Communications and Computer Engineering II

FPGA Application

Hiroki Nakahara
Tokyo Institute of Technology

Outline

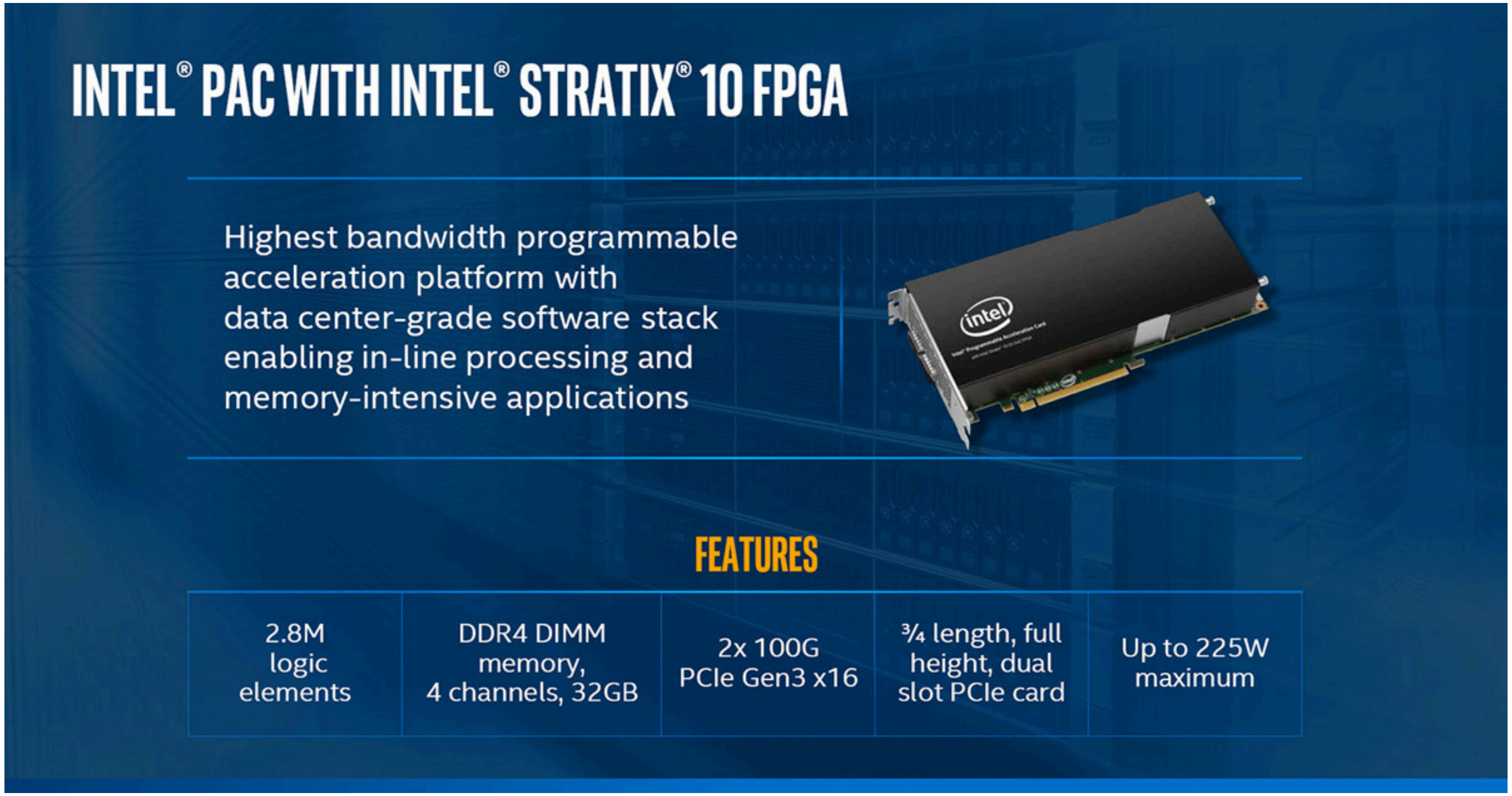
- Trends
- Killer Applications
- AI (Deep-Learning) Accelerator
  - Trends
  - Optimization Techniques
- Summary

1. Trends

Intel Acquisition of Altera

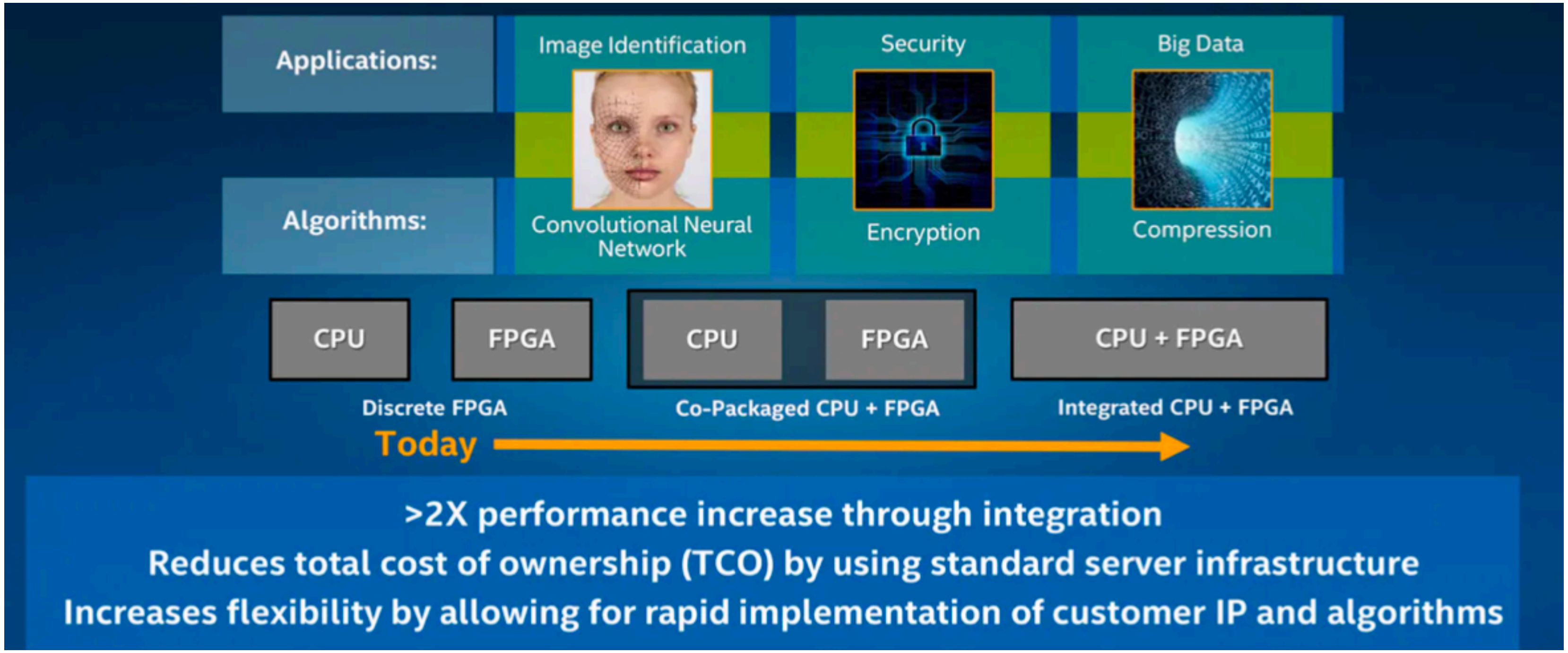
now part of Intel

- CPU market reaches to the end of growing?
- FPGA ”potential” for non-Neumann model
- Stratix 10 series (toward data center)



Data Center FPGA Acceleration

- Up to 1/3 of cloud service provider nodes to use FPGAs by 2020
- AI (Neural network), security, big-data



