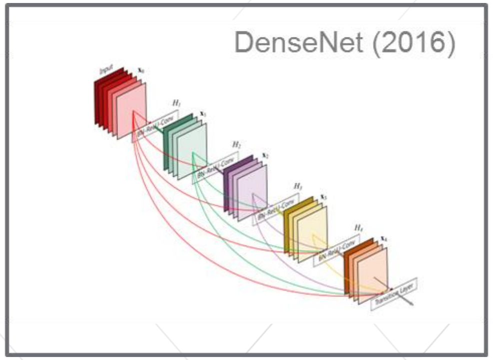
- **Challenge 1:**
  - Challenging compute and memory requirements
- **Challenge 2:**
  - Huge variation in applications, requirements and design targets
- **Challenge 3:**
  - Neural Networks Change @ Increasing Rate

C3: Neural Networks Change @ Increasing Rate

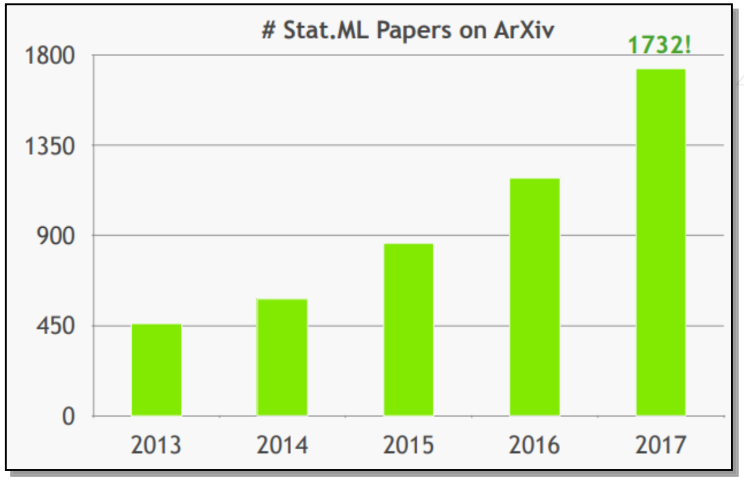
> Graph connectivity, number and types of layers are changing







> Increasing stream of research



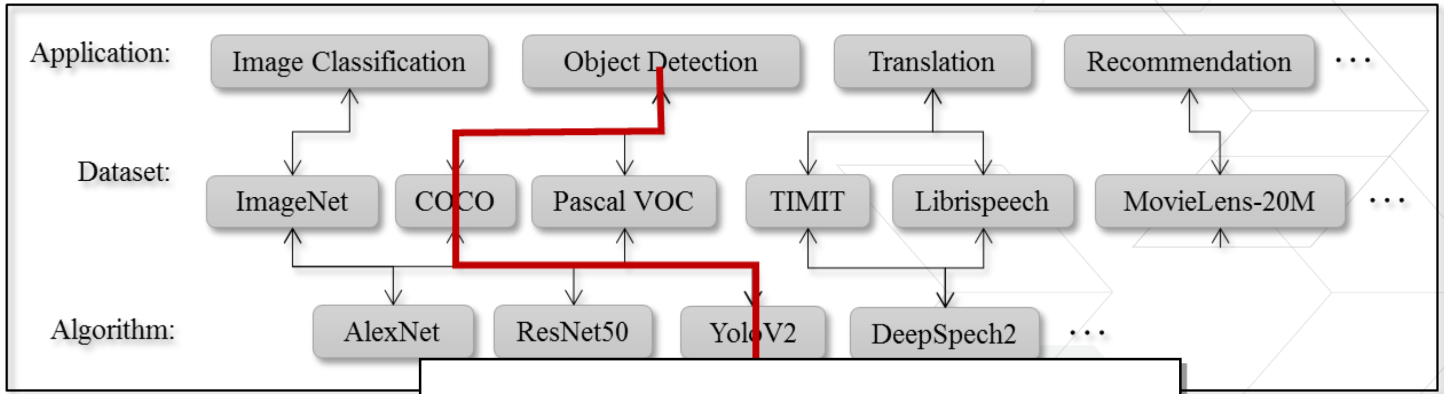
Ce Zhang, ETH Zurich, Systems Retreat 2018

XILINX

Challenges in Summary

> **Machine Learning is a very demanding use case, compute and memory intensive**
>> High variation

> **Complicated design space**
>> Different applications
>> Different and changing algorithms
>> Different figures of merits



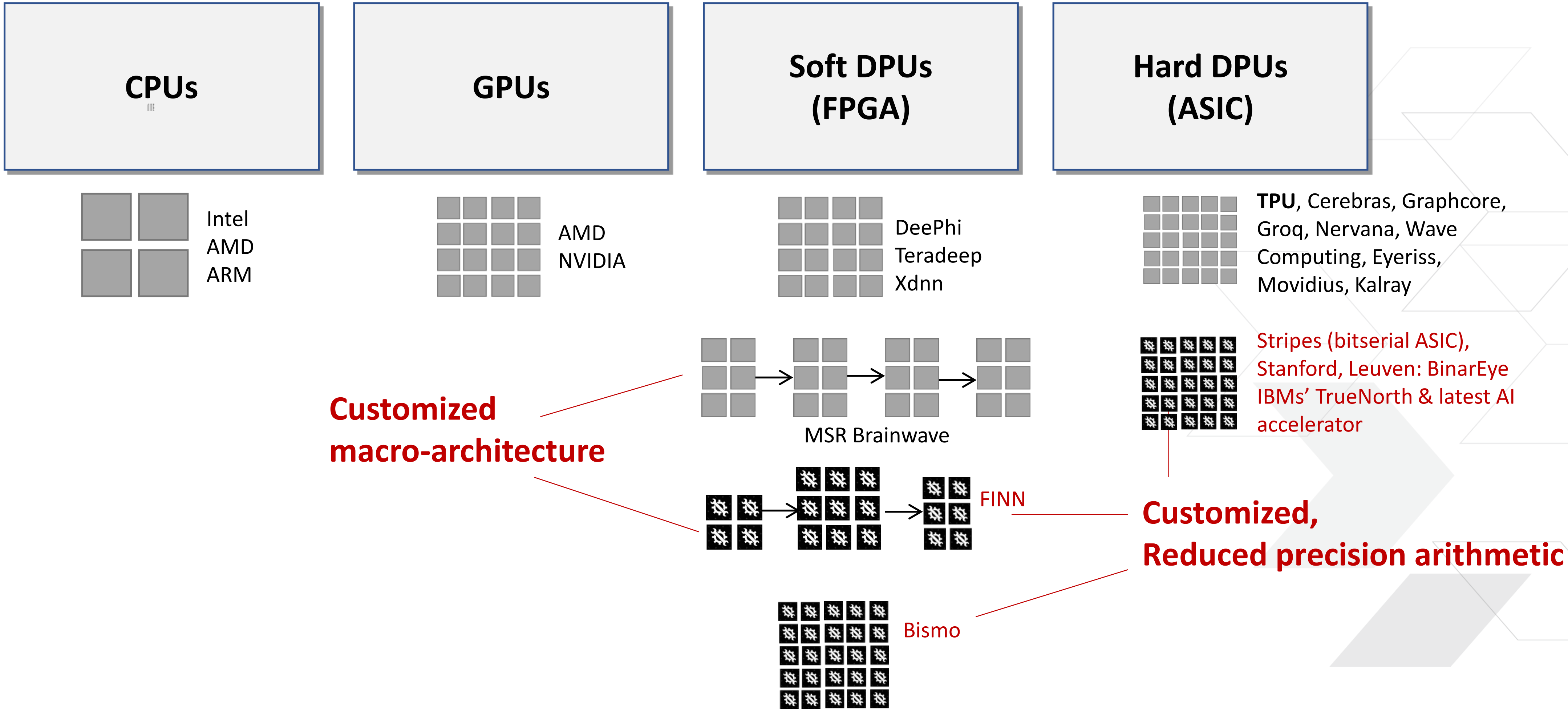
Each Combination delivers different results regarding the design targets:
Throughput, power, latency, cost,…

