The Emerging Computational Landscape of Neural Networks

Michaela Blott
Principal Engineer, Xilinx Research
August 2018

XILINX

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

XILINX

Current Xlabs Dublin Team

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

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

Plus a Very Active Internship Program

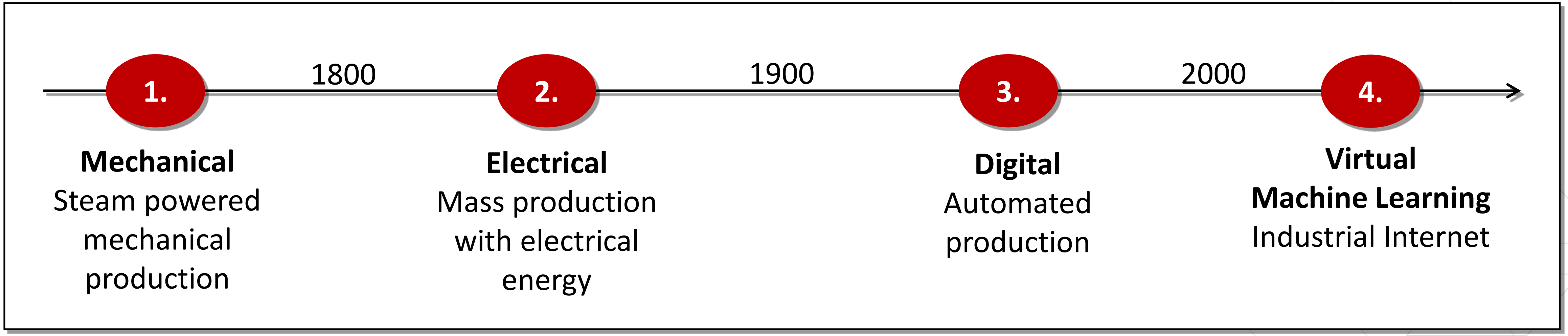
- **On average 4-6 interns at any given time**
  - From top universities all over the world
  - We are always looking for talent ;-)

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

Machine Learning,
Neural Networks & its Challenges

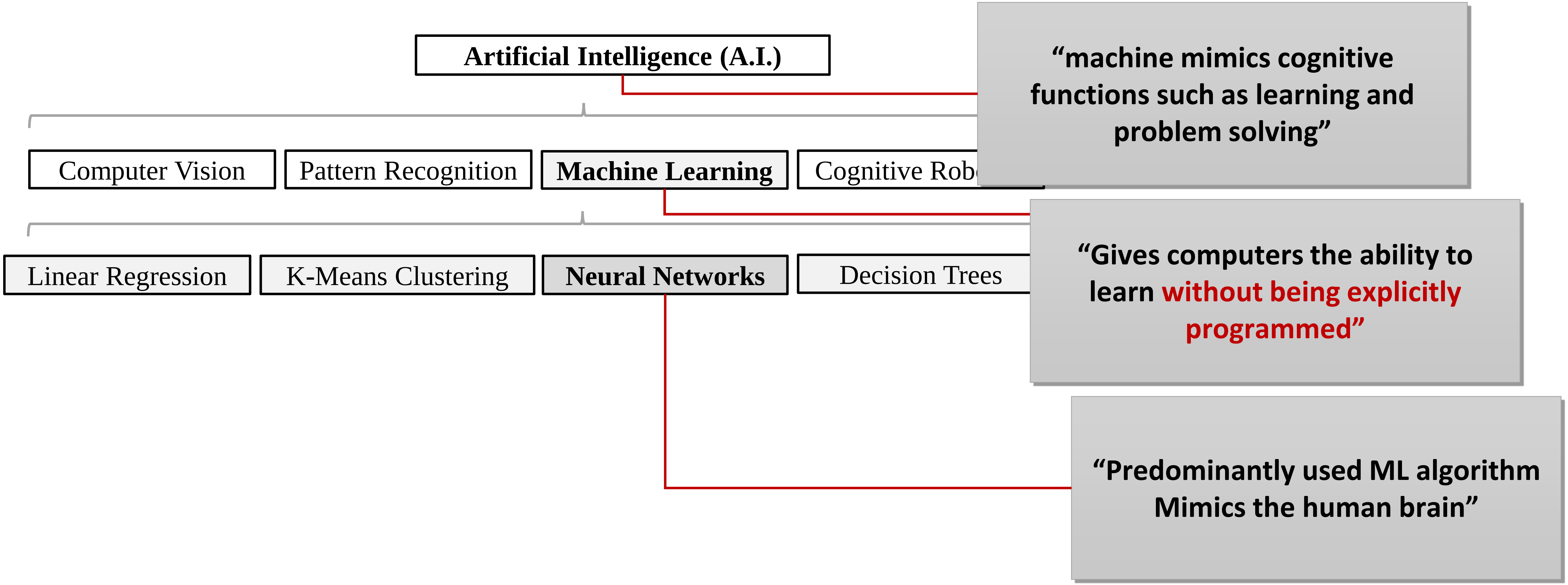
XILINX

The Rise of The Machine (Learning Algorithms)



- **Potential to solve the unsolved problems**
  - Making solar energy economical, reverse engineering the brain (Jeff Dean, Google Brain 2017)
- **Many difficult ethical questions**
  - Will machines destroy jobs? AI apocalypse?
- **History has shown: We are going through cycles of inventions followed by society adjustments**
  - All of this has happened before and will happen again (Battlestar Galactica, 2014)
- **Let’s look at what the technology can do, and how we FPGA designers & computer architects broaden its adoption**

A.I. – Machine Learning - Neural Networks



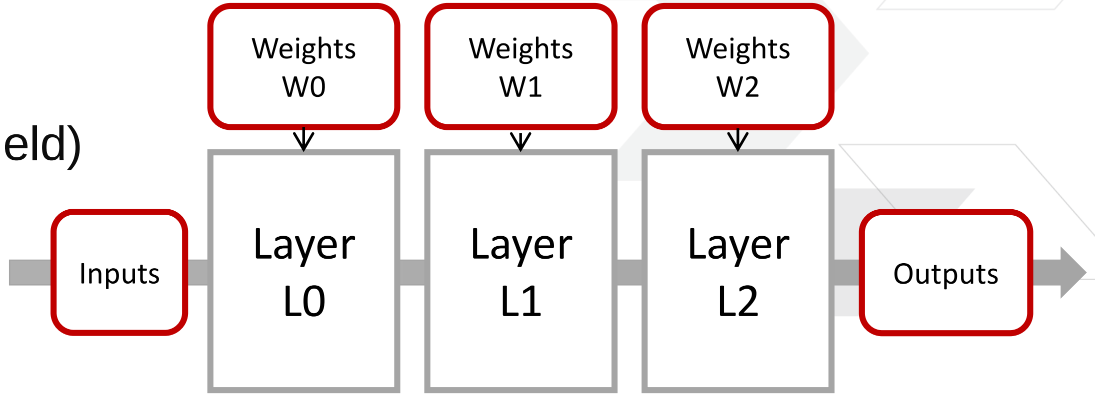
Convolutional Neural Networks (CNNs)
from a computational point of view

- **CNNs are usually feed forward* computational graphs constructed from one or more layers**
  - Up to 1000s of layers

- **Each layer consists of neurons *ni* which are interconnected with synapses, associated with weights *wij***

$$n0 = Act(w00*i0 + w10*i1)$$

- **Each neuron computes:**
  - Typically linear transform (dot-product of receptive field)
  - Followed by a non-linear "activation" function

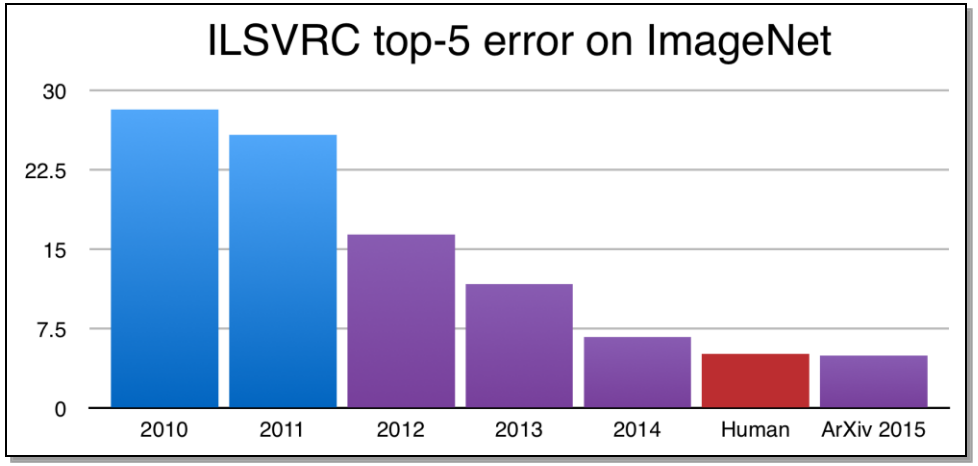


* With exception of RNNs

XILINX

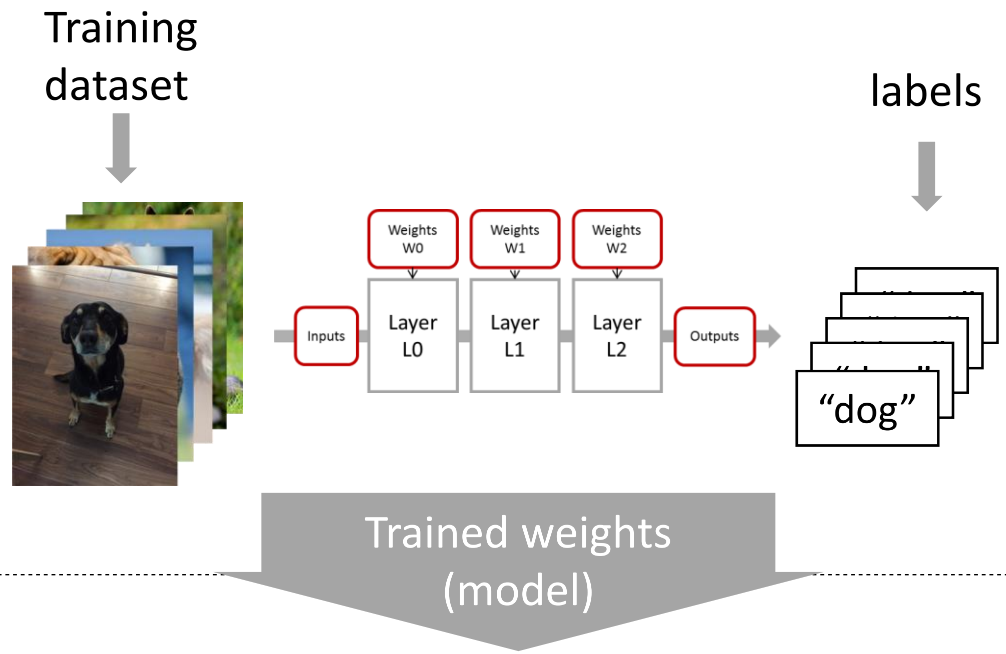
Convolutional Neural Networks (CNNs)
Why are they so popular?