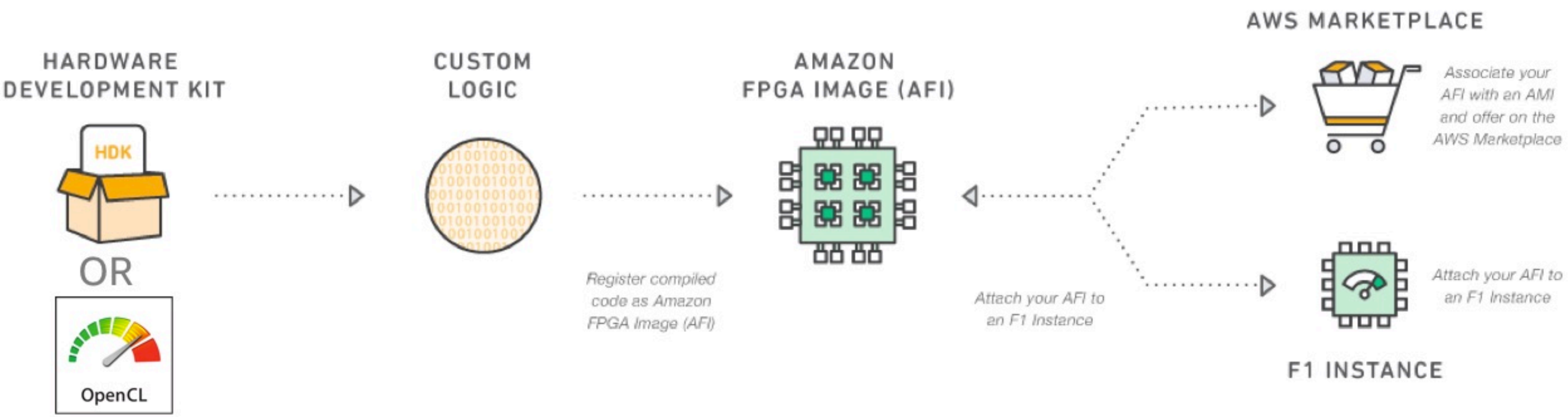
Requirements for AI Computing

Cloud	Embedded
Many classes (1000s)	Few classes (<10)
Large workloads	Frame rates (15-30 FPS)
High efficiency (Performance/W)	Low cost & low power (1W-5W)
Server form factor	Custom form factor

J. Freeman (Intel), “FPGA Acceleration in the era of high level design”, HEART2017

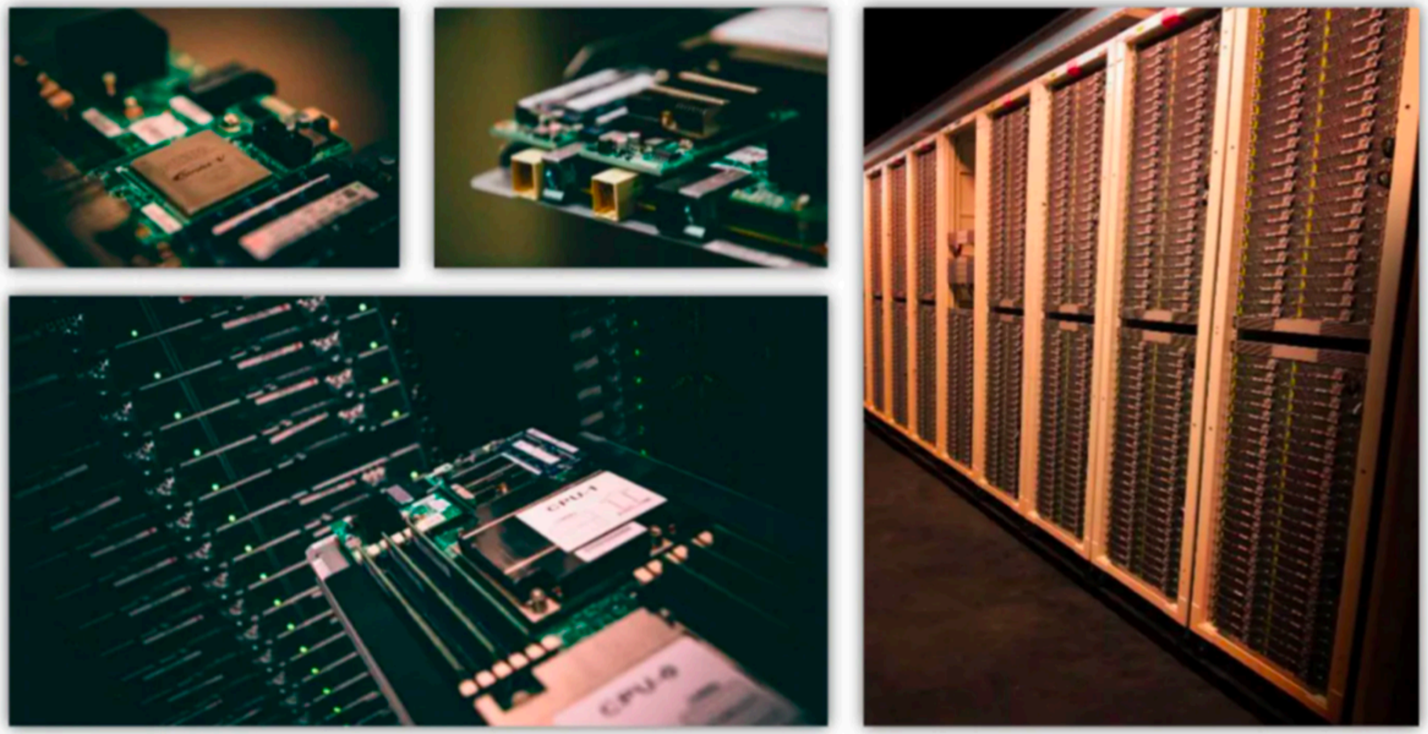
AWS supports FPGA Instance

- As an EC2 Instances
  - Xilinx FPGA
- OpenCL-based programming
  - SDAccel 2019.1



Microsoft Datacenter Server

- Catapult project
  - Bing and Azure deployed new multi-FPGA
- Arria10 FPGAs on Azure cloud system

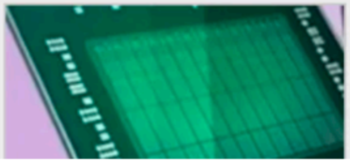


https://www.microsoft.com/en-us/research/project/project-catapult/

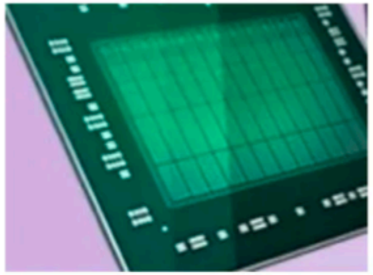
IBM put big data FPGA design in Cloud

RECOMMENDED ARTICLES

ARM TechCon: FPGA module has 16Gbit/s I/O



IBM's cloud service will host the Xilinx SDAccel development environment which will allow developers to describe their algorithms in OpenCL, C, and C++ and then compile directly to Xilinx FPGA-based acceleration boards.