> **Changing requirements**

> **Need to be addressed through architectural and algorithmic innovation**

Spectrum of New Architectures for Deep Learning

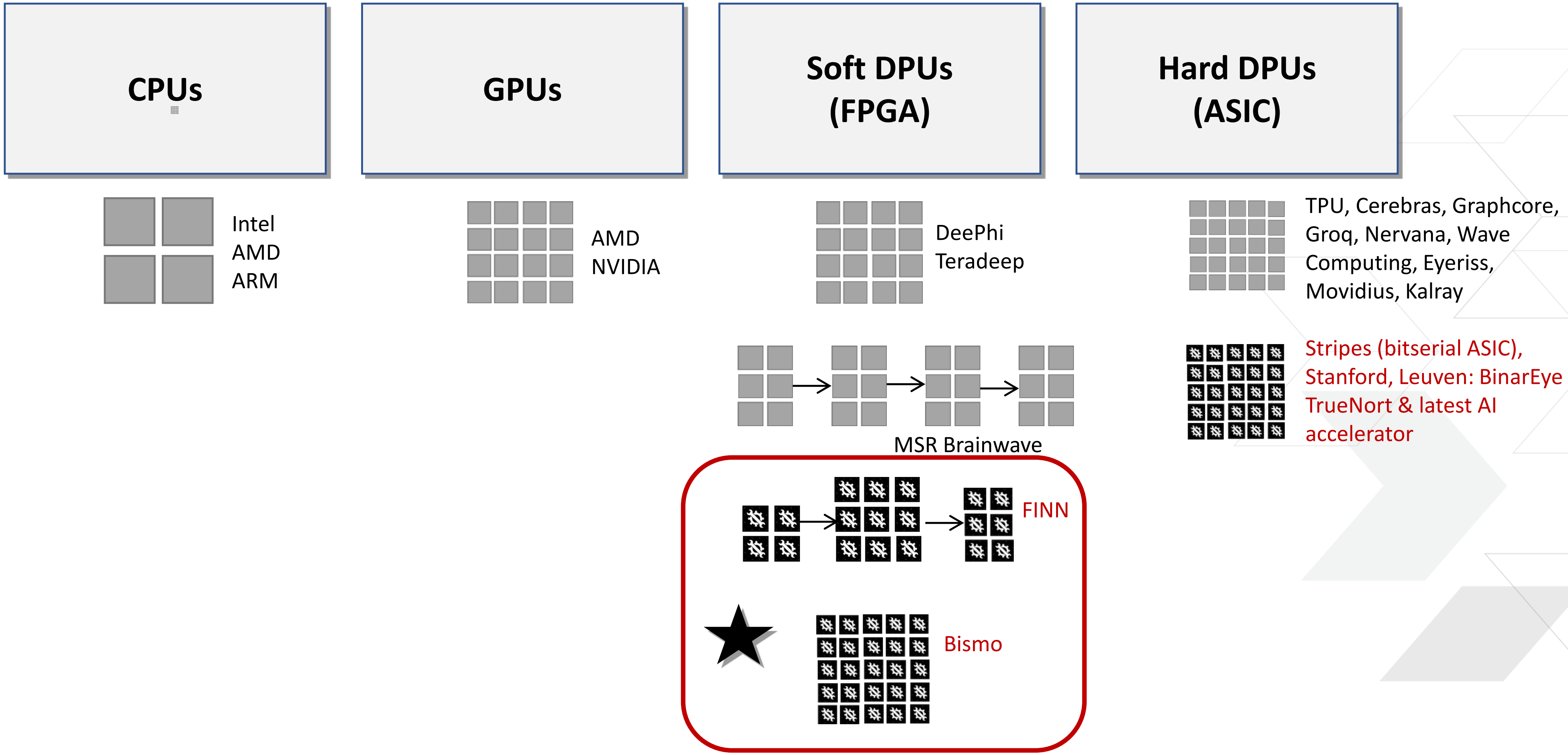
Exciting Times in Computer Architecture Research!



DPU: Deep Learning Processing Unit

Spectrum of New Architectures for Deep Learning

Efficiency vs Flexibility



XILINX

Agenda

- Background – Xilinx Research
- Machine Learning
- **Research Efforts**
- Summary & Outlook

XILINX

Our Research Effort

> **Changing neural network algorithm by reducing precision in data types to provide performance scalability, compute efficiency**
>> Numerical representations, precision, quantization

> **Customizing architecture to hit specific design targets**
>> On micro and macro level

> **Through automated tool flow (FINN) and open source platforms (PYNQ and AWS) to provide ease of use**

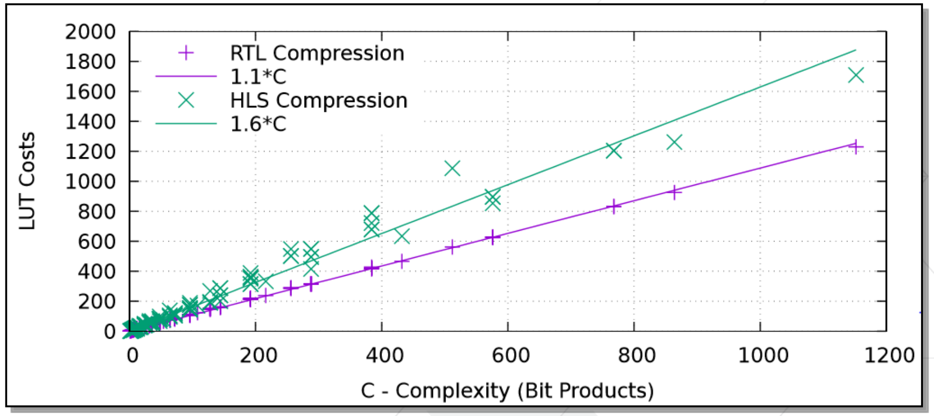
Reducing Precision

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

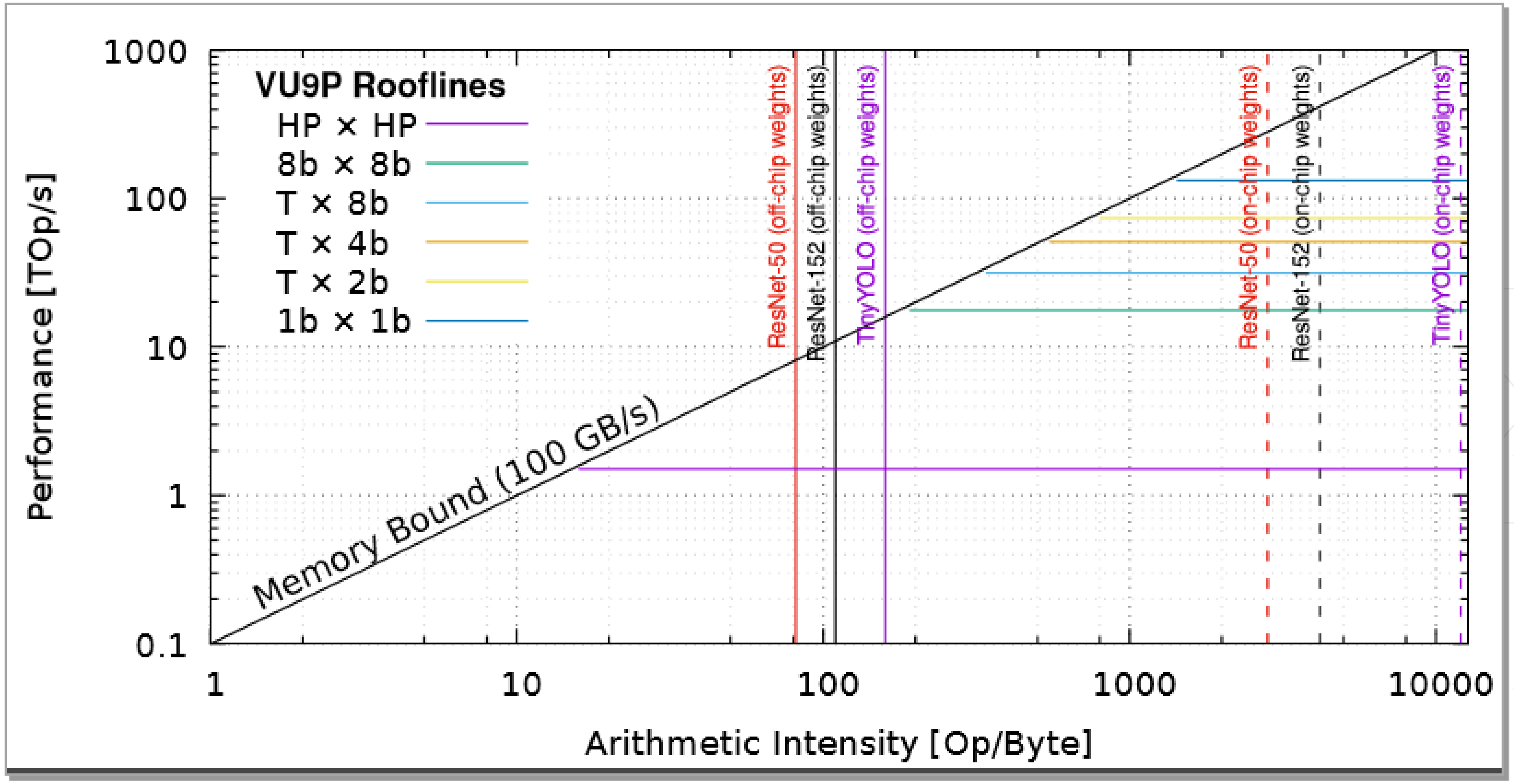
Precision	Modelsize [MB] (ResNet50)
1b	3.2
8b	25.5
32b	102.5