- 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



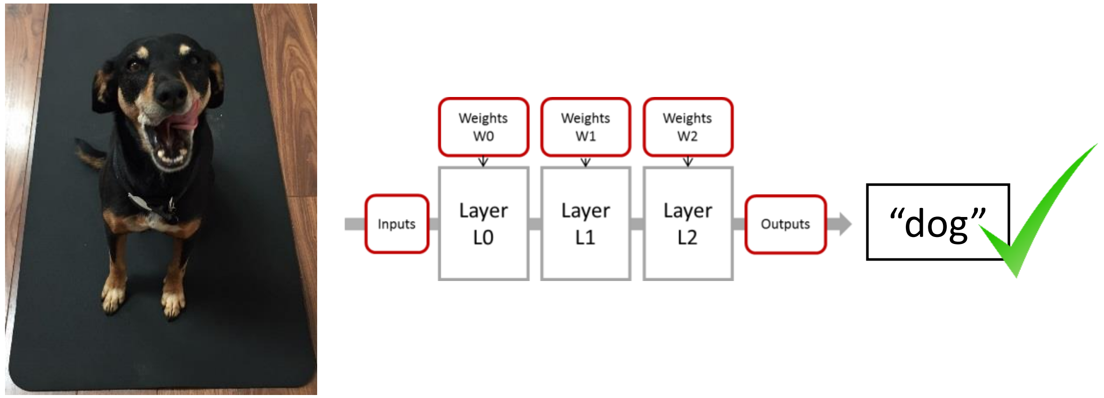
- Will increasingly replace other algorithms
  - unless for example simple rules can describe the problem
- Solve problems previously unsolved by computers
- And solve completely unsolved problems

From Training to Inference



Training
Process for a machine to ***learn*** by optimizing models (weights) from labeled data.

Typically computed in the cloud



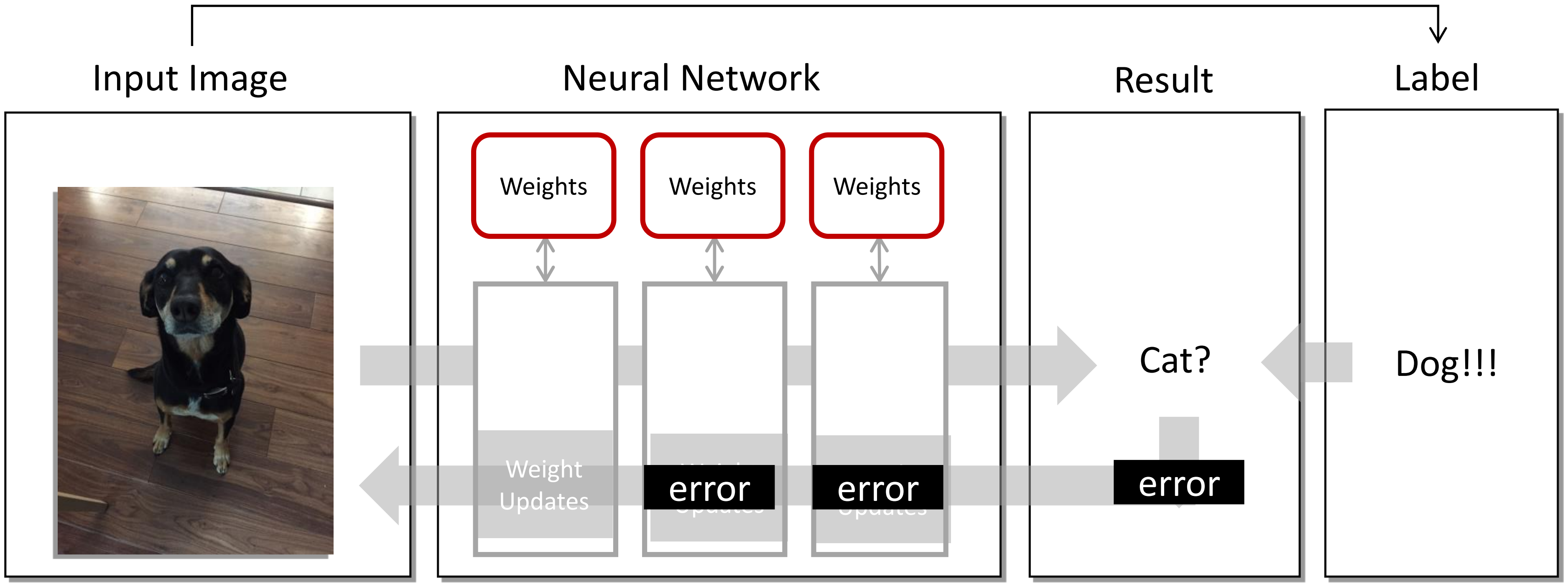
Inference
Using trained models to predict or estimate outcomes from new inputs.

Deployment at the edge

What is the Challenge?

Example: ResNet50

Backpropagation – 1 Image



For ResNet50:

23 Billion operations

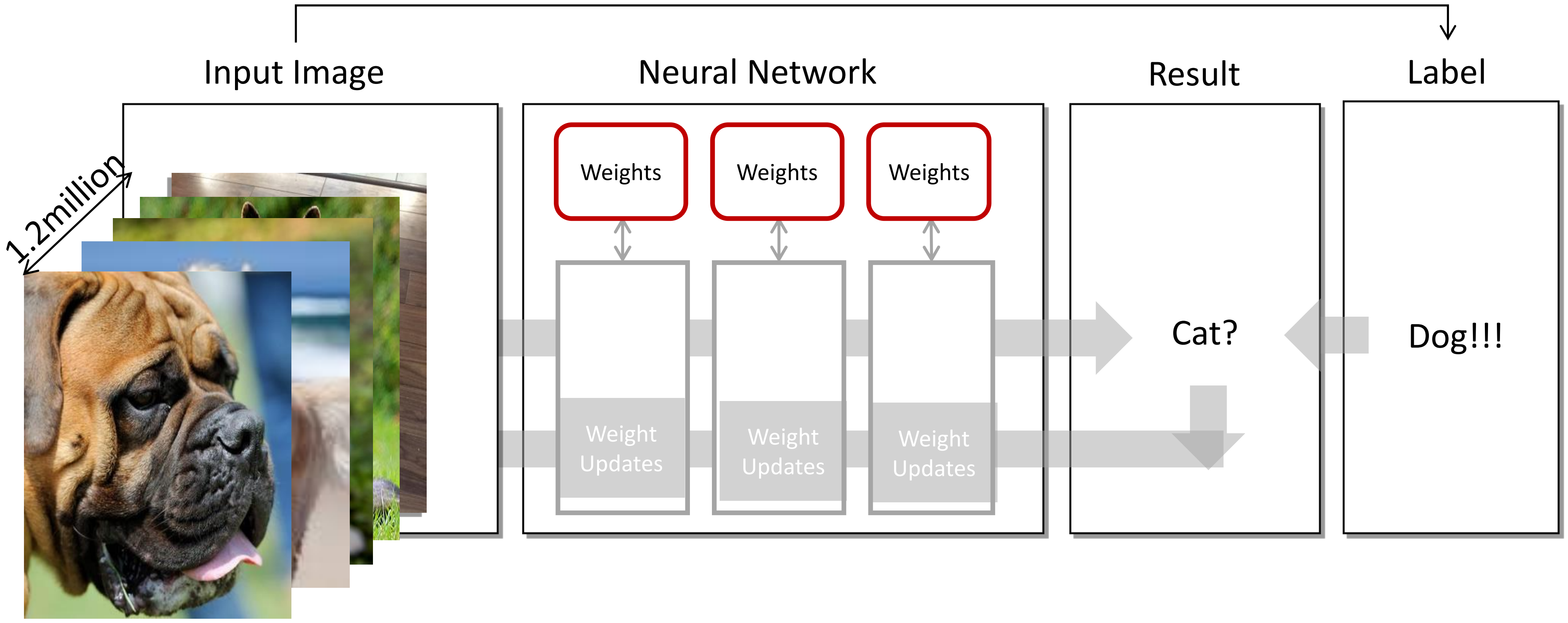
weights, weight gradients, updates: 303MBytes of storage (3-5x)

activations, gradients: 80 MBytes

**Assuming 32b SP*

Example: ResNet50

Training – 1.2 Million Images for 1 epoch

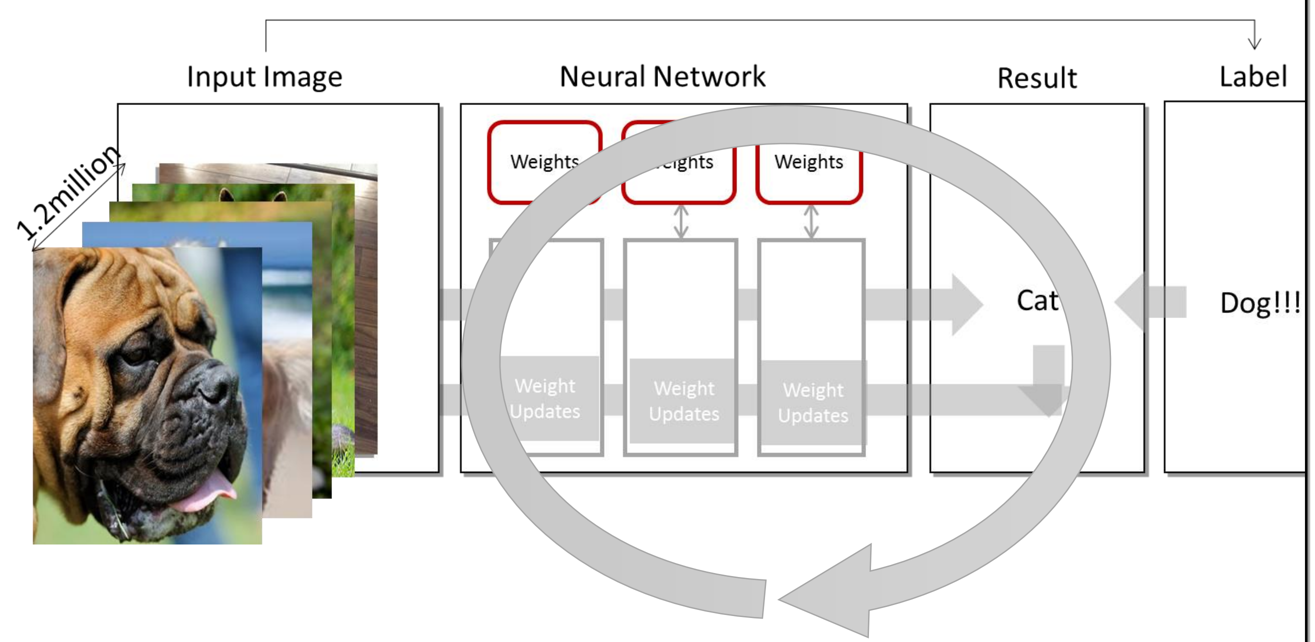


For ResNet50: 1 epoch takes 1.2M * 23 Billion operations = 23 * 10^{15} operations (peta)