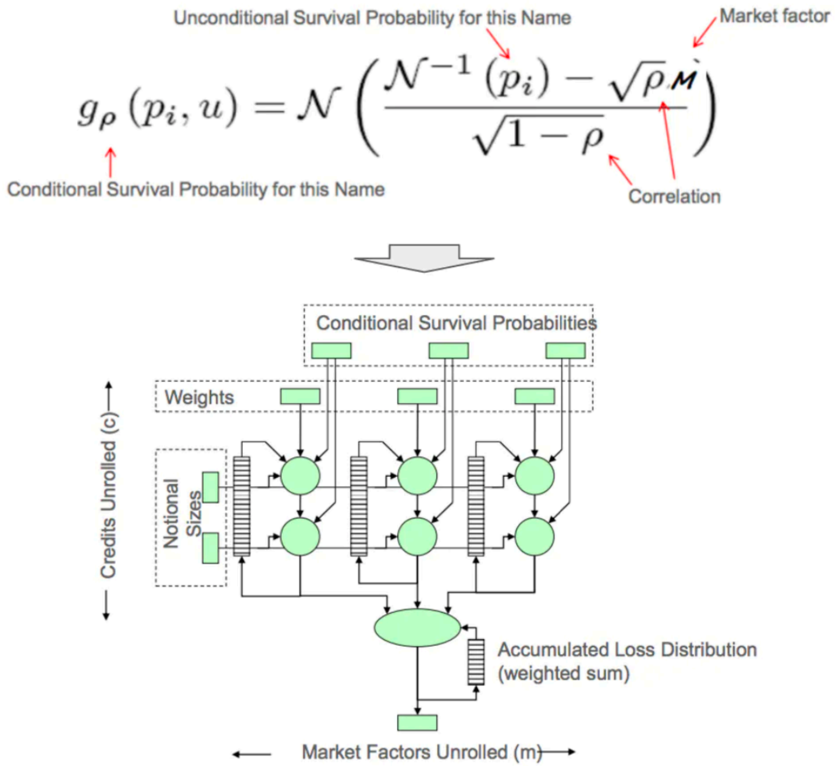
This is an open access cloud service, called SuperVessel, which can be used by application developers, system designers, and academic researchers to create, test and pilot their FPGA designs for big data analytic processors and even data gathering IoT node devices.

http://www.electronicsweekly.com/news/xilinx-and-ibm-put-big-data-fpga-design-in-the-cloud-2016-04/

2. Killer Applications

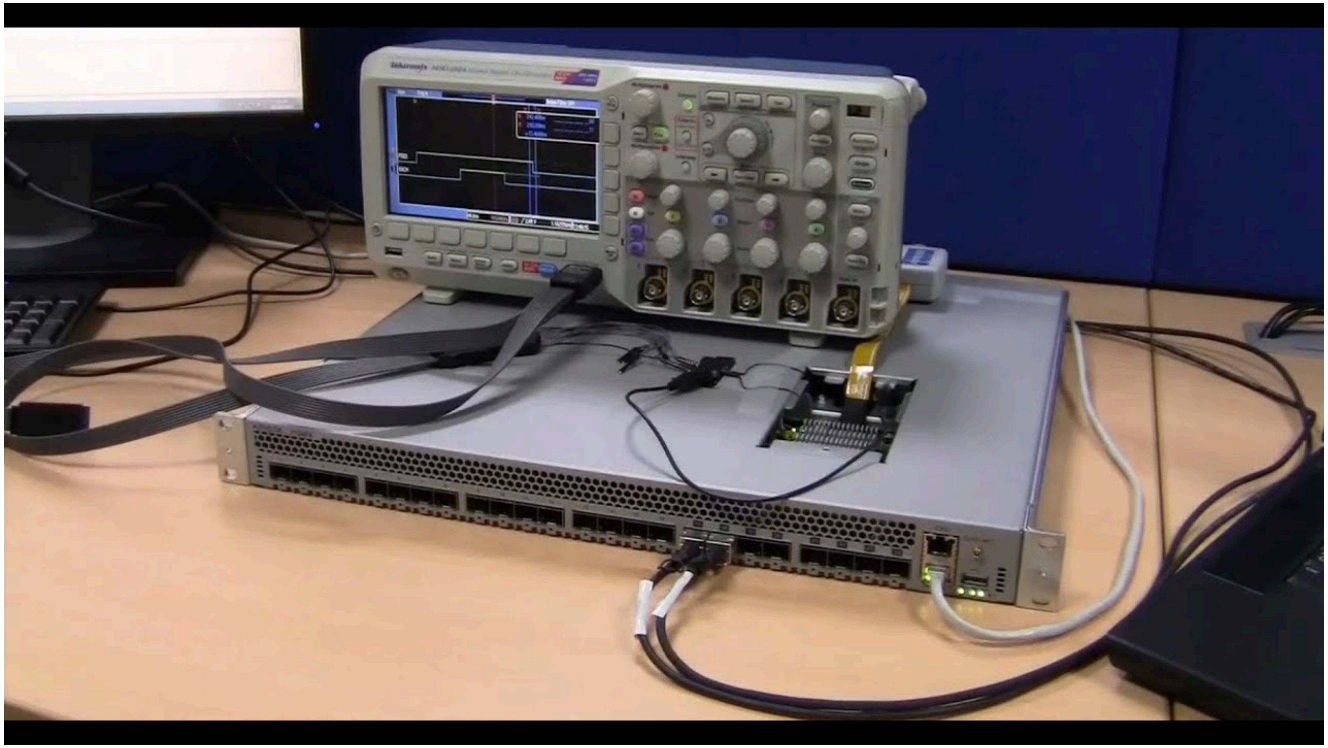
JP Morgan

- FPGA implementation of derivative risk analysis
- Reduced company-wide risk analysis from 8H to 4min



High Frequency Trading (HFT)

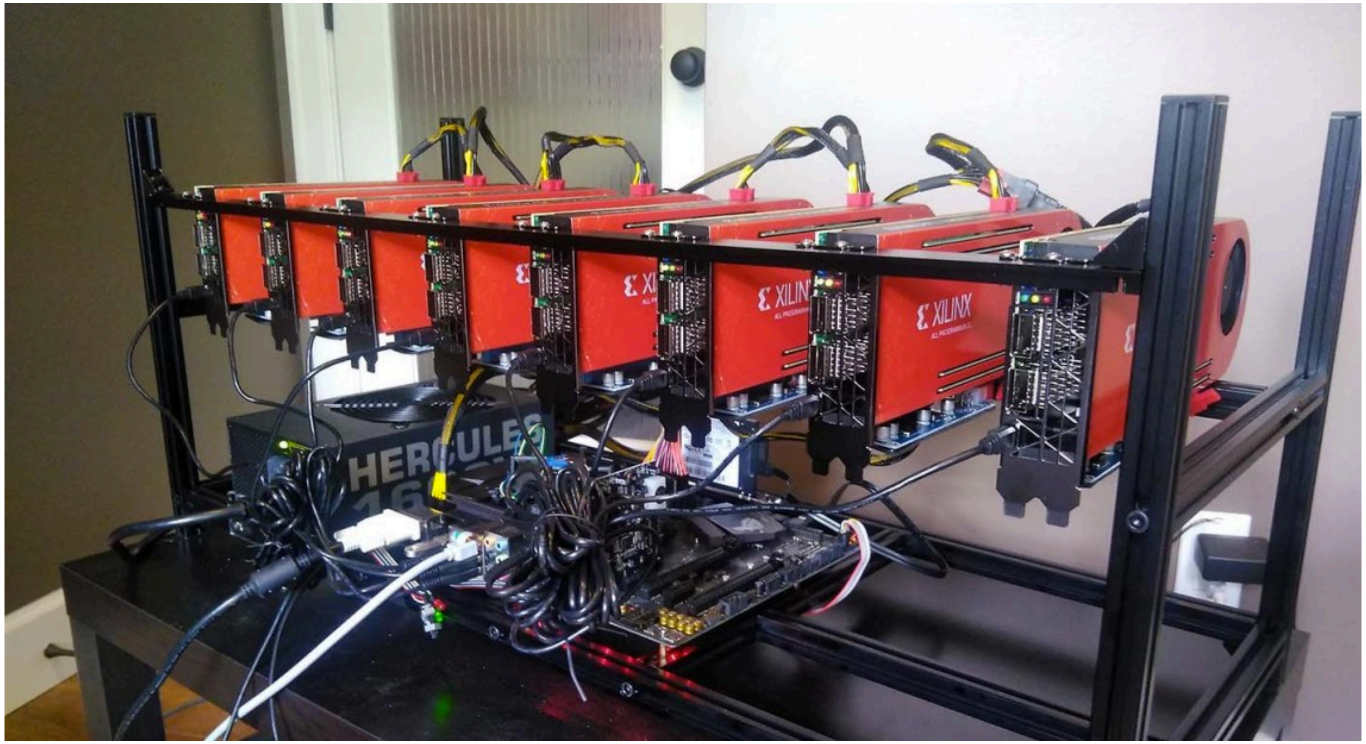
- Buy and sell in microseconds
  - Not in time for software
- Send trading packets while receiving stock price packets