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

Reducing Precision provides Performance Scalability

Example: ResNet50, ResNet152 and TinyYolo



Theoretical Peak Performance for a VU9P with different Precision Operations

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

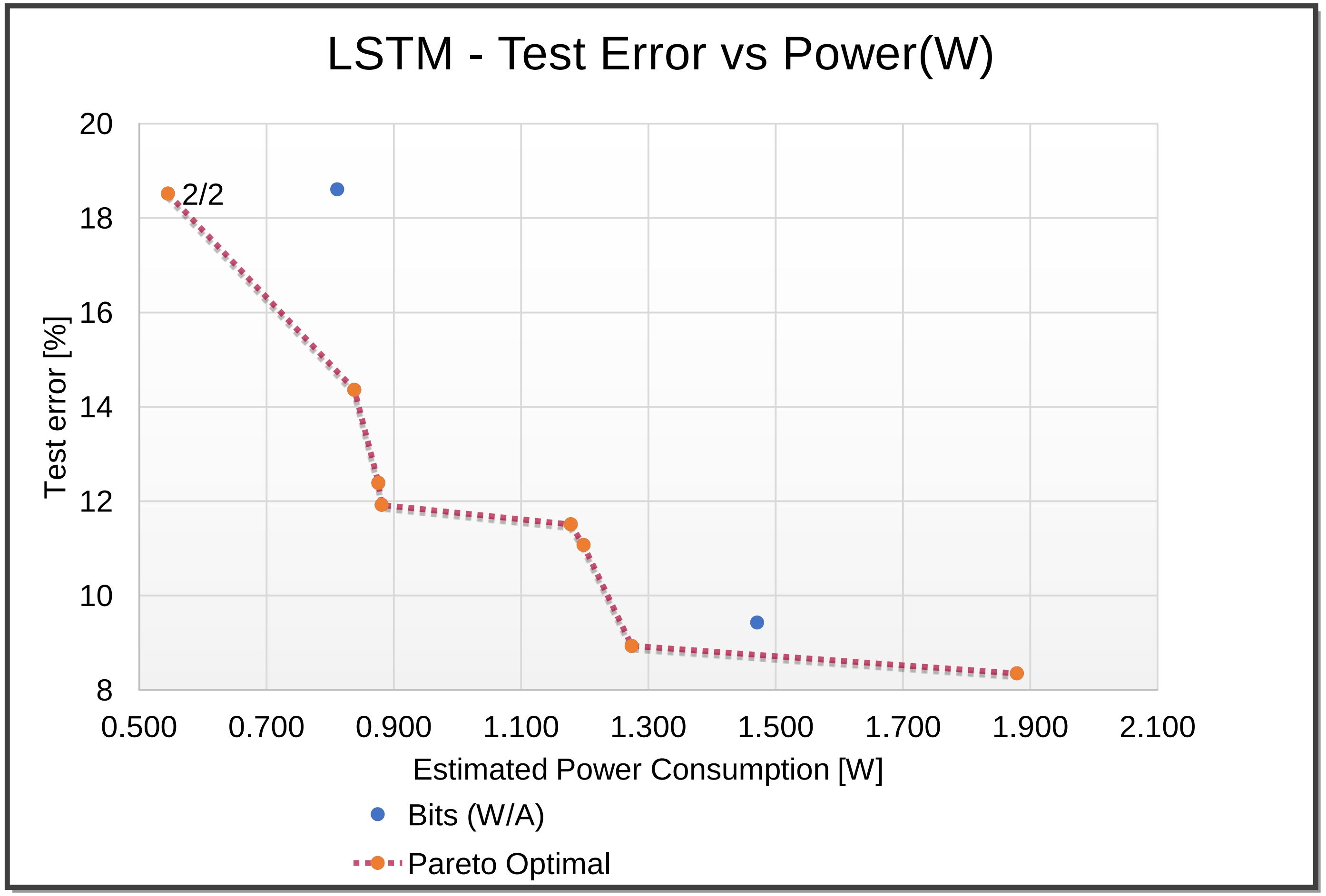
RP scales compute performance

RP reduces model size=> to stay on-chip

Up to 100x

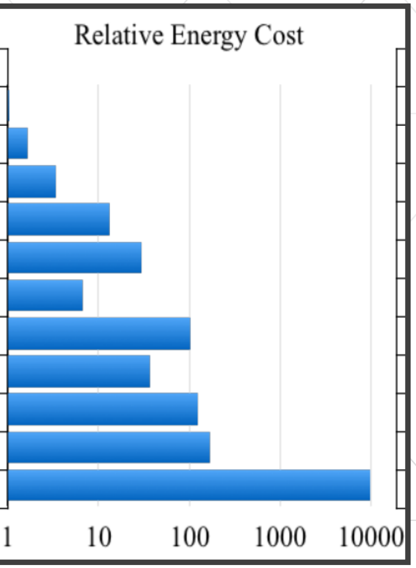
XILINX

Reduced Precision Inherently Saves Power



Target Device ZU7EV ● Ambient temperature: 25 °C ● 12.5% of toggle rate ● 0.5 of Static Probability ● Power reported for PL accelerated block only