XILINX

Example: ResNet50

Training – Approximately 100 Epochs



For ResNet50: $100 * 23\ 10^{15} = 2.3 * 10^{18}$ (exa)

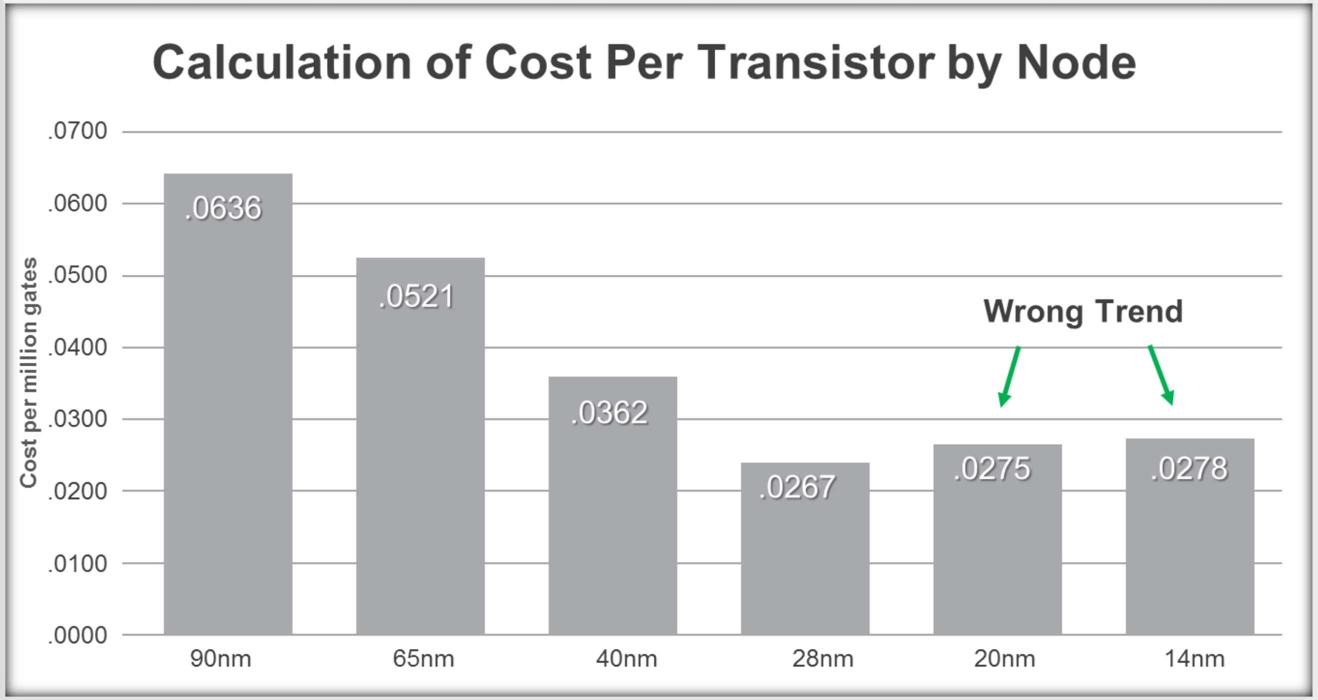
Single P40 GPU (12TFLOPS): 11days @ 100%, usually ~2 weeks

ResNet50:

- **For inference: Billions of operations, and 10s of MegaBytes**
- **For training: Quintillions/Exa of operations, and 100s of MegaBytes**

XILINX

Challenge 1



Source: IBS

- Huge amount of compute and memory
- While compute performance is no longer scaling and becomes more expensive

What else?