https://www.youtube.com/watch?v=uDy_8Q0GdTk

Bitcoin Mining

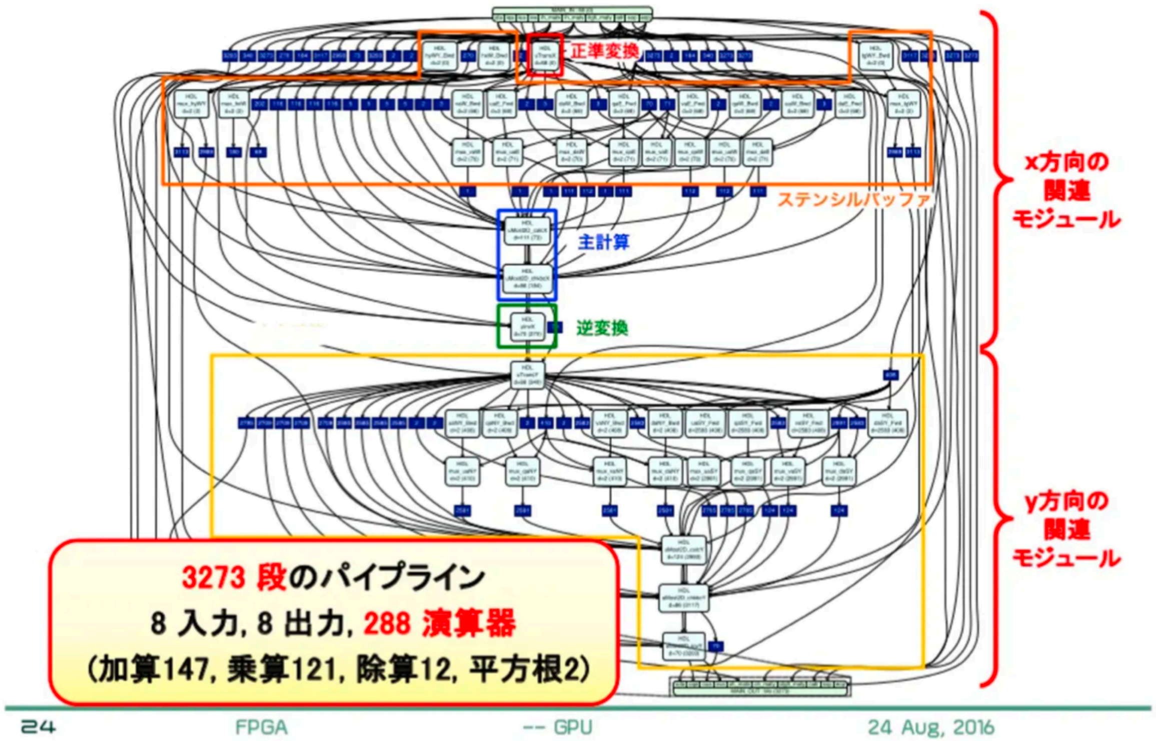
- Brute force hash value
- Flexible response to specification changes



https://en.bitcoin.it/wiki/Open_Source_FPGA_Bitcoin_Miner

Tsunami Simulator

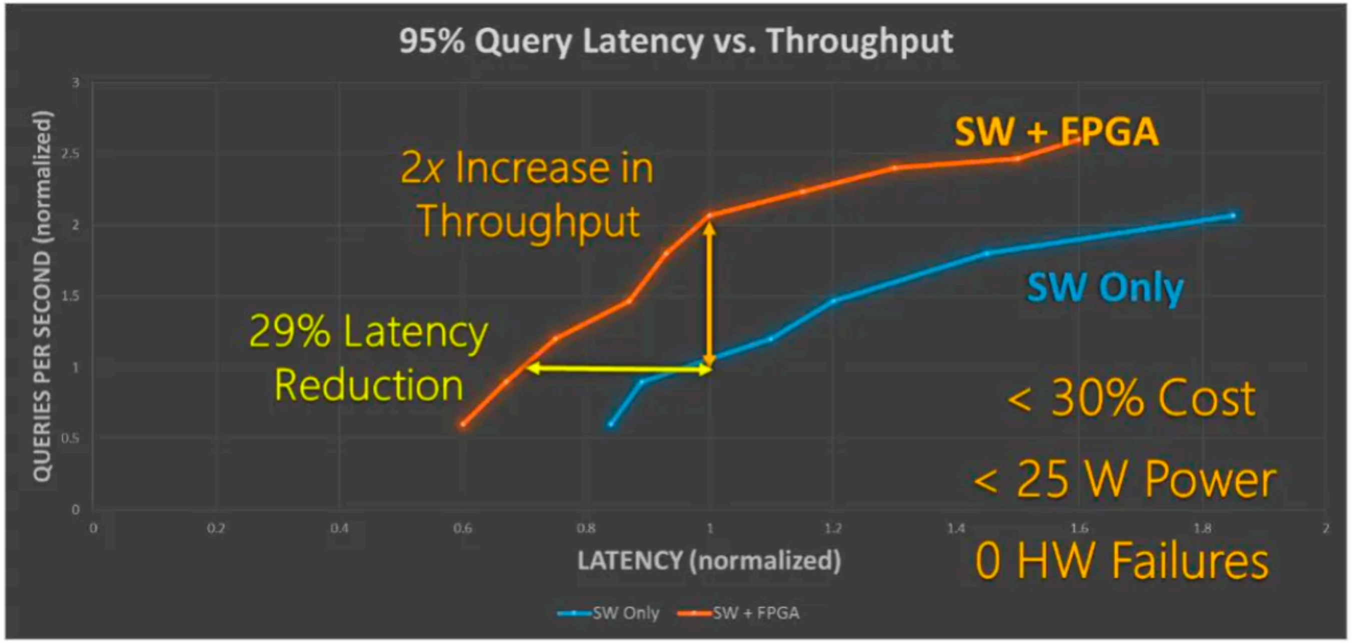
- Tsunami prediction by grid method
- Outperforms the GPU with a 3000-stage pipeline



Source: K. Sano (RIKEN) et al.