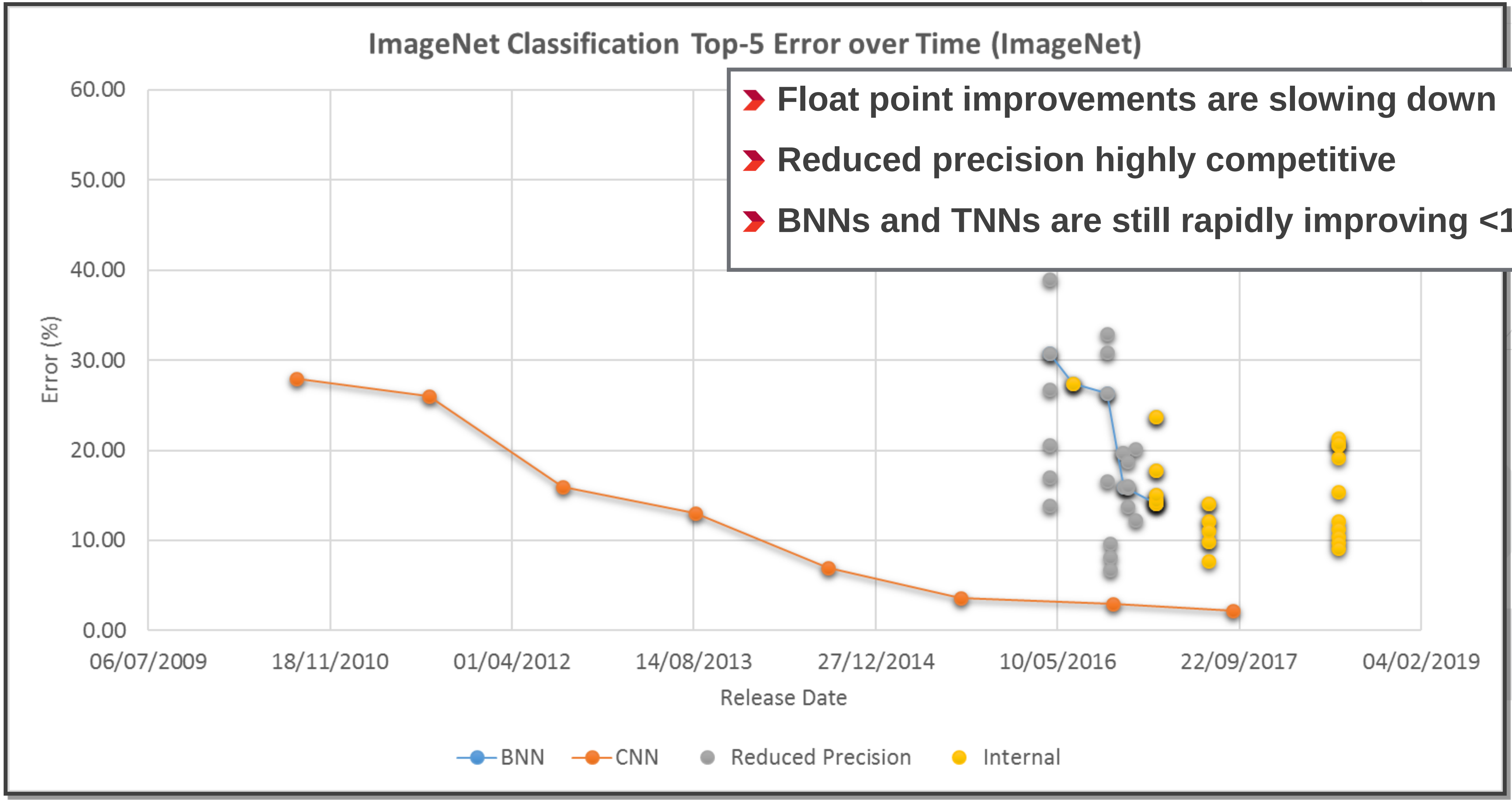
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640



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

What are the downsides of reduced precision?

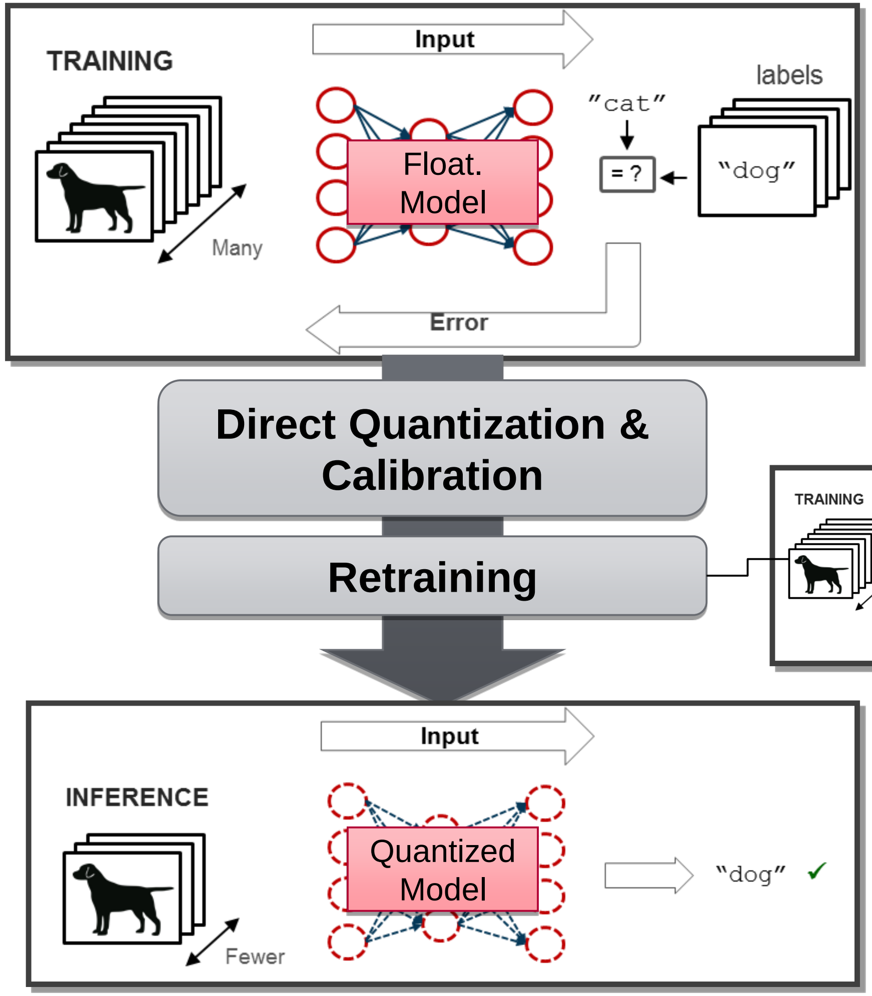
RPNNs: Closing the Accuracy Gap

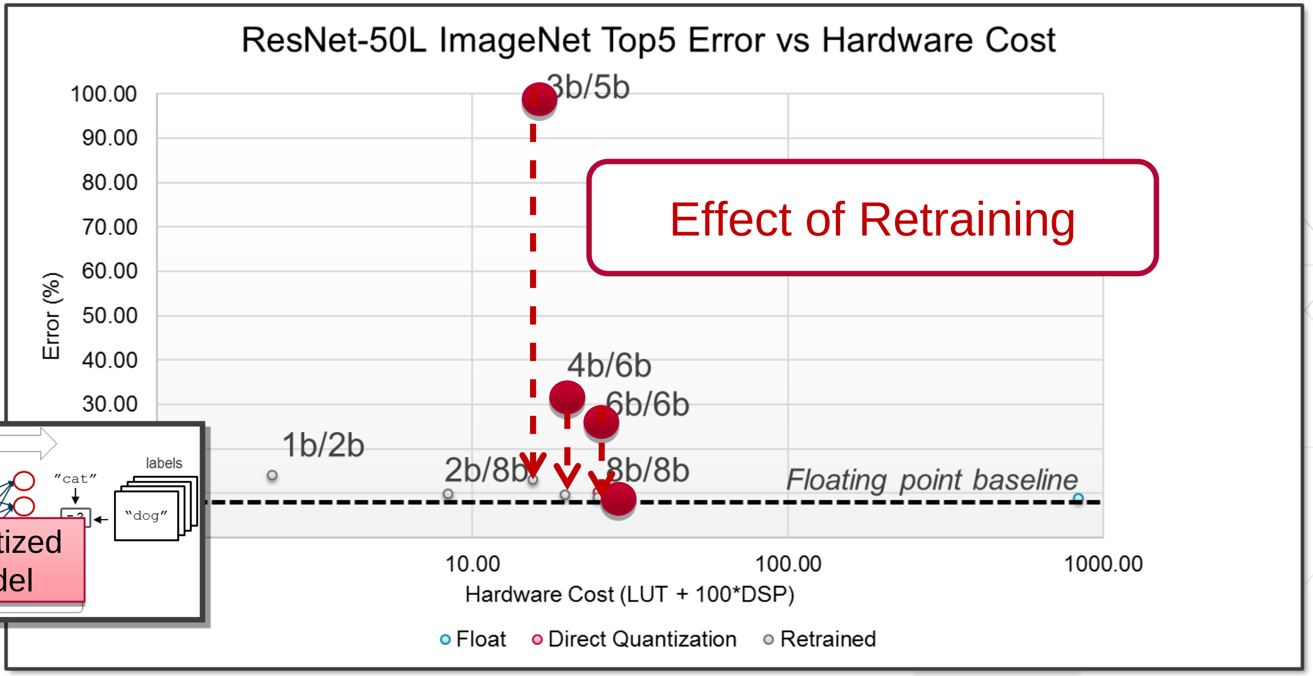


- Float point improvements are slowing down
- Reduced precision highly competitive
- BNNs and TNNs are still rapidly improving <10% top5