Many Applications Require Different Networks

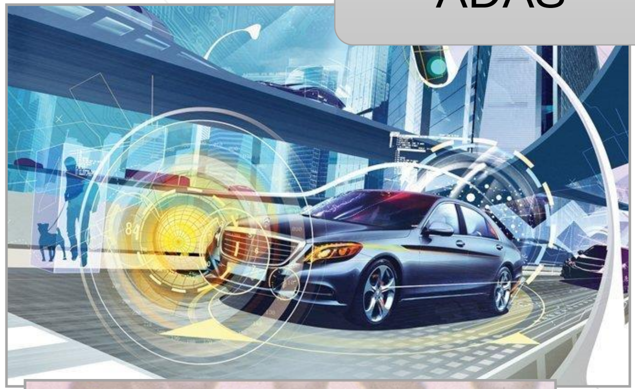
Translation Service

AlphaGo

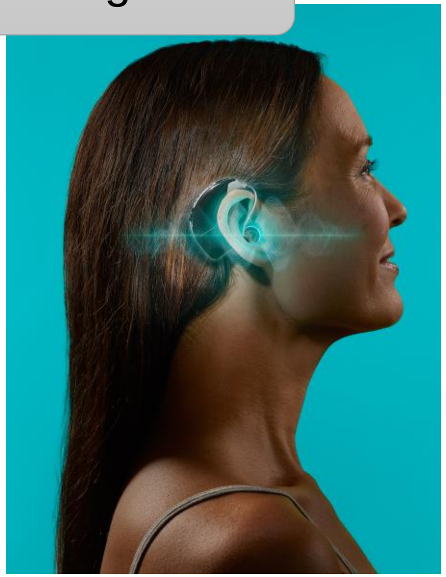
Gaming strategy

3D reconstruction from drone images

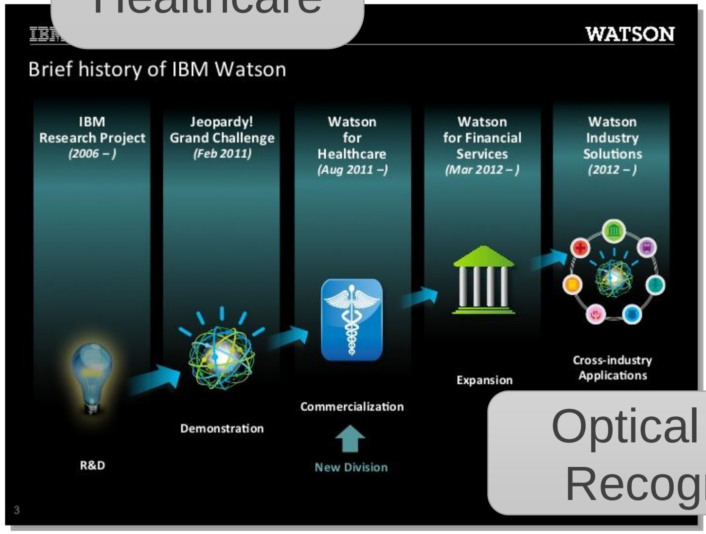
ADAS



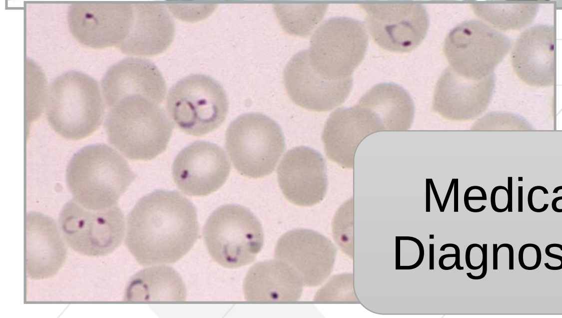
Hearing Aids



Data Analysis for Healthcare



Real-time, sensor-based-control



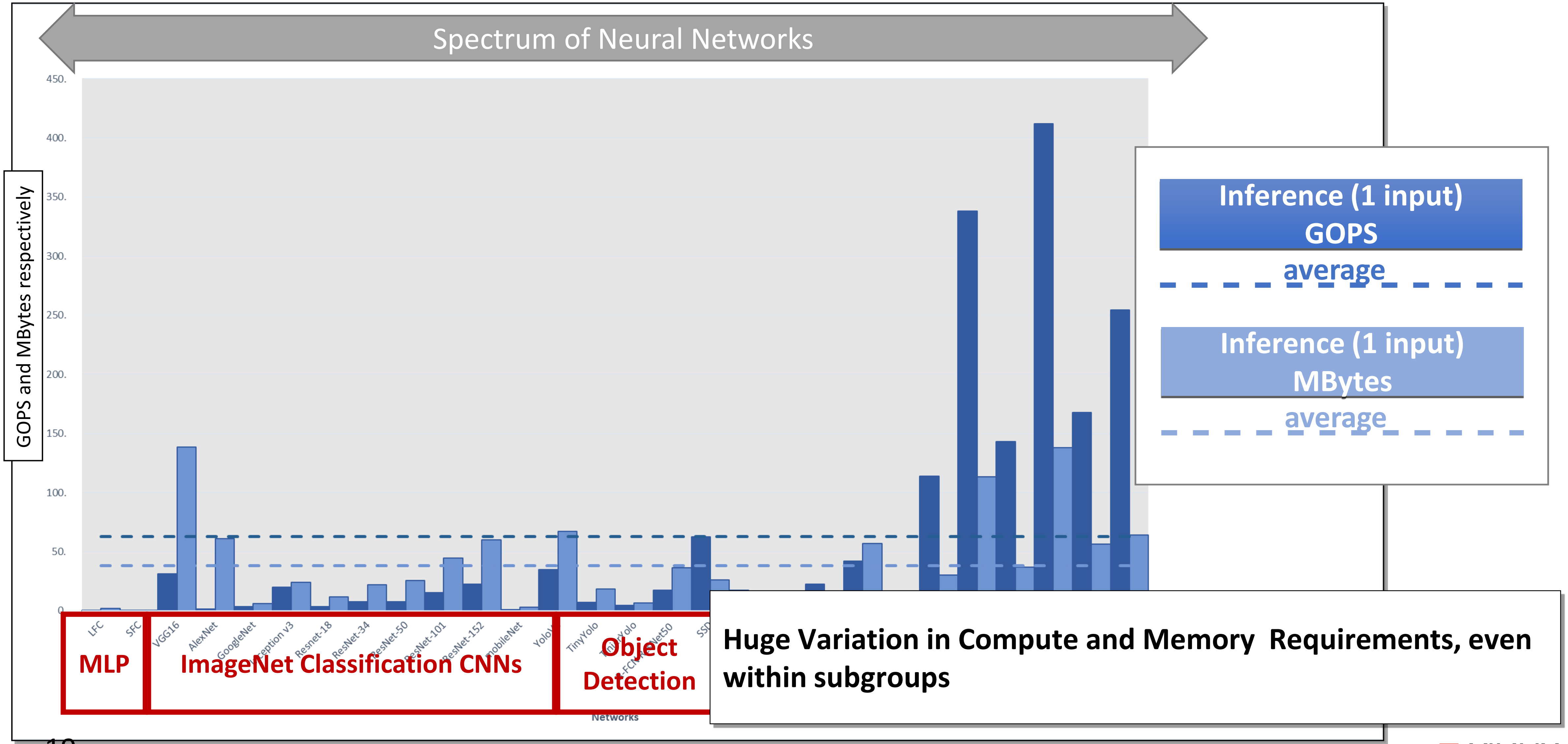
Medical Diagnoses

Optical Char. Recognition

Recommender Systems

Challenge 2: Inference Compute and Memory
Variation Across a Spectrum of Neural Networks

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

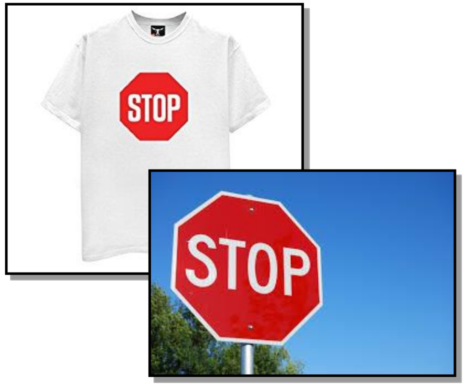


Huge Variation in Compute and Memory Requirements, even within subgroups

Anything else?

Challenge 3:
Different Use Cases, Different Design Targets

Accuracy, speed, power, latency, cost



- **ADAS:**
  - Accuracy
  - High throughput

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

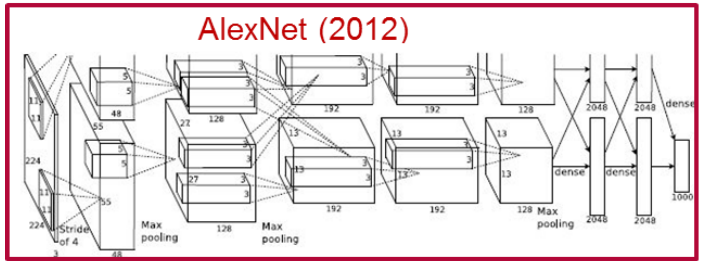
- **AR**
  - High throughput
  - Low latency
  - Low power

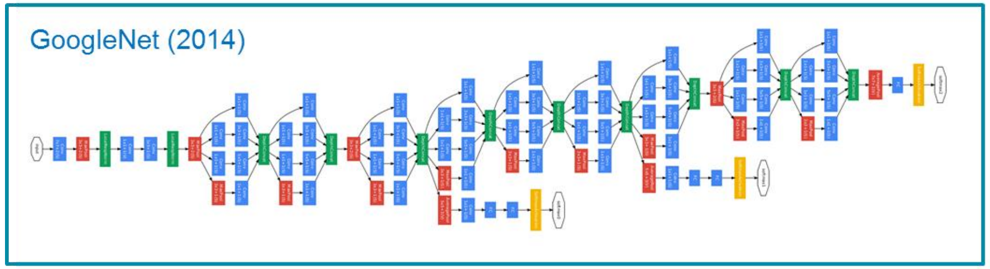
- **3D reconstruction of HR images**
  - High throughput
  - Offline

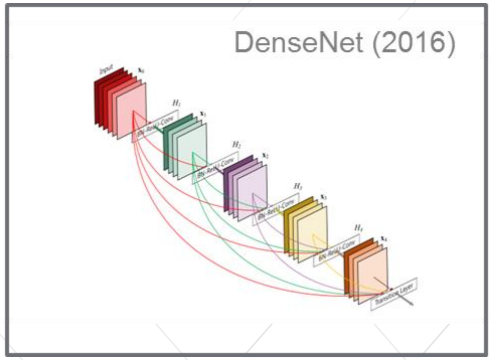
Finally,…

Challenge 4:
Neural Networks Change @ Increasing Rate

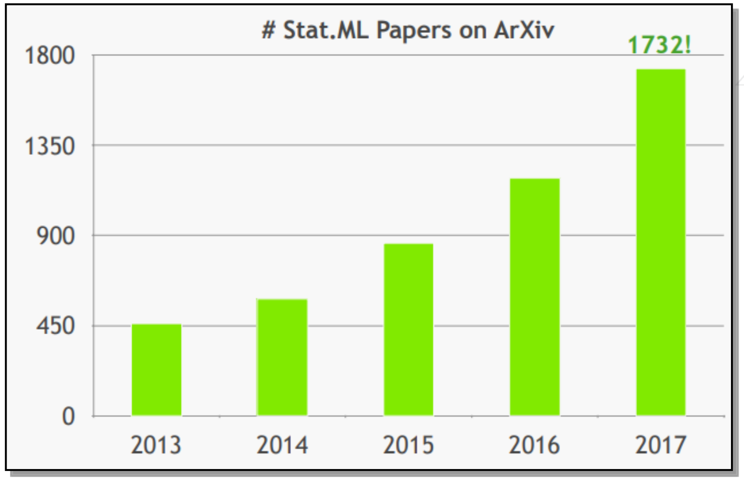
> Graph connectivity, number and types of layers are changing







> Increasing stream of research



Ce Zhang, ETH Zurich, Systems Retreat 2018

XILINX

In Summary: CNNs are associated with…

> **Significant amounts of memory and computation**

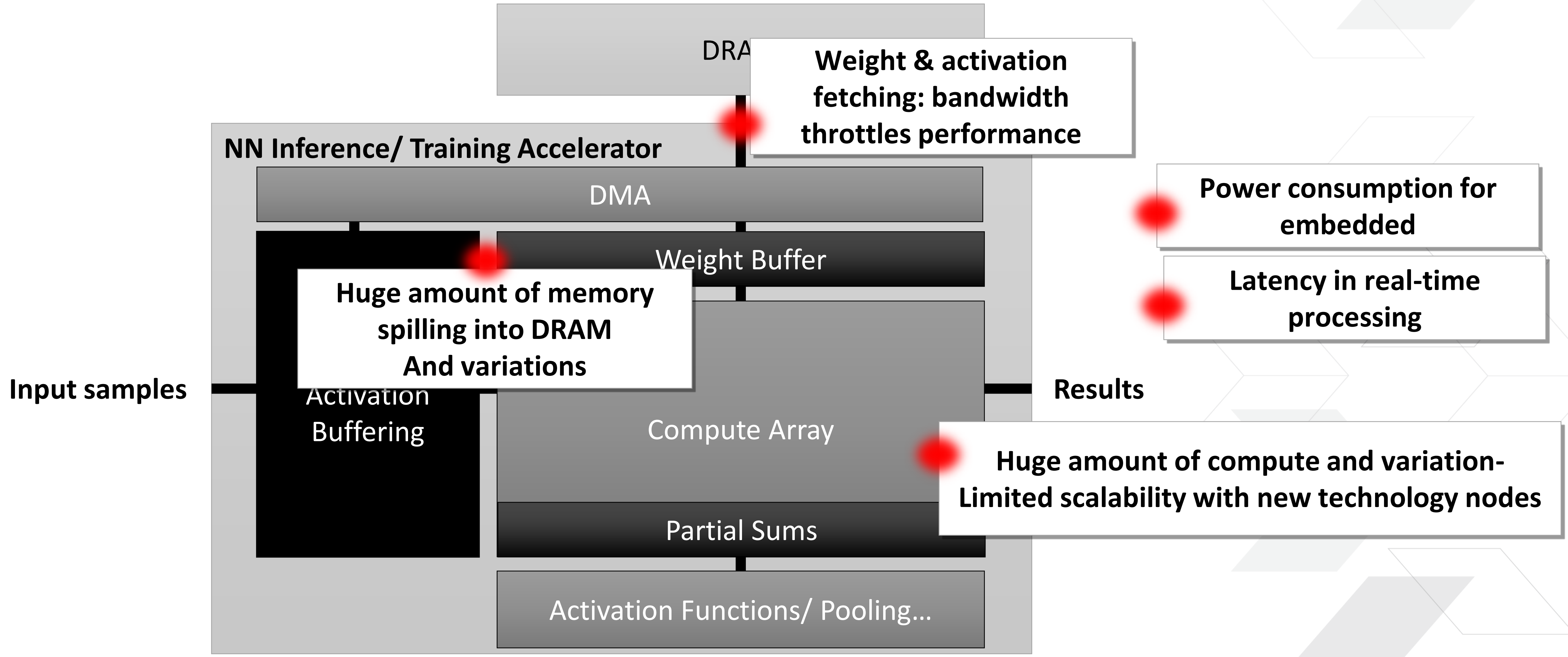
> **Huge variation between topologies and within them**

> **Broad spectrum of applications with different design targets**

> **Fast changing algorithms**

> **However, incredibly parallel!**
>> For convolutions: filter dimensions, feature map dimensions, input & output channels, batches, layers, and even precisions