Bing Search by Microsoft

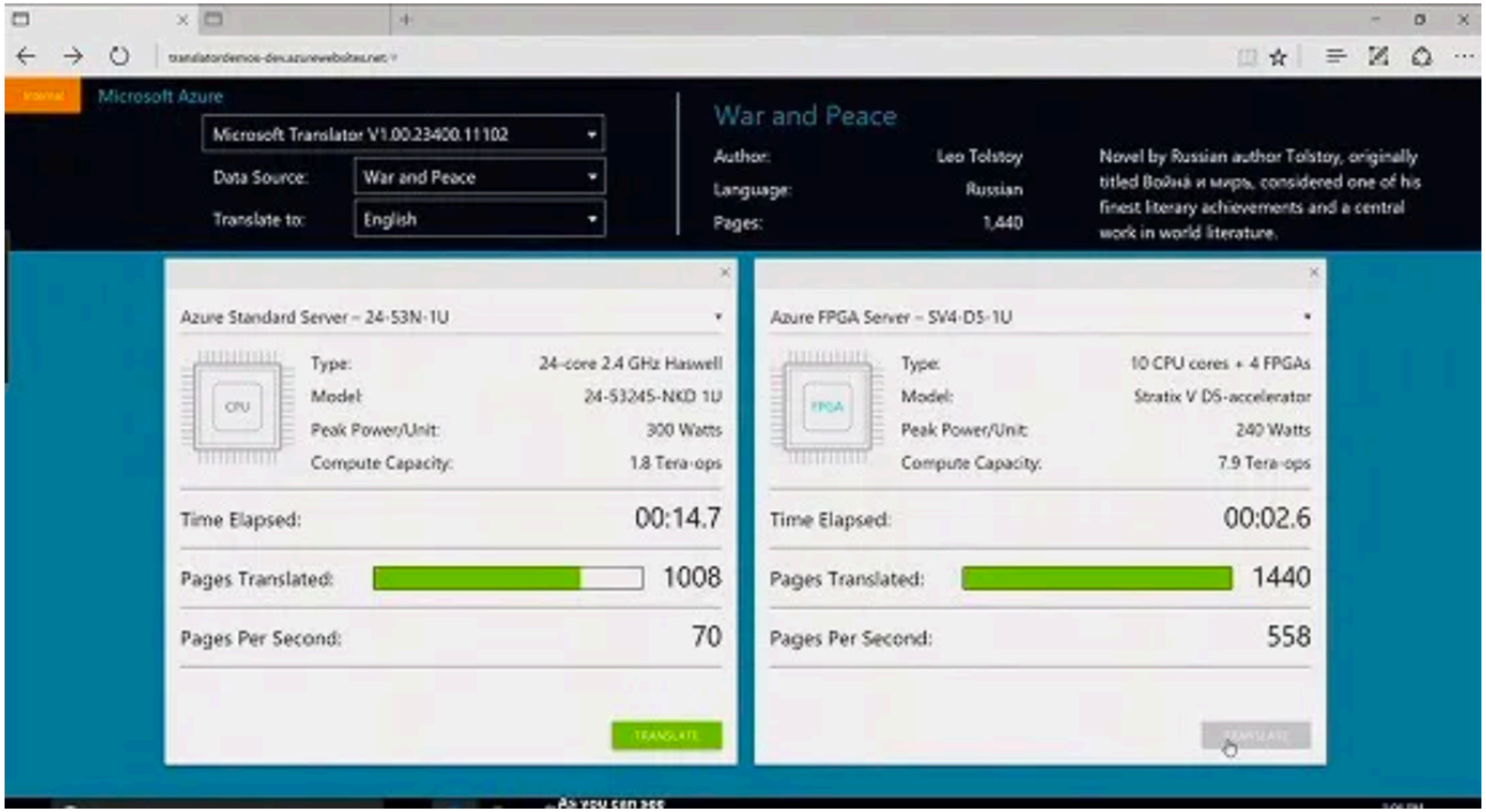
- Feature extraction and neural network inference
- 2x increase in Throughput



https://www.microsoft.com/en-us/research/publication/a-reconfigurable-fabric-for-accelerating-large-scale-datacenter-services/?from=http%3A%2F%2Fresearch.microsoft.com%2Fpubs%2F212001%2Fcatapult_isca_2014.pdf

Azure Translation Service

- CPU: 14 seconds, FPGA: 2.6 secs

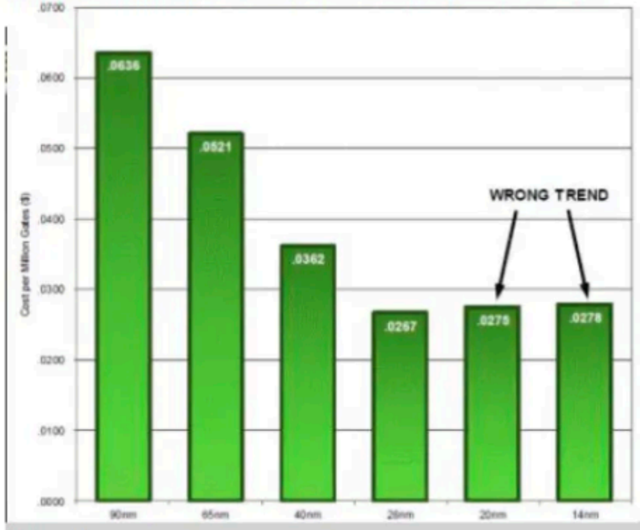


Source: Microsoft Ignite, CEO keynote (26/Sep./2016)

Why?

- Microsoft thinks that the Moore's low reaches to the end
- Hardware specialization
  - Economics will increasingly drive silicon ecosystem
  - Number of leading-edge fab vendors shrinking
  - Cost of performance growth will increase
  - Hardware specialization will be critical

Chart 8: IBS Calculation of Cost per Transistor by Node



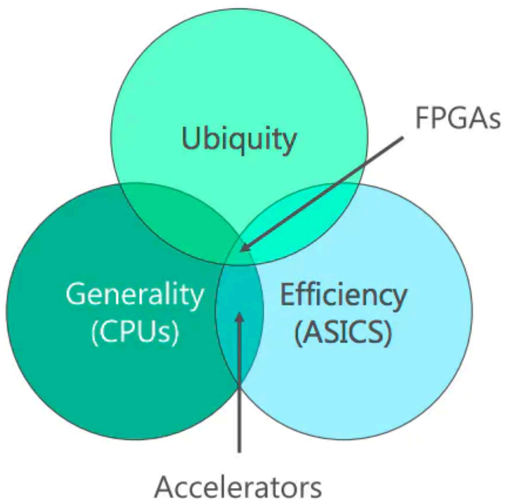
Source: IBS.
http://embedded.com/discussion/other/4238315/Feature-dimension-reduction-slowdown

What's next of CPUs?

- ASIC
  - Mass production costs tens of millions to hundreds of millions of yen, development period is months to years
  - Best performance and power
- GPU
  - Very good at performance a large amount of floating-point arithmetic and SIMD arithmetic throughput
  - Software engineers can develop relatively easily with CUDA and OpenCL
  - Flexible circuit design like ASIC and FPGA is not possible, it is not good at application specified
- FPGA
  - The upper limit of the clock is about several hundred MHz, and the circuit scale that can be assembled is much smaller than that of ASIC and GPU
  - Development is not as easy as GPU
  - Circuit configuration can be freely rewritten according to the application, so, some applications can get a great effect
  - Compared to ASIC, the development period is short and it is strong against application specification changes

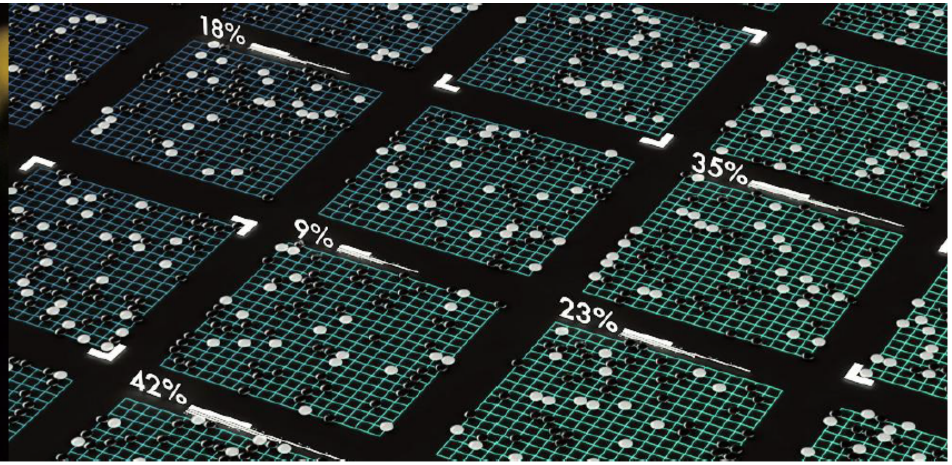
Microsoft Strategy

- With ASIC, development costs and time are large
  - Development and operation in units of 5 years
  - Prediction (additional functions and load) after 5 years is impossible
- There are 200 other cloud services besides Bing
- FPGA that can update circuit design every day
  - Flexibility to adapt to various application requirements and changes
  - High efficiency of dedicated hardware
    - Not as good as ASIC