XILINX

Retraining:
From Floating Point to Reduced Precision NNs

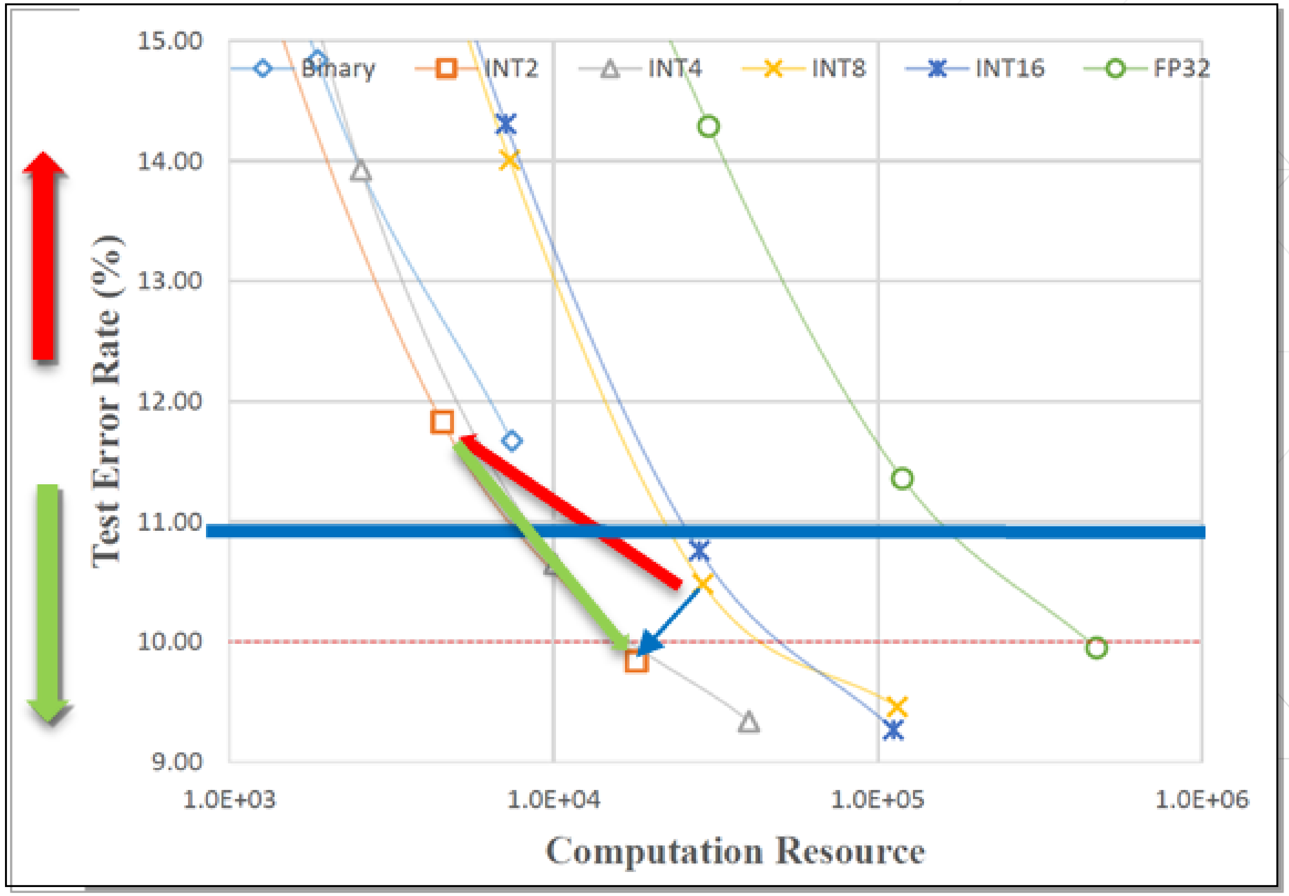




- **Direct quantization & calibration**
  - **Deploying a different model to the one we trained**
  - **Works surprisingly well for 8b**
- **<8bit: retraining helps a lot, but takes time**

How to recuperate accuracy?

- > **Recuperate accuracy by increasing network size**
- > **Topological changes**
- > **New training techniques**
  - >> Knowledge distillation



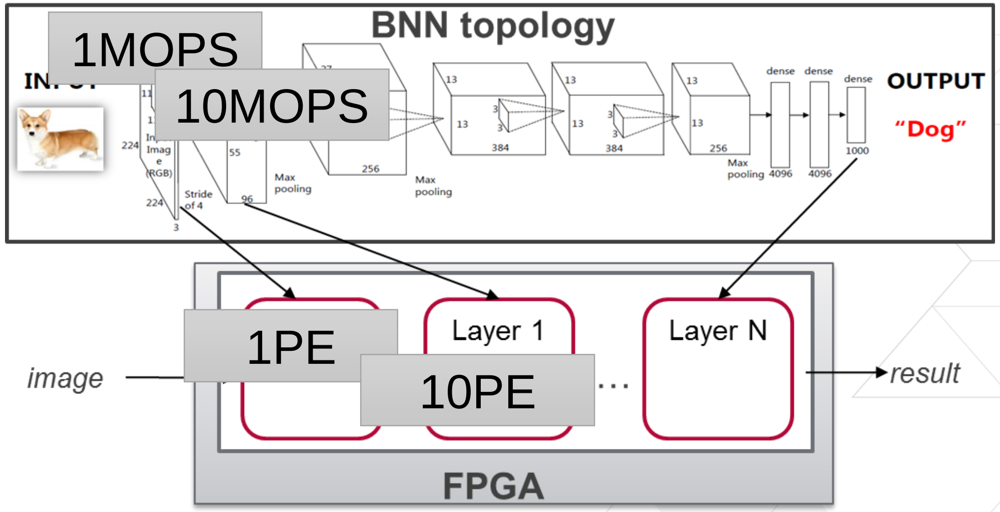
Automating & Customization

FINN: Custom-Tailored Hardware Architectures

> Customized feed-forward dataflow architecture to match network topology

> Customized to meet design requirements

> Customized data types (n-bit)



Automatically generated from CNN description

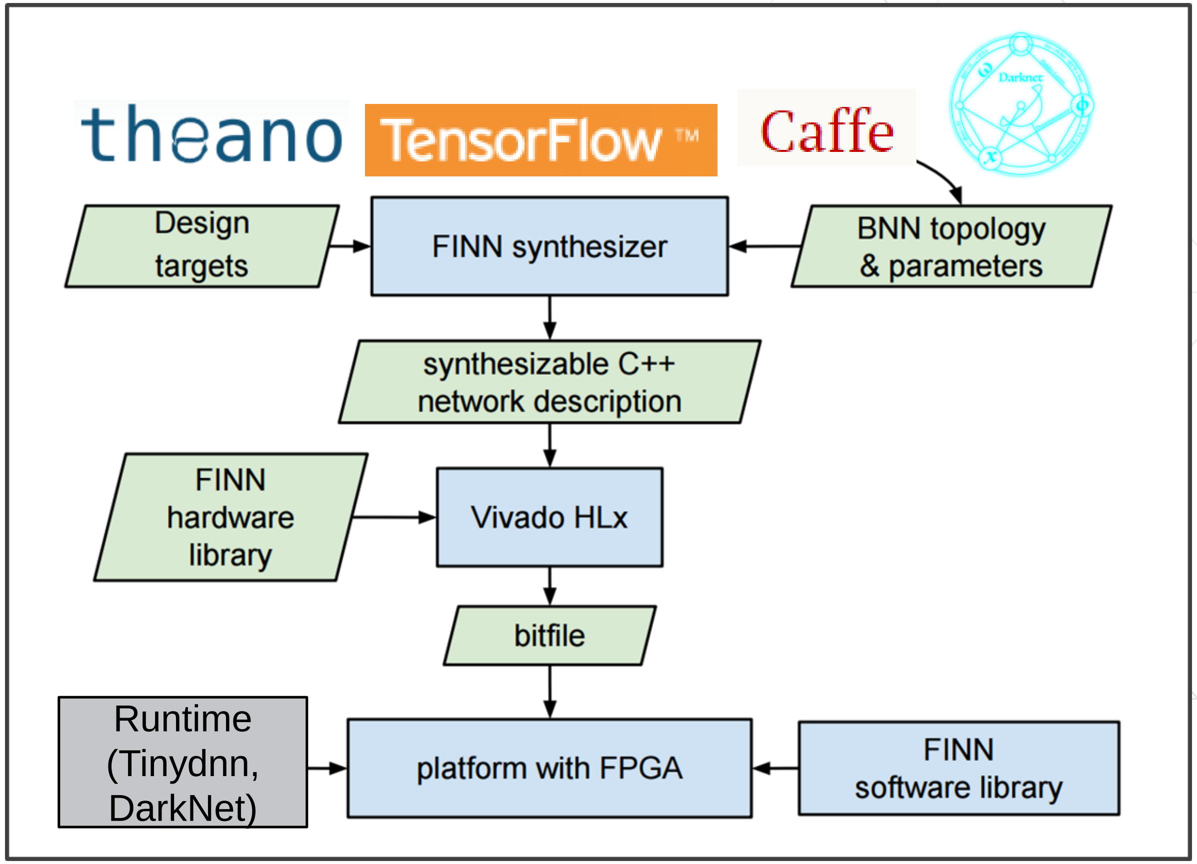
> Uses a synthesizable C++ NN description