Architectural Challenges/ Pain Points

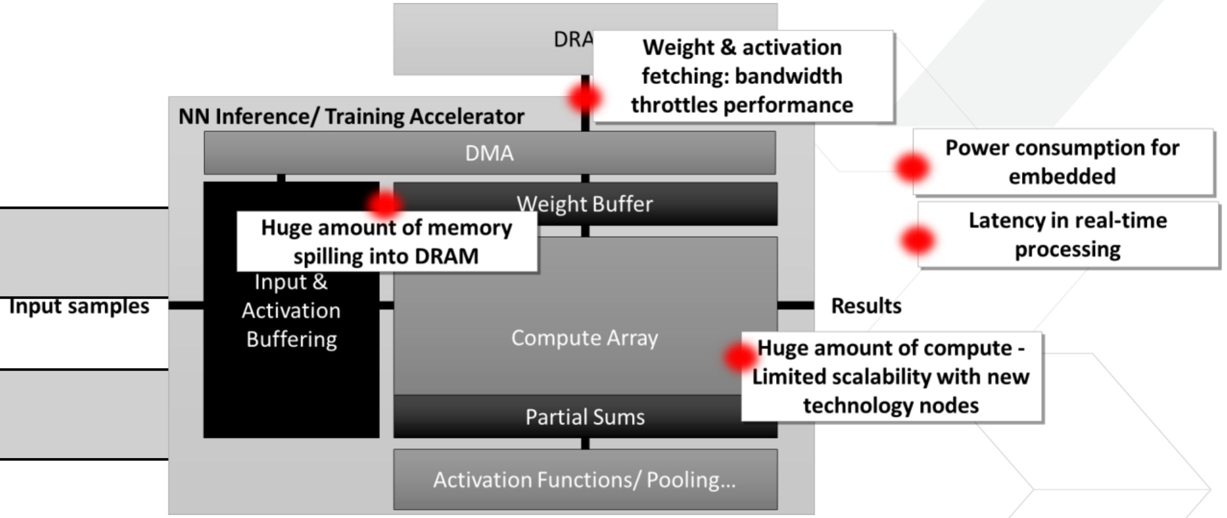


Requires algorithmic & architectural innovation

XILINX

Algorithmic Optimization Techniques

Optimization Techniques



- **Loop transformations to minimize memory access***
- **Pruning**
- **Compression**
- **Winograd, Strassen and FFT**
- **Novel layer types (squeeze, shuffle, shift)**
- **Numerical Representations & Reducing Precision**

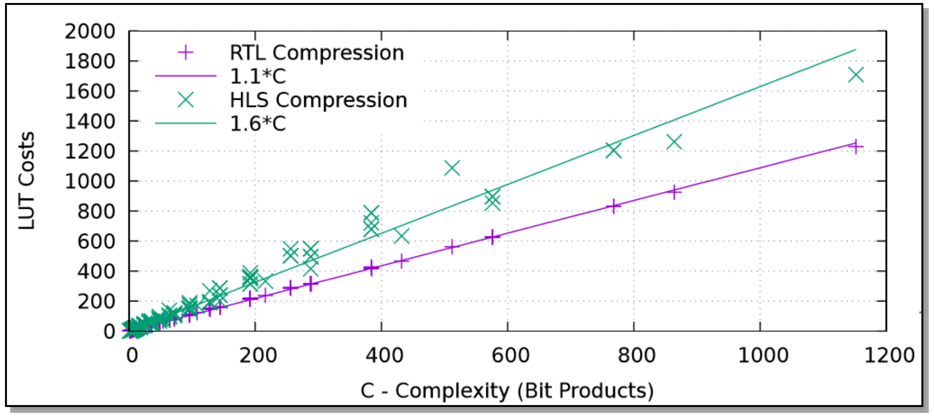
**Chen, Y.H., Krishna, T., Emer, J.S. and Sze, V., 2017. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE Journal of Solid-State Circuits, 52(1), pp.127-13*

Example: Reducing Bit-Precision

- **Linear reduction in memory footprint**
  - Reduces weight fetching memory bandwidth
  - NN model may even stay on-chip

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

- **Reducing precision shrinks inherent arithmetic cost in both ASICs and FPGAs**
  - Instantiate **100x** more compute within the same fabric and thereby scale performance

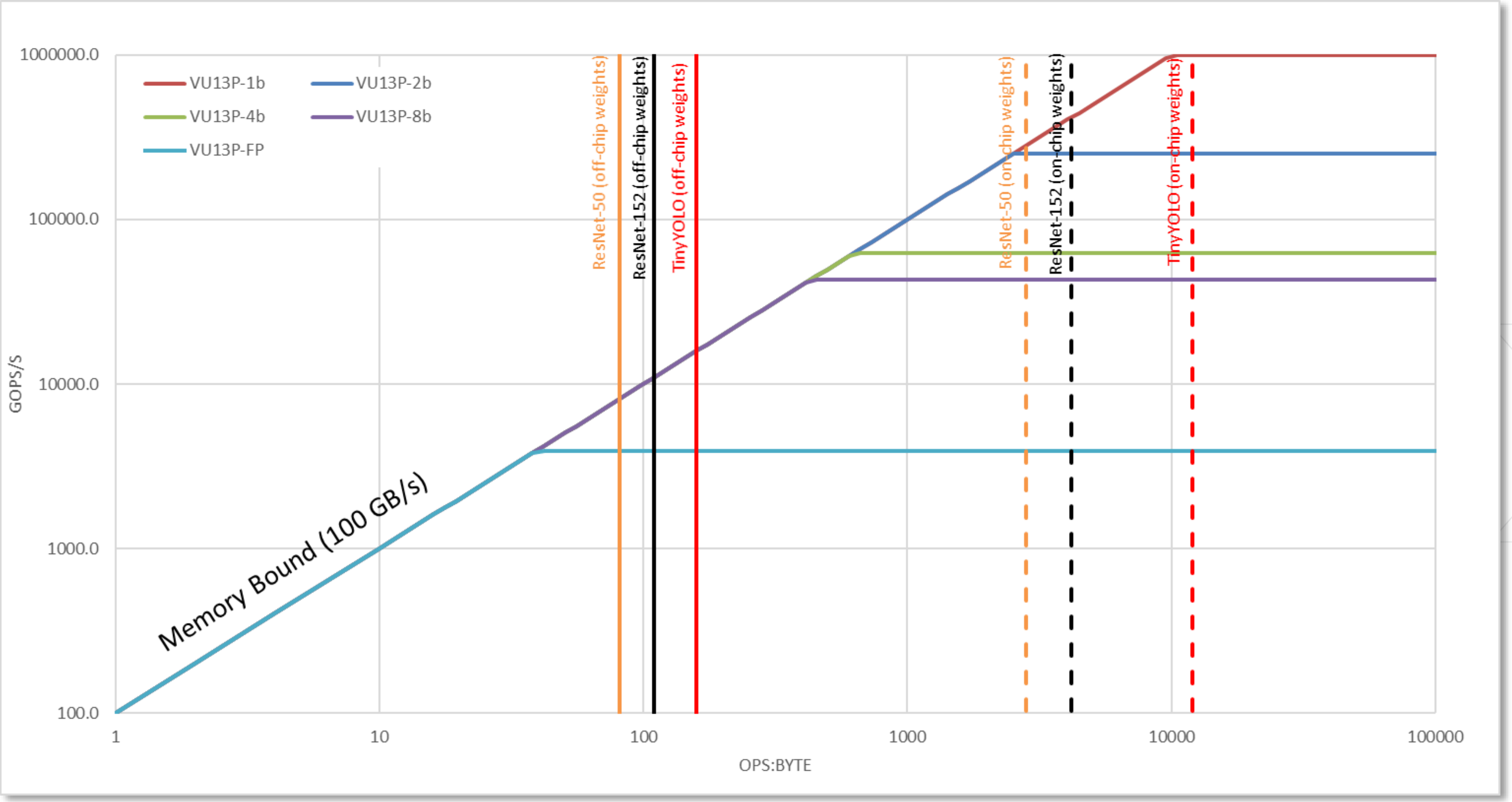


C= size of accumulator * size of weight * size of activation

(to appear in ACM TRETS SE on DL, FINN-R)

Reducing Precision provides Performance Scalability

Example: ResNet50, ResNet152 and TinyYolo



RP scales compute performance

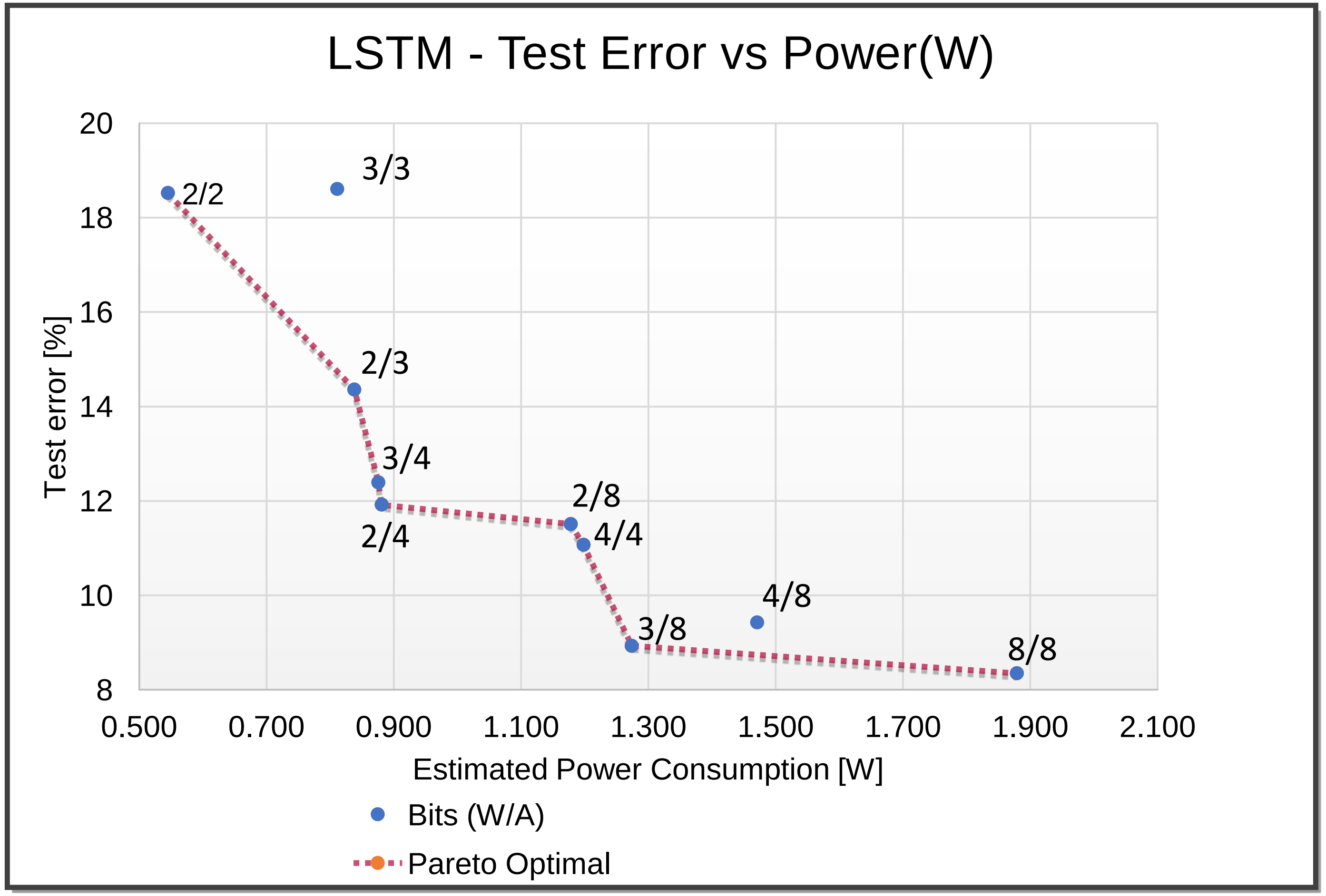
Theoretical Peak Performance for a VU13P with different Precision Operations
Assumptions: Application can fill device to 90% (fully parallelizable) 710MHz