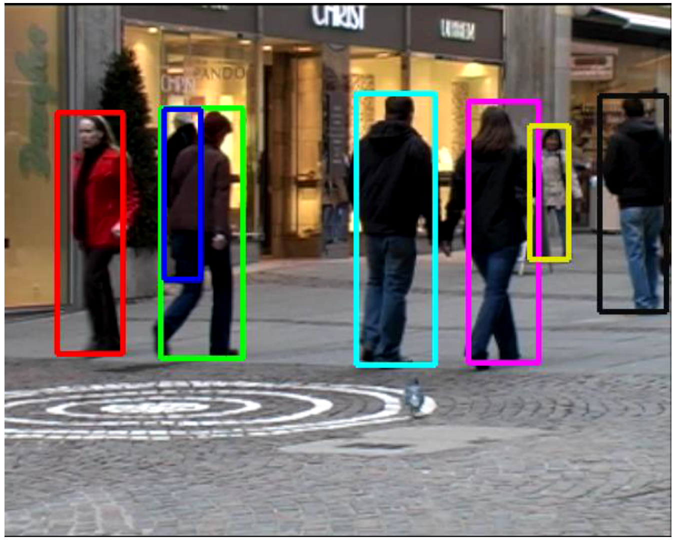


3. AI (Deep-Learning) Accelerator

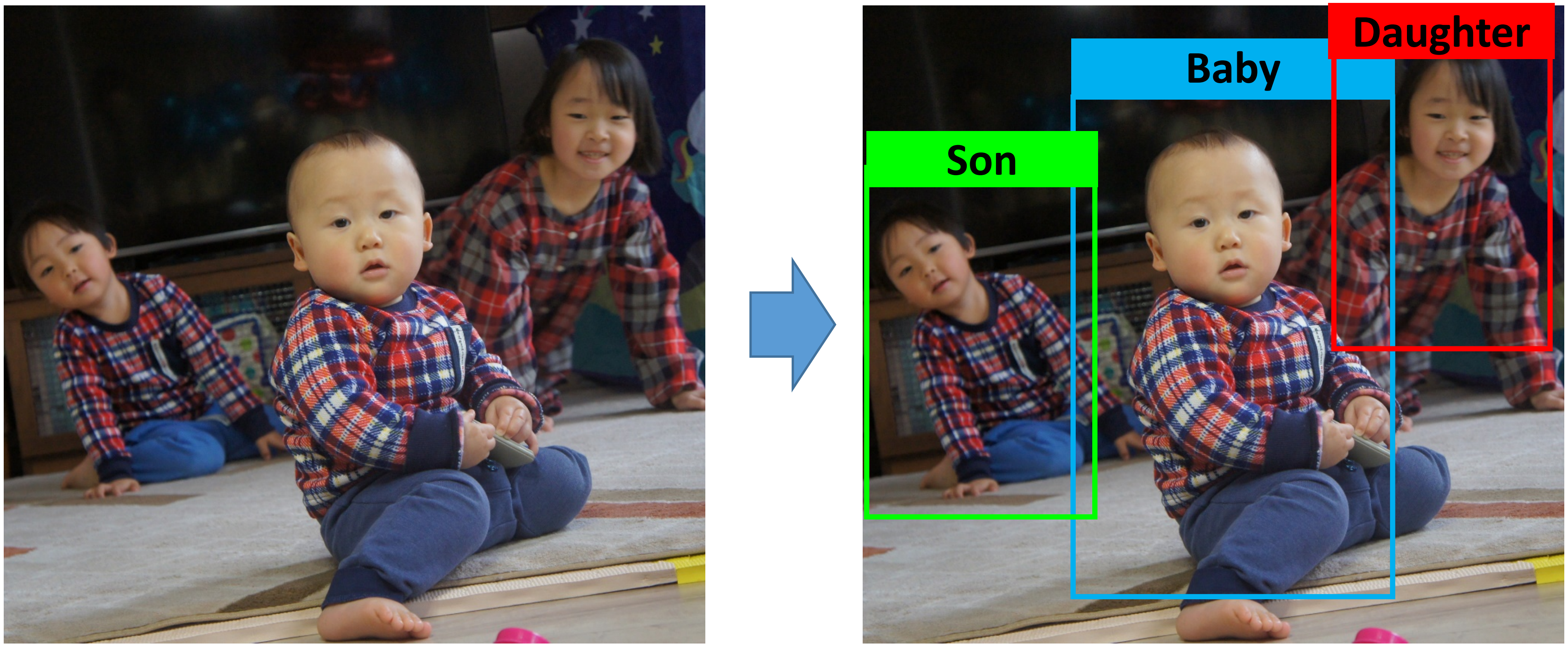
Artificial Intelligence is everywhere



Deep-Learning for Embedded Vision System

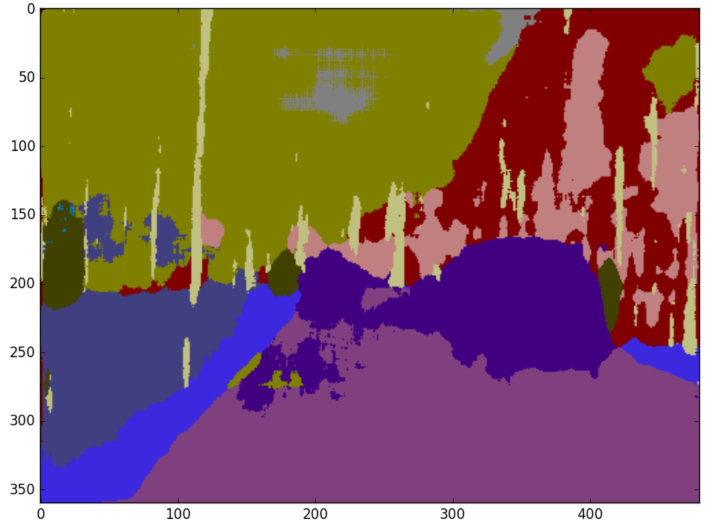


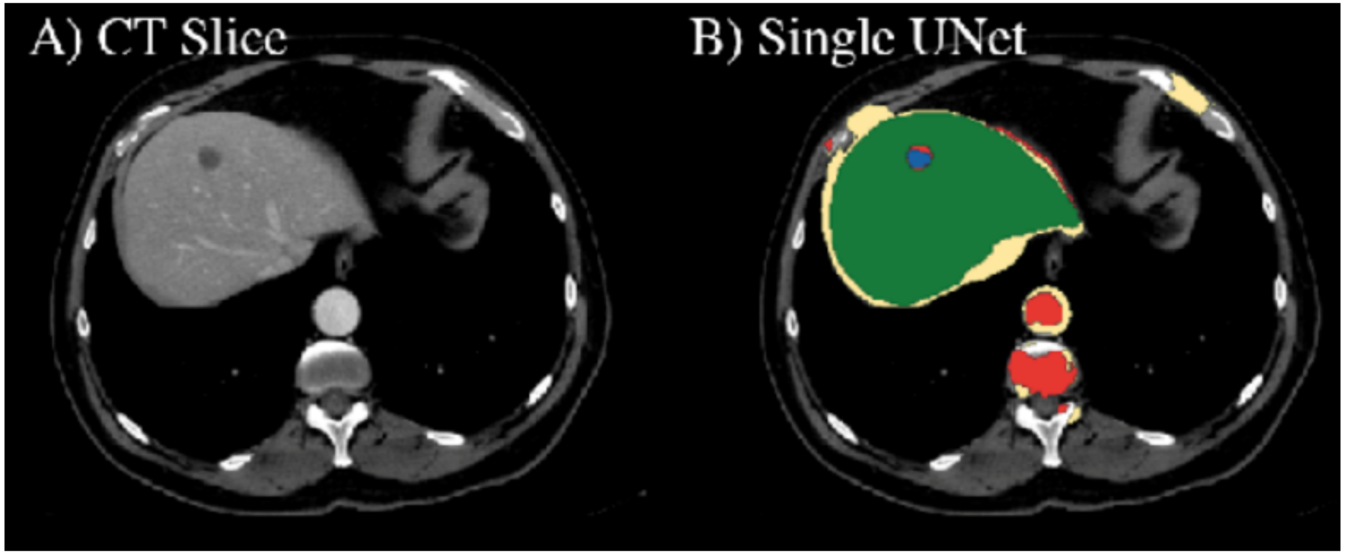
Object Detection



J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv, 2018