> Enables flexibility & scalability and supports portability, rapid exploration

Synthesizable CNN Description

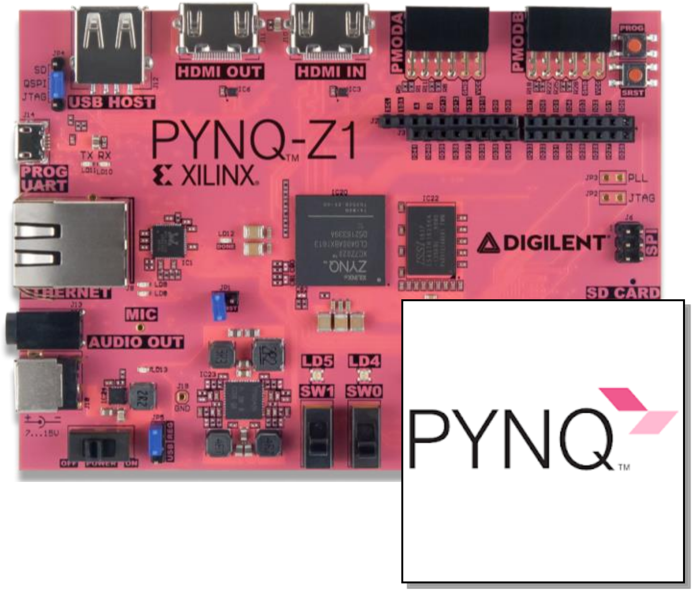
```
void DoCompute(ap_uint<64> * in, ap_uint<64> * out) {
#pragma HLS DATAFLOW
   stream<ap_uint<64> > memInStrm("memInStrm");
   stream<ap_uint<64> > InStrm("InStrm");
          .
          .
          .
   stream<ap_uint<64> > memOutStrm("memOutStrm");

   Mem2Stream<64, inBytesPadded>(in, memInStrm);
   StreamingMatrixVector<L0_SIMD, L0_PE, 16, L0_MW, L0_MH, L0_WMEM, L0_TMEM>
          (InStrm, inter0, weightMem0, thresMem0);
   StreamingMatrixVector<L1_SIMD, L1_PE, 16, L1_MW, L1_MH, L1_WMEM, L1_TMEM>
          (inter0, inter1, weightMem1, thresMem1);
   StreamingMatrixVector<L2_SIMD, L2_PE, 16, L2_MW, L2_MH, L2_WMEM, L2_TMEM>
          (inter1, inter2, weightMem2, thresMem2);
   StreamingMatrixVector<L3_SIMD, L3_PE, 16, L3_MW, L3_MH, L3_WMEM, L3_TMEM>
          (inter2, outstream, weightMem3, thresMem3);
    StreamingCast<ap_uint<16>, ap_uint<64> >(outstream, memOutStrm);
    Stream2Mem<64, outBytesPadded>(memOutStrm, out);
}
```