RP reduces model size=> to stay on-chip

XILINX

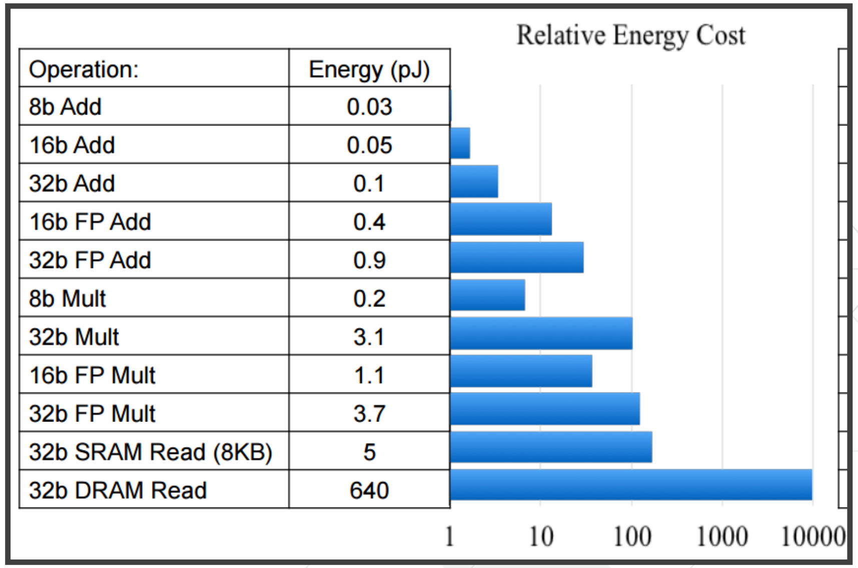
Reducing Precision Inherently Saves Power

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:



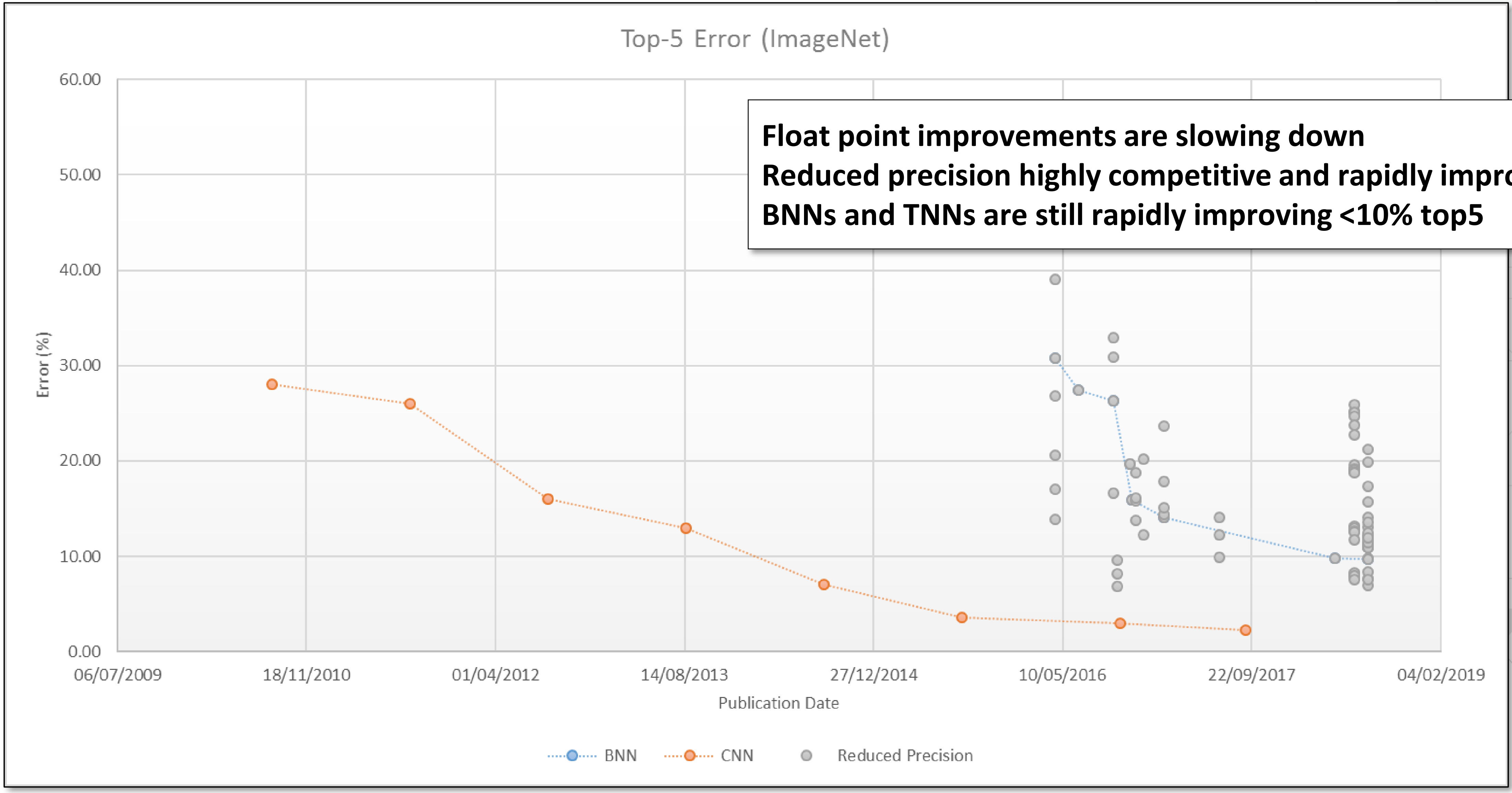
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Mult	0.2
32b Mult	3.1
16b FP Mult	1.1
32b FP Mult	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640

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

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."

What are the downsides of reduced precision?

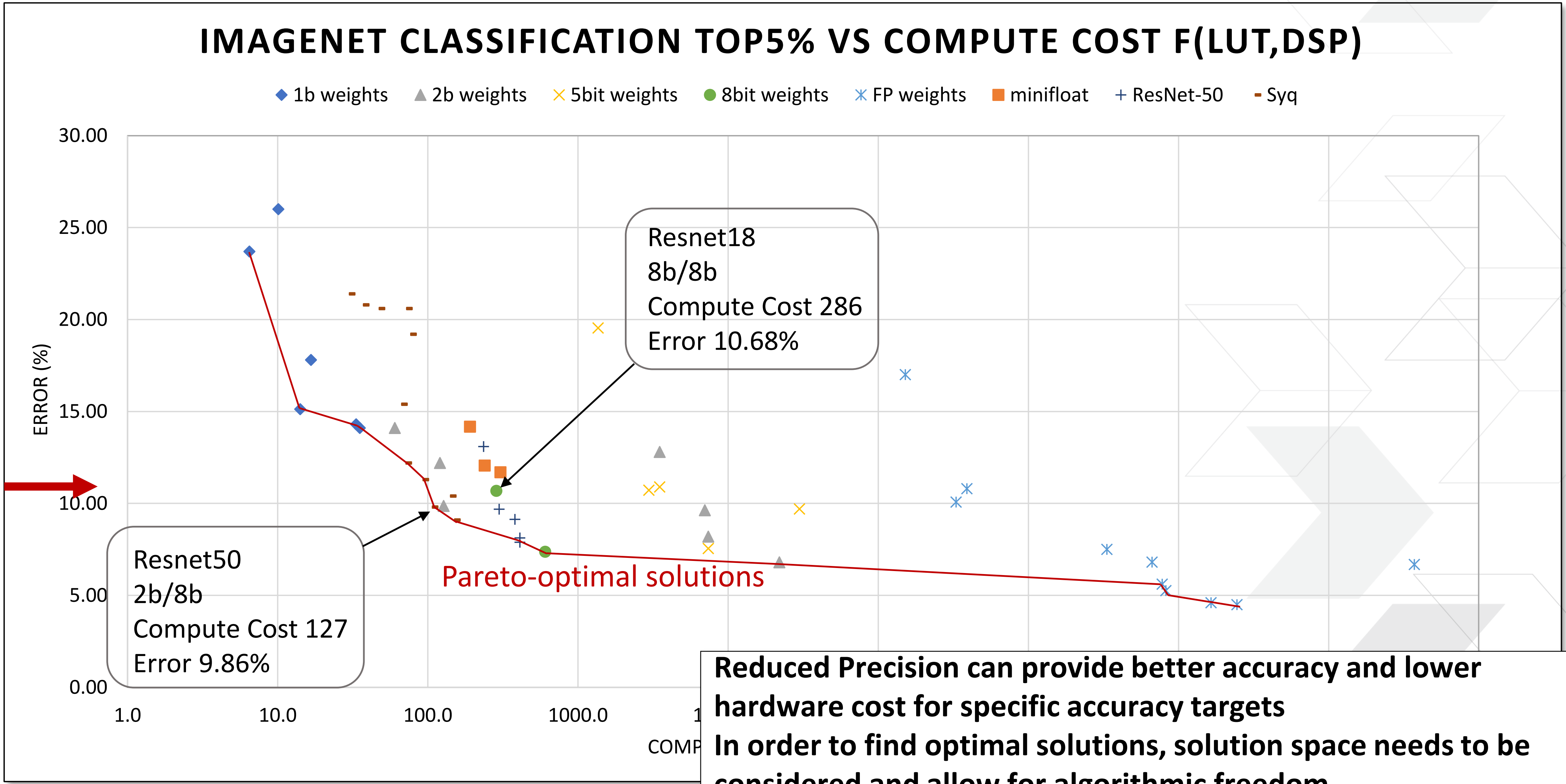
RPNNs: Closing the Accuracy Gap



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

Latest numbers: Dongqing Zhang∗, Jiaolong Yang∗, Dongqiangzi Ye∗, and Gang Hua "LQ-Nets: Learned Quantization for Highly Accurate and Compact Deep Neural Networks"

Design Space Trade-Offs



Reduced Precision can provide better accuracy and lower hardware cost for specific accuracy targets
In order to find optimal solutions, solution space needs to be considered and allow for algorithmic freedom

The Emerging Computational Landscape of Neural Networks

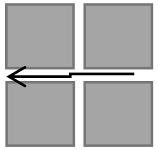
Exciting Times in Computer Architecture Research!