Semantic Segmentation



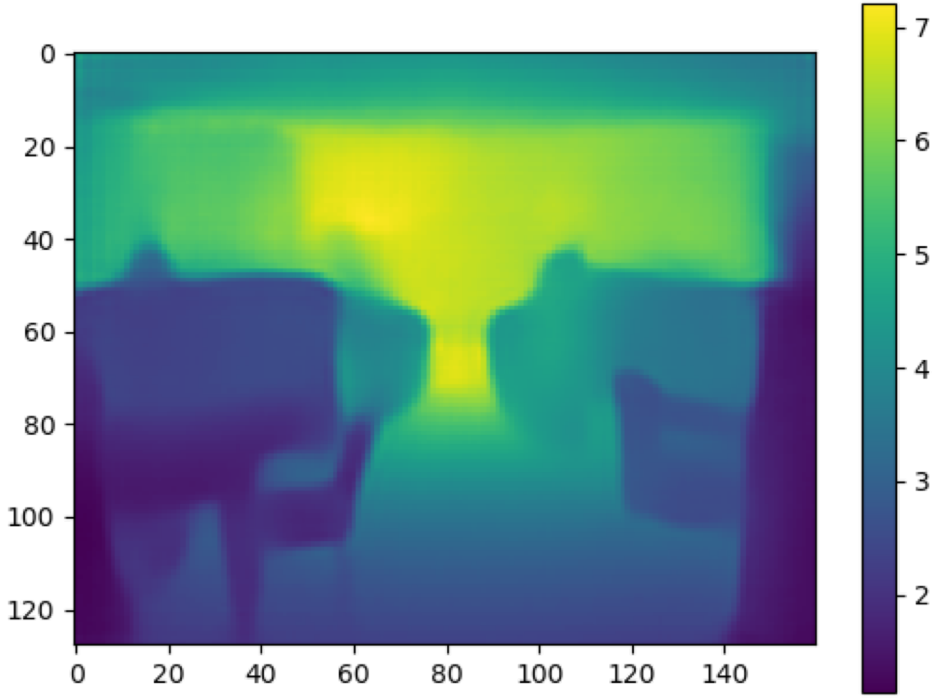


E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.39, No.4, 2017, pp. 640 - 651.

Pose Estimation

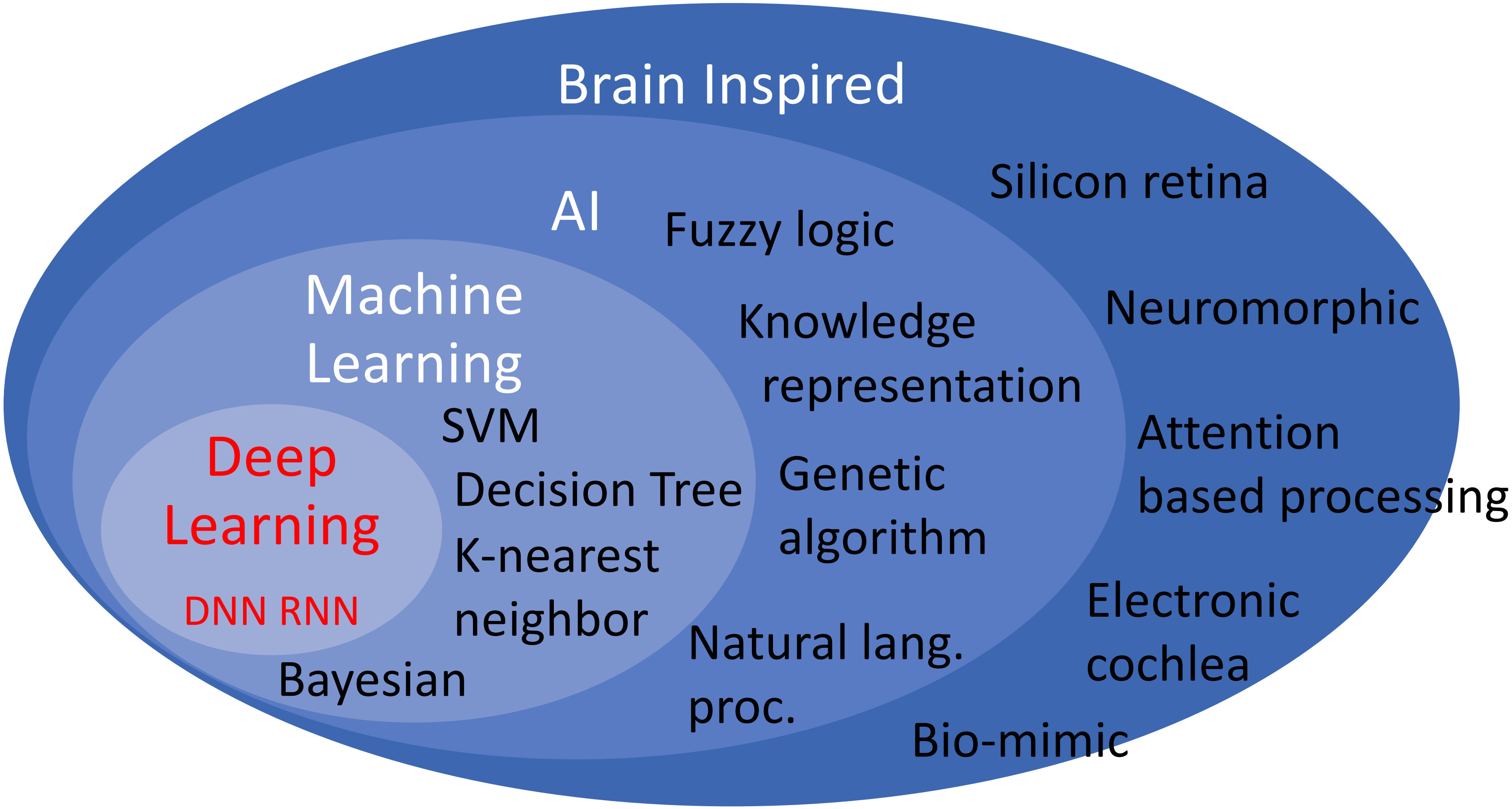
Z. Cao, T. Simon, S.-E. Wei and Y. Sheikh, " Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," CVPR, 2017.