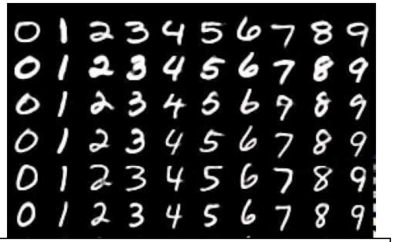


XILINX

Numerous Platforms – From Embedded to Cloud



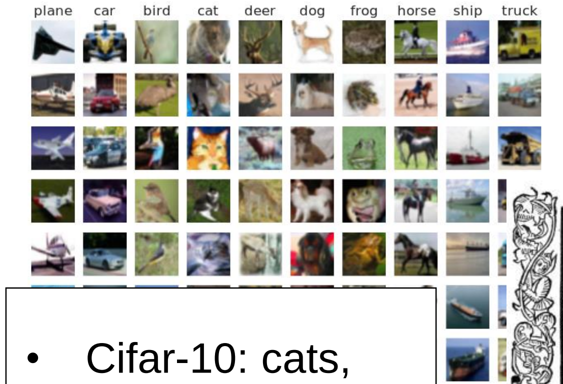
Numerous Platforms



- MNIST handwritten digits

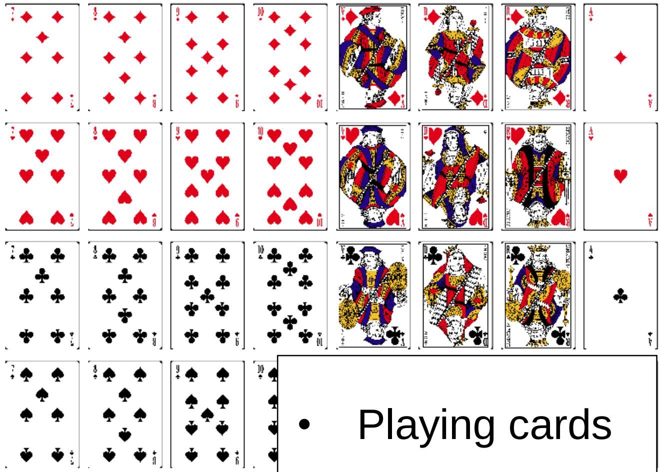
- Streetview house numbers

- German road signs

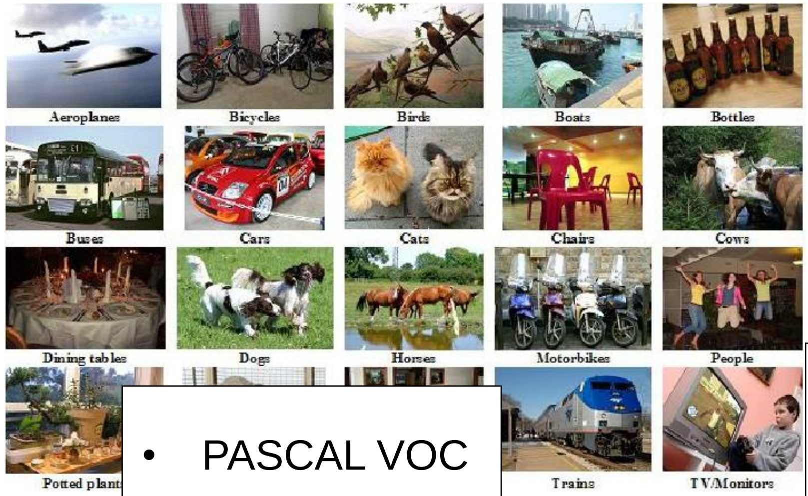


- Cifar-10: cats, dogs, etc

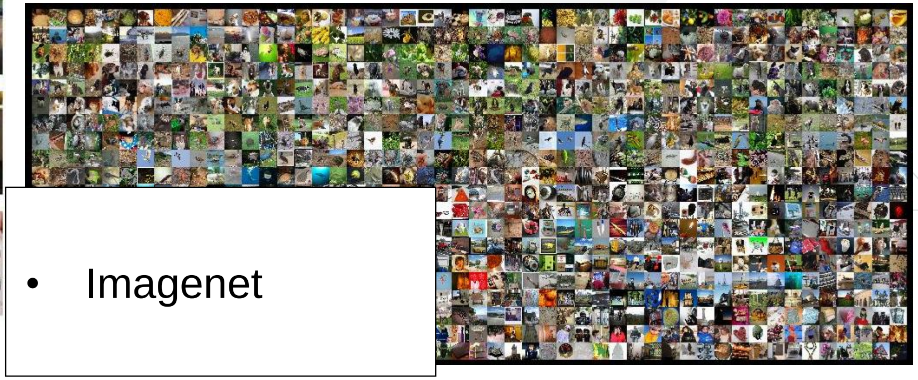
- Fraktur



- Playing cards

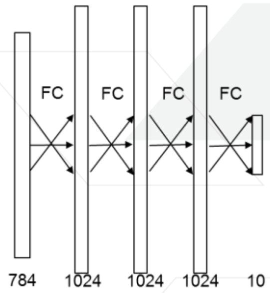


- PASCAL VOC

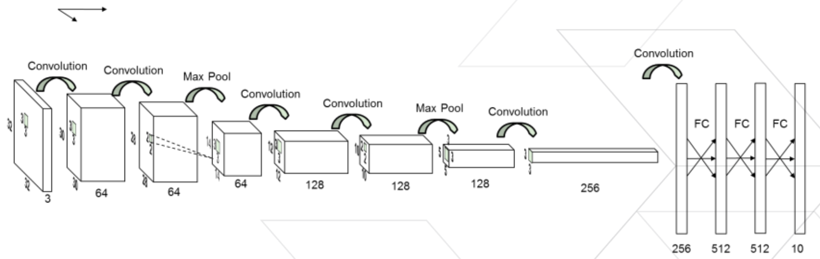


- Imagenet

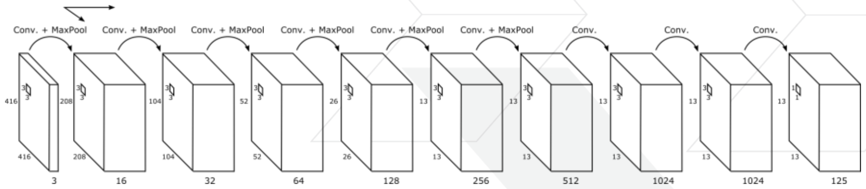
Numerous Test Networks



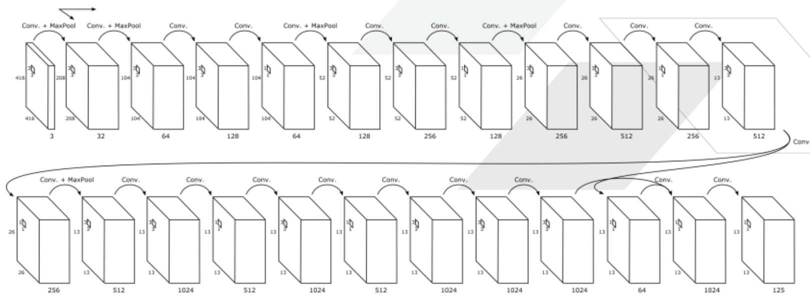
- Multilayer Perceptron (1b weights, 1b act), MNIST
  - Up to 5.8MOPS/frame



- **VGG-16 derivative (1b weights, 1b act), SVHN, CIFAR-10, traffic signs, playing cards**
  - Up to 1.2GOPS/frame



- **DorefaNet – AlexNet derivative (mostly 1b weights, 2b act) (ImageNet)**
  - Up to 3.9GOPS/frame



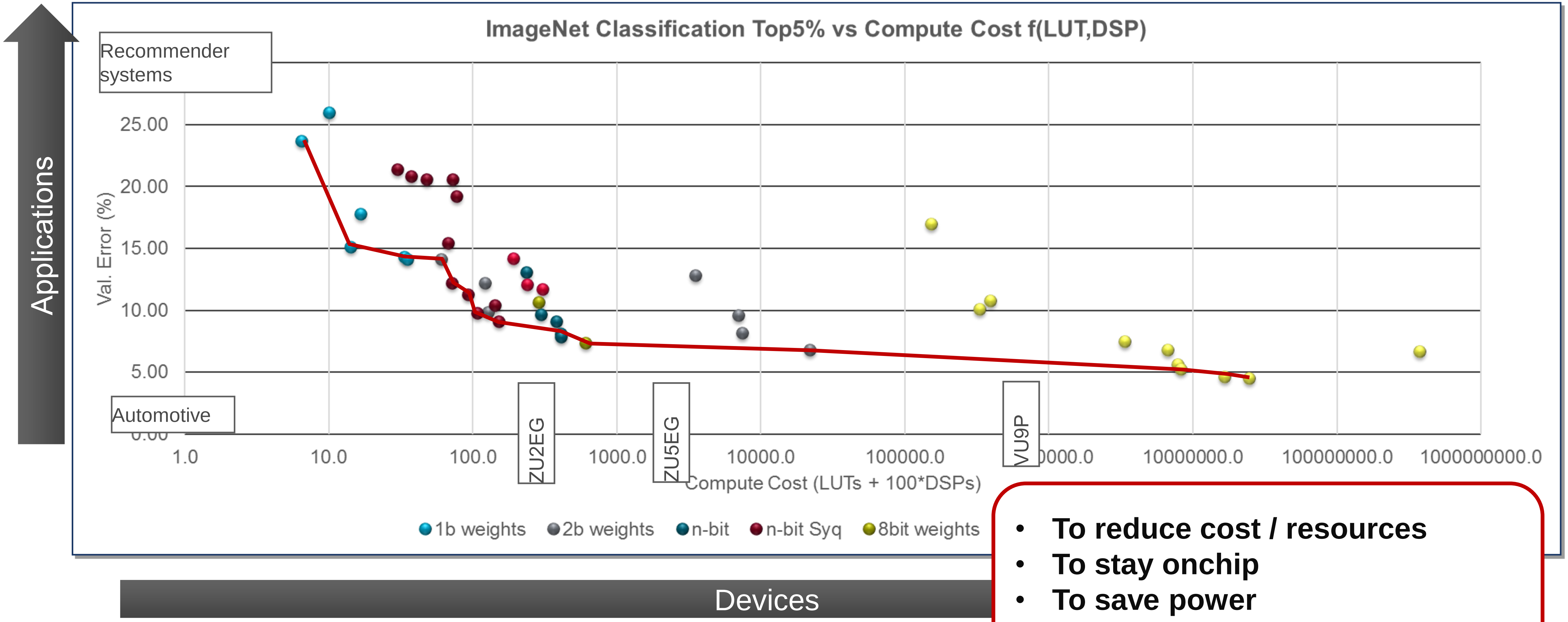
- **YoloV2, Yolo9000, TinyYolo (1b weights, 8b act) (VOC, COCO)**
  - 34.9, 19 and 7.0GOPS/frame

- **LSTM, for OCR on Fraktur**

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

XILINX

Design Trade-offs with Reduced Precision NNs



- To reduce cost / resources
- To stay onchip
- To save power
- To scale performance

FINN Results

> **Performance**
  - VOC Object recognition: Quantized TinyYolo @ **55fps @ 7Watt** (batch=1) for embedded (ZU3EG)
  - ImageNet Classification: Dorefanet @ **11 TOPS on AWS F1** instance
  - Scaled binary operations to **51TOPS on AWS F1** and **5.2 TOPS on ZU3EG & 1000x over Raspberry Pi**

> **Energy efficiency: measured 433GOPS/Watt**

> **Flexibility & Scalability**
  - Different platforms can easily be targeted from embedded to cloud
  - Different use cases, networks & training data sets

> **While being sufficiently accurate**
  - <10% top5 for ImageNet classification