XILINX

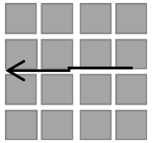
Spectrum of New Architectures for Deep Learning

DPU: Deep Learning Processing Unit

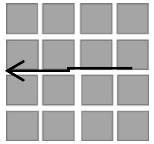
CPUs	GPUs	Soft DPUs (FPGA)	Hard DPUs (ASIC)	*In-Memory Compute*



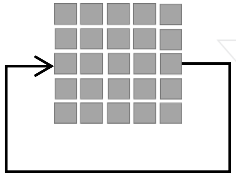
Intel
AMD
ARM



AMD
NVIDIA



DeePhi
Teradeep
XDNN



TPU, Cerebras, Graphcore, Groq, Nervana, Wave Computing, Eyeriss, Movidius, Kalray

Using non-volatile resistive memories or stacked DRAM*

ISAAC, Tetris, Neurcube

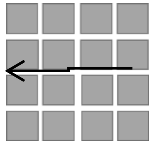
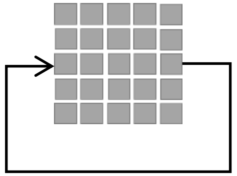
Vector-based SIMD processors
becoming increasingly customized for Deep Learning
(Tensor Cores, Reduced Precision,…)

**Shafiee, A., Nag, A., Muralimanohar, N., Balasubramonian, R., Strachan, J.P., Hu, M., Williams, R.S. and Srikumar, V., 2016. ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. ACM SIGARCH*

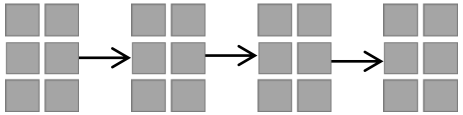
Chi, P., Li, S., Xu, C., Zhang, T., Zhao, J., Liu, Y., Wang, Y. and Xie, Y., 2016, June. Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory. In ACM SIGARCH

Chen, Y., Luo, T., Liu, S., Zhang, S., He, L., Wang, J., Li, L., Chen, T., Xu, Z., Sun, N. and Temam, O., 2014, December. Dadiannao: A machine-learning supercomputer. In Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (pp. 609-622). IEEE Computer Society.

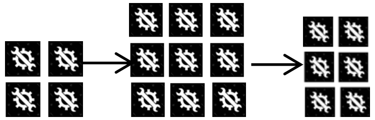
Architectural Choices – Macro-Architecture

	Soft DPUs (FPGA)	Hard DPUs (ASIC)
Matrix of PE	 DeePhi, Teradeep, XDNN	 **TPU**, Cerebras, Graphcore, Groq, Nervana, Wave Computing, Eyeriss, Movidius, Kalray

Customized macro-architecture (Synchronous Dataflow)



MSR Brainwave*

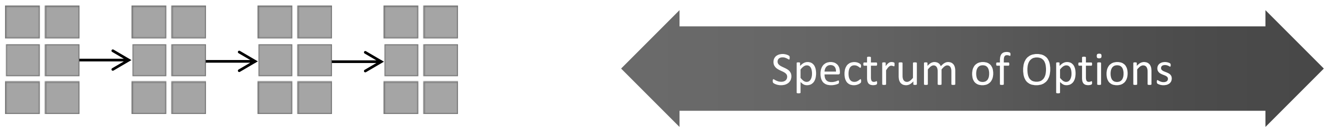


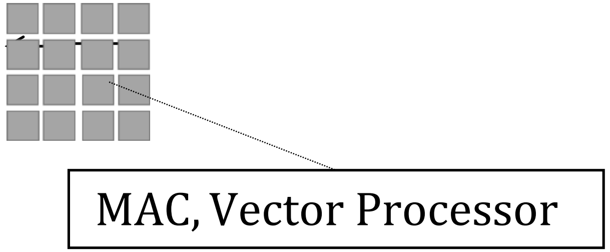
FINN**

Chung, E., Fowers, J., Ovtcharov, K., Papamichael, M., Caulfield, A., Massengill, T., Liu, M., Lo, D., Alkalay, S., Haselman, M. and Abeydeera, M.Serving DNNs in Real Time at Datacenter Scale with Project Brainwave. IEEE Micro, 38(2) https://www.microsoft.com/en-us/research/uploads/prod/2018/06/ISCA18-Brainwave-CameraReady.pdf

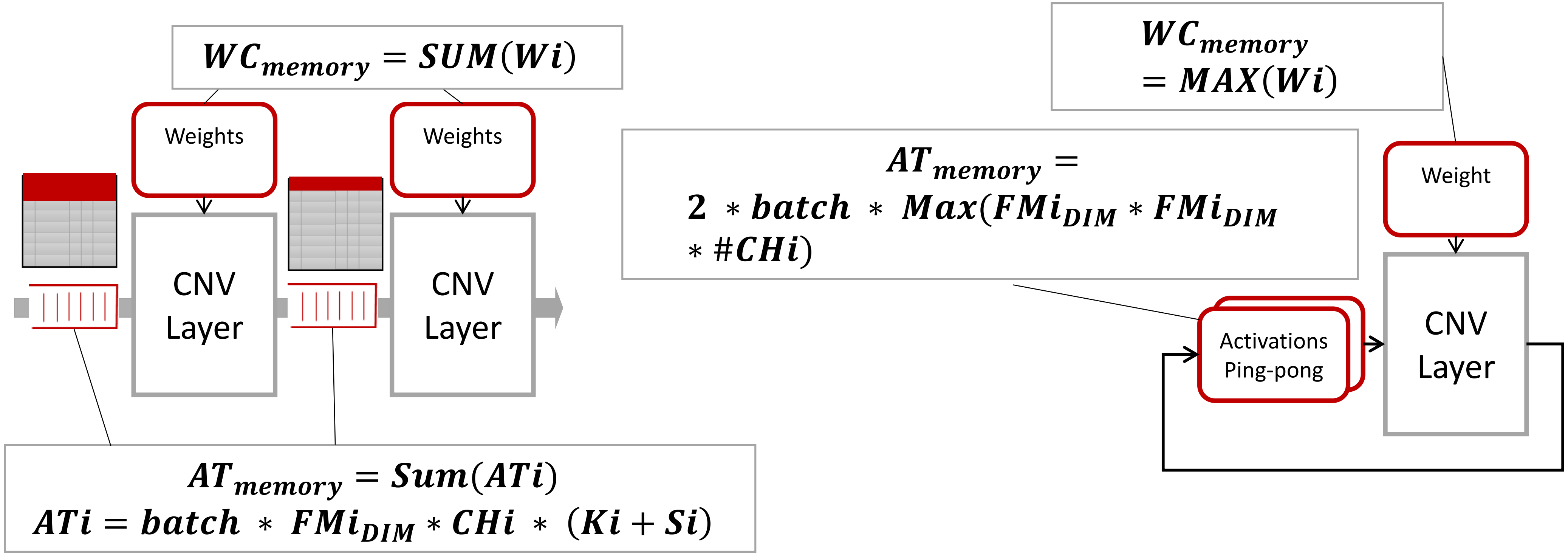
***Umuroglu, Yaman, Umuroglu, Y., Fraser, N.J., Gambardella, G., Blott, M., Leong, P., Jahre, M. and Vissers, K. "FINN: A framework for fast, scalable binarized neural network inference." ISFPGA'2017*

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



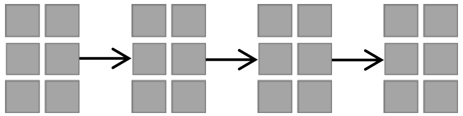


>> End points are pure layer-by-layer compute and feed-forward dataflow architecture

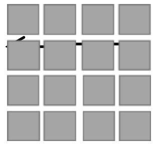


Lin, X., Yin, S., Tu, F., Liu, L., Li, X. and Wei, S. LCP: a layer clusters paralleling mapping method for accelerating inception and residual networks on FPGA. DAC'2016
Alwani, M., Chen, H., Ferdman, M. and Milder, P. Fused-layer CNN accelerators. MICRO 2016.