Agenda

Background – Xilinx Research

Machine Learning

Research Efforts

Summary & Outlook

Summary

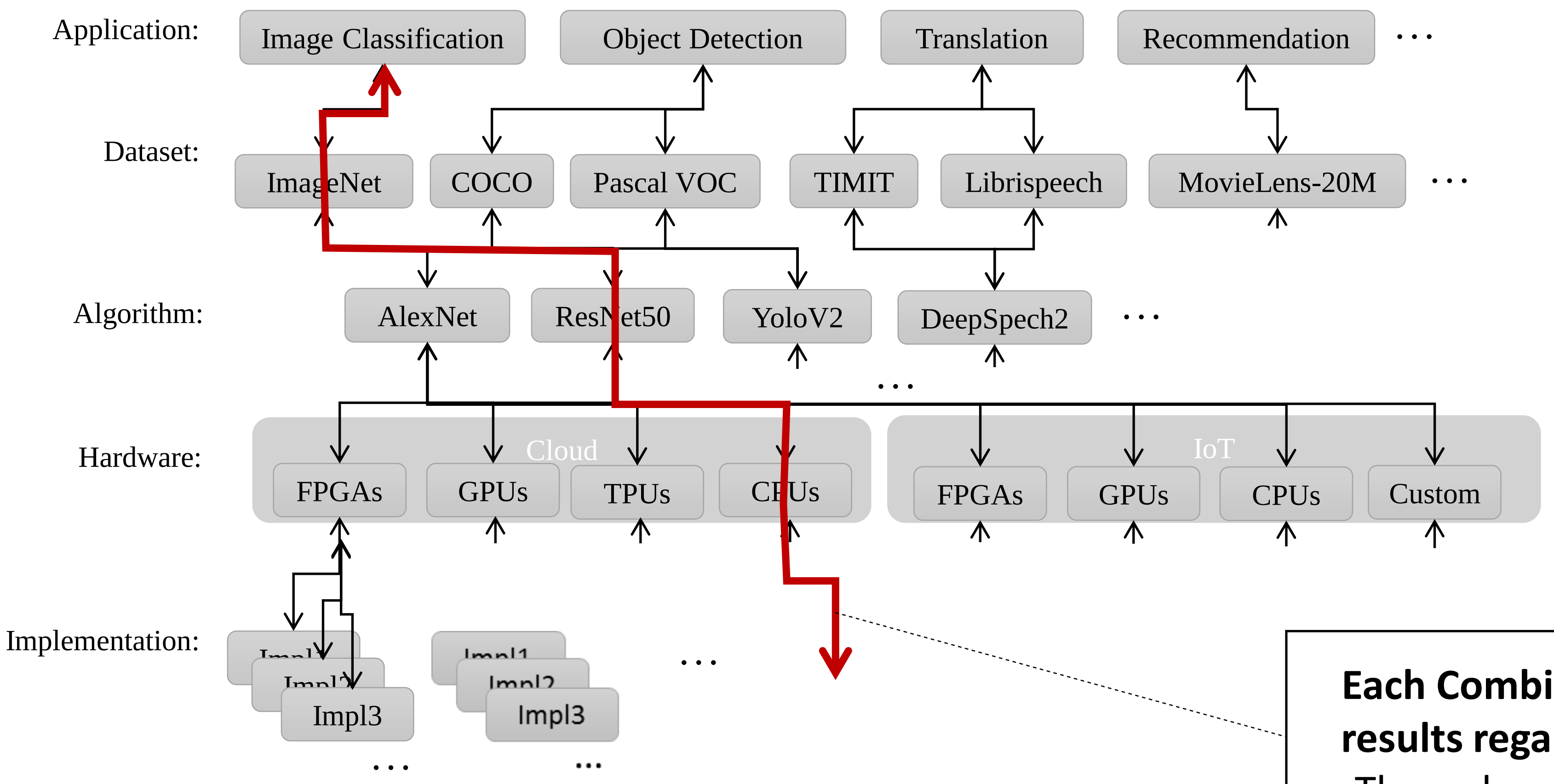
> ML has the potential to address many of the grand engineering challenges of this century

> However, compute & memory requirements are huge and flexibility and scalability are key

> New, customized computer architecture are emerging

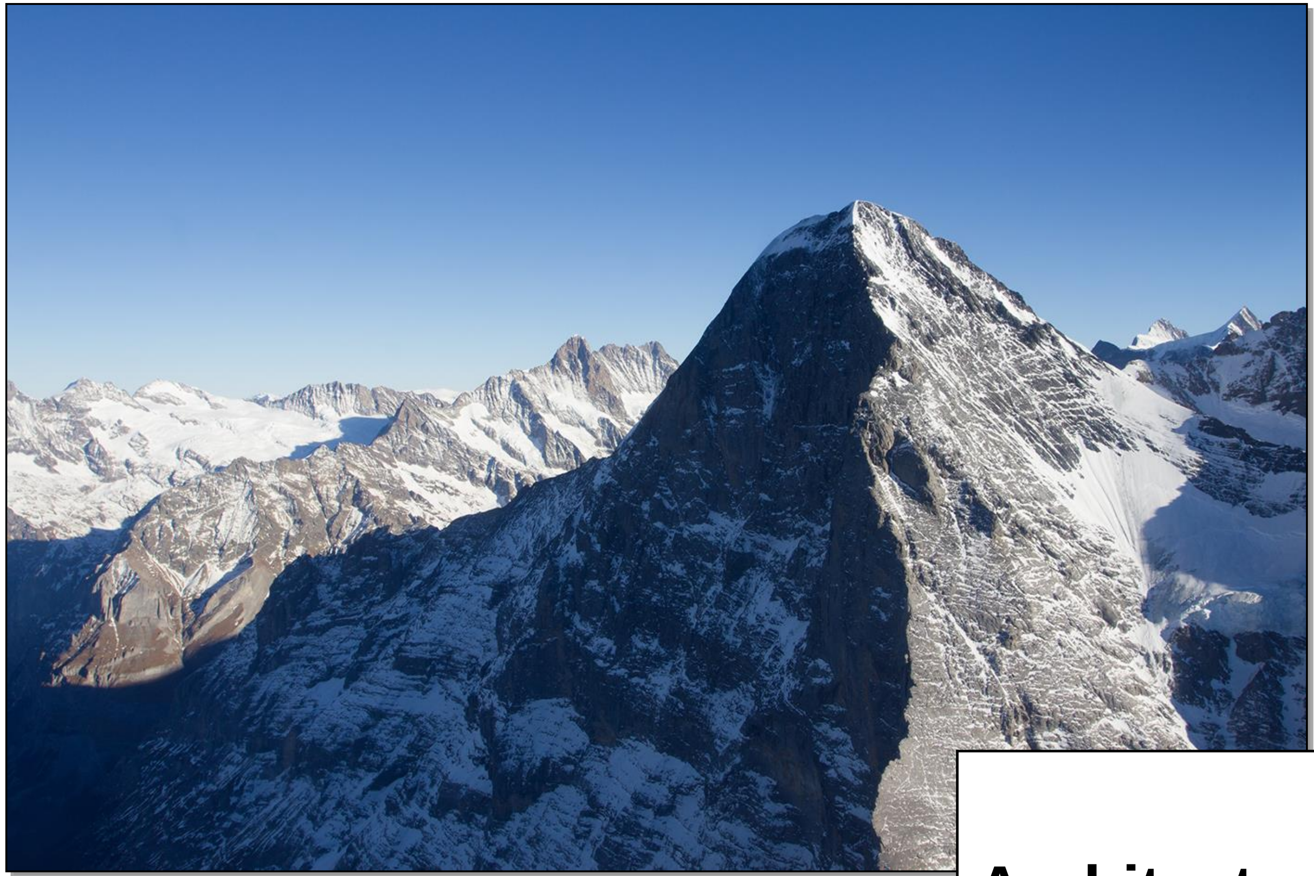
> FPGAs can play an important role here, in particular in conjunction with reduced precision and customized macro architectures
>> Orders of magnitude improvement in performance, resources and power consumption

Exciting Times for our Community:
Finding Optimal Solutions within a Complex Design Space



Each Combination delivers different results regarding the design targets:
Throughput, power, latency, cost,…

Outlook



Architecture Exploration

- **Help understand the choices!**

Adaptable.
Intelligent.

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