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



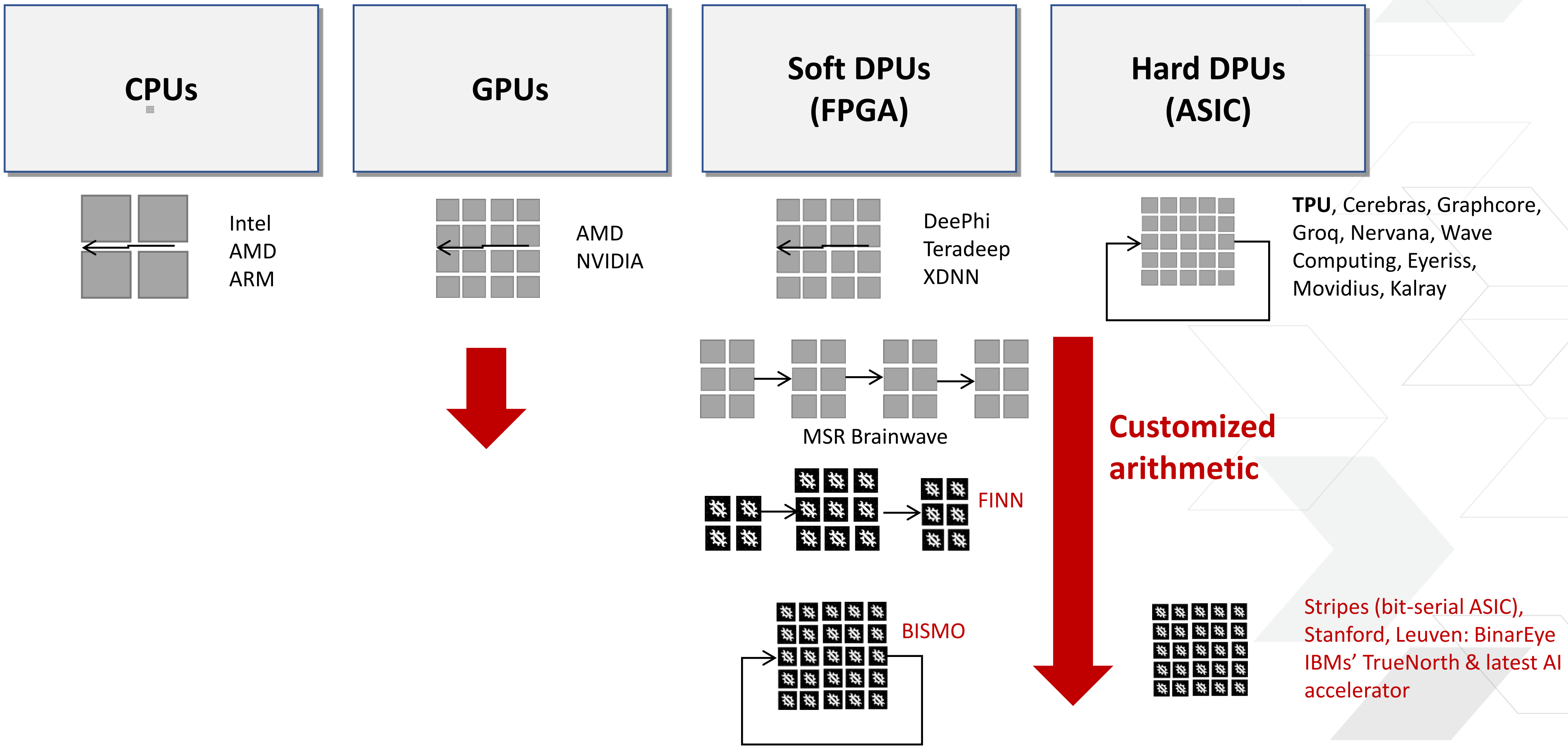
Degree of parallelization across layers



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

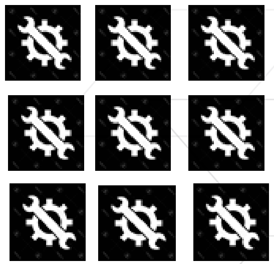
- Requires less on-chip weight memory, but more activation buffers
- Efficiency of memory for weights and activations depends on how well balanced the topology is
- Flexible hardware, which can scale to arbitrary large networks
- Compute efficiency is a scheduling problem
  => generating sophisticated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

Architectural Choices – Micro-Architecture



Judd, P., Albericio, J., Hetherington, T., Aamodt, T.M. and Moshovos, A., 2016, October. Stripes: Bit-serial deep neural network computing. MICRO'2016
Moons, B., Bankman, D., Yang, L., Murmann, B. and Verhelst, M. BinarEye: An always-on energy-accuracy-scalable binary CNN processor with all memory on chip in 28nm CMOS, ICC'2018
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

Micro-Architecture:
Customized Arithmetic for Specific Numerical Representations



- **Customizing arithmetic compute allows to maximize performance at minimal accuracy loss**
  - Flexpoint, Microsoft Floating Point formats, Binary & Ternary, Bfloat16

- **Which do we focus on?**

- **What's more, non-uniform arithmetic can yield more efficient hardware implementations for a fixed accuracy***
  - Run-time programmable precision: Bit-Serial

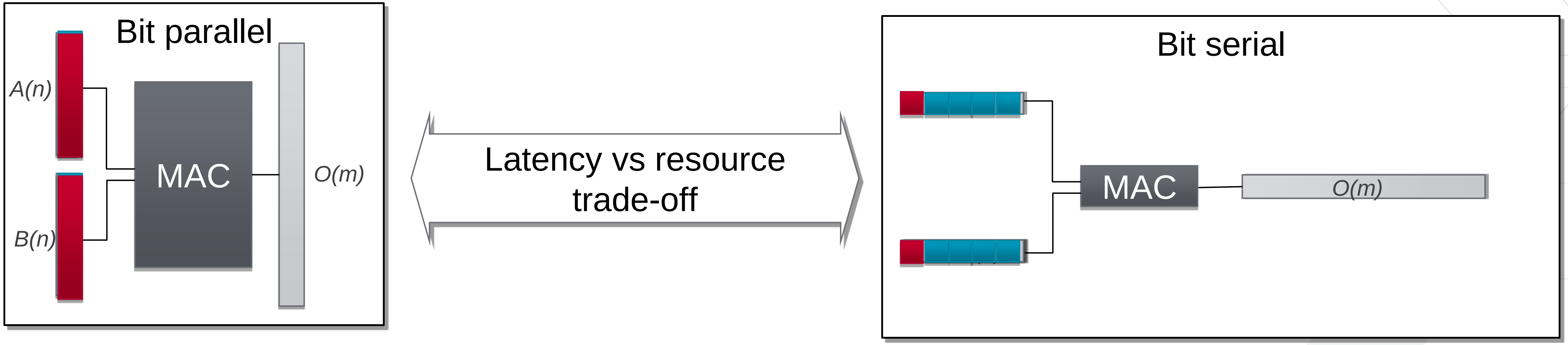
	DEC	INC	CONCAVE	CONVEX
Top-1 [%]	53.79	50.35	54.45	54.33
Top-5 [%]	77.59	74.89	76.43	78.20

Table 2. Accuracy comparison of our approach under different styles of layer-wise quantization.

**Eunhyeok Park, Junwhan Ahn, and Sungjoo Yoo, "Weighted-Entropy-based Quantization for Deep Neural Networks" CVPR'2017]*

Micro-Architecture:
Bit-Parallel vs Bit-Serial

> **Bit-serial can provide run-time programmable precision with a fixed architecture**
>> ASIC* or FPGA** overlay



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

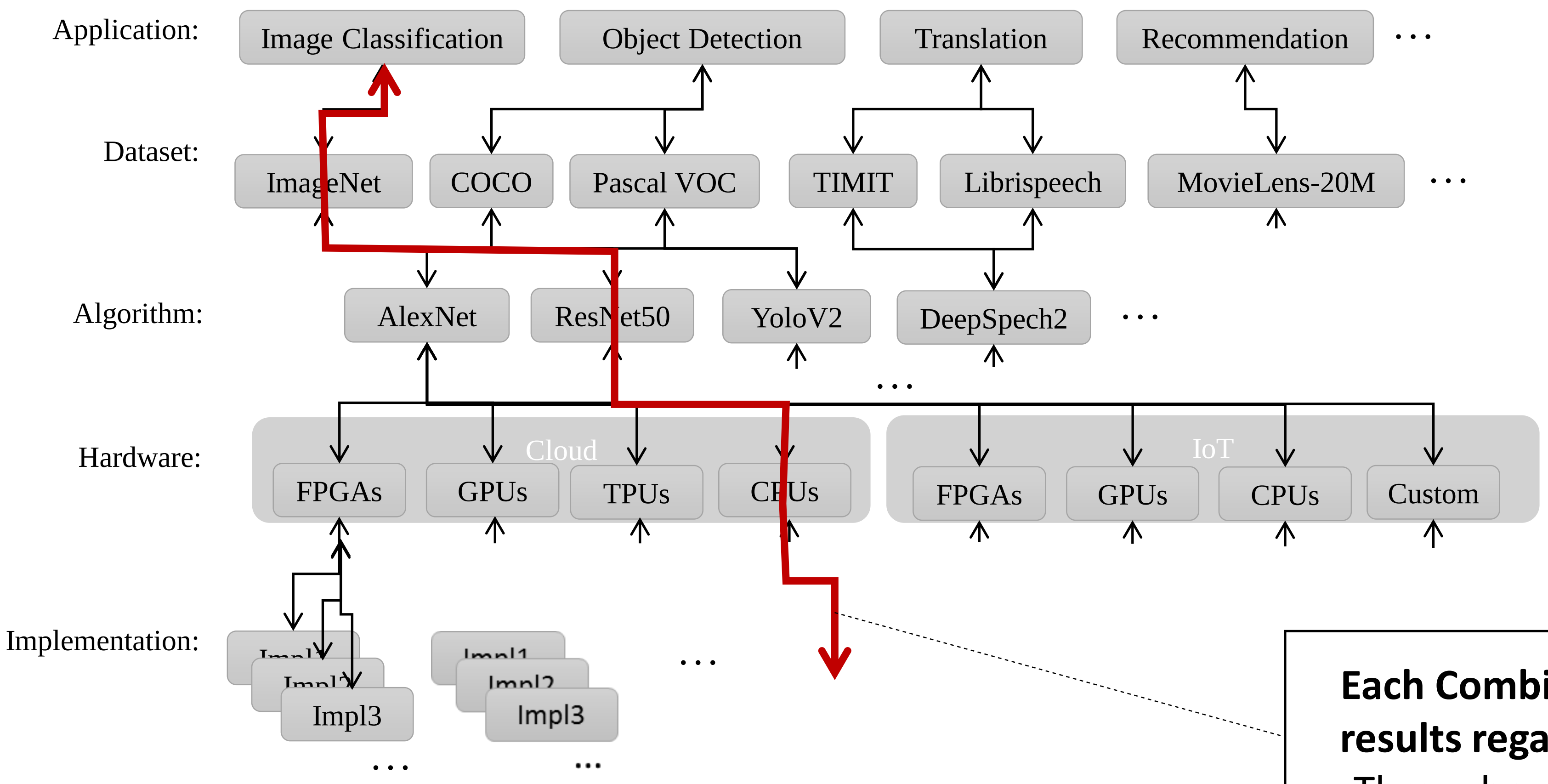
**Judd, P., Albericio, J., Hetherington, T., Aamodt, T.M. and Moshovos, A., 2016, October. Stripes: Bit-serial deep neural network computing. MICRO'2016*
***Umuroglu, Rasnayake, Sjalander"BISMO: A Scalable Bit-Serial Matrix Multiplication Overlay for Reconfigurable Computing." FPL'2018*
https://arxiv.org/pdf/1806.08862.pdf

Summary

Summary

- 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,…

THANK YOU!

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