Depth Estimation



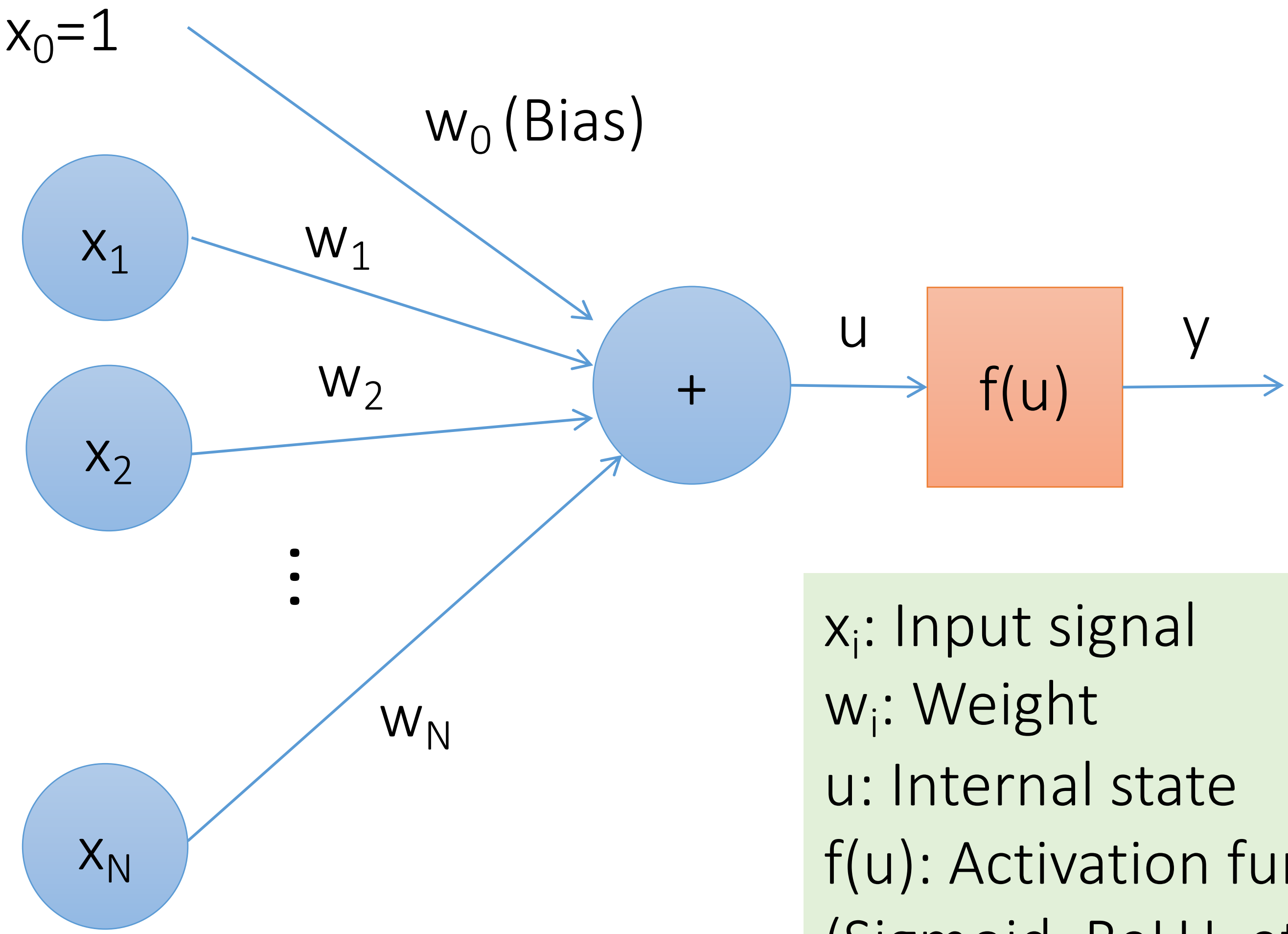
D. Eigen, C. Puhrsch and R. Fergus, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," arXiv:1406.2283 , 2014.

Intelligence and Deep Learning



J. Park, "Deep Neural Network SoC: Bringing deep learning to mobile devices," Deep Neural Network SoC Workshop, 2016.

Artificial Neuron (AN)

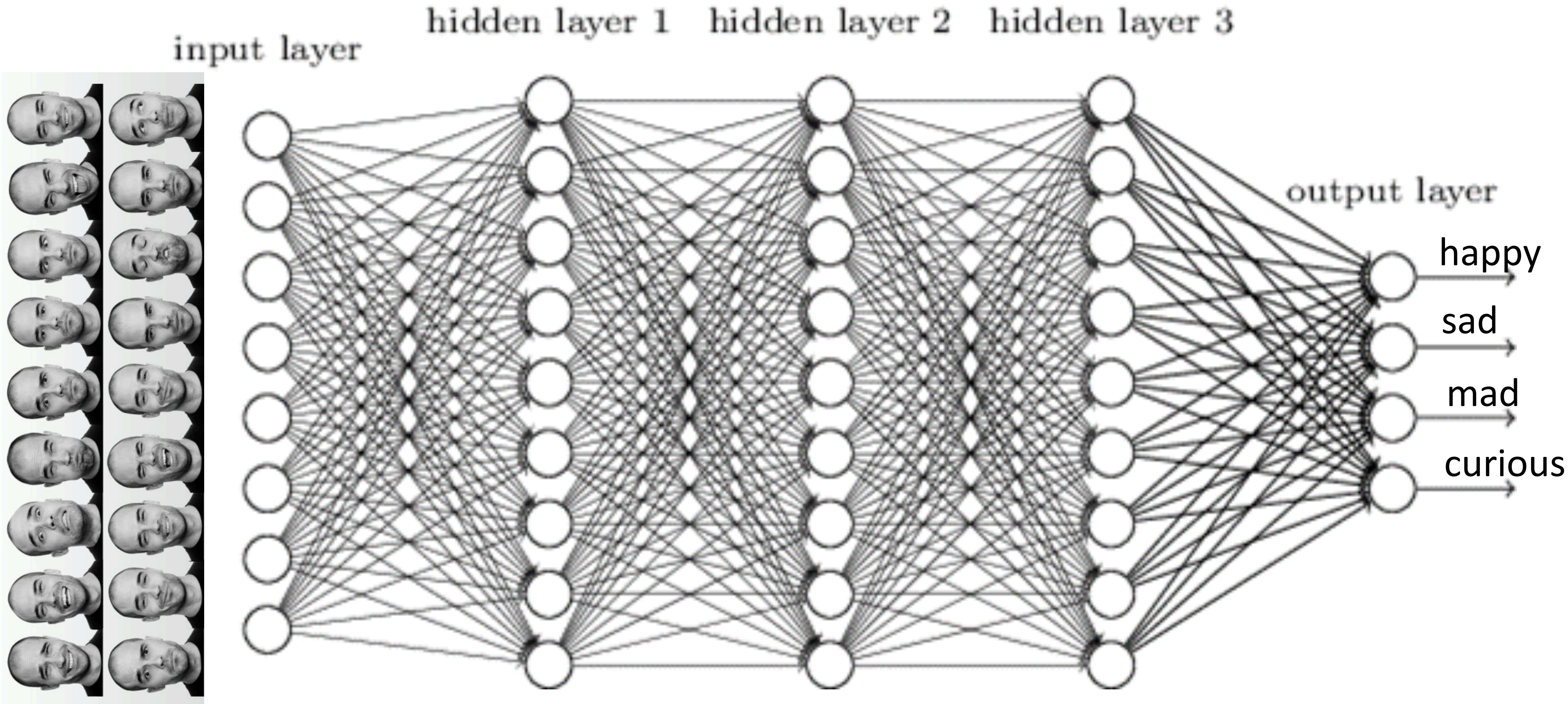


$$y = f(u)$$

$$u = \sum_{i=0}^{N} w_i x_i$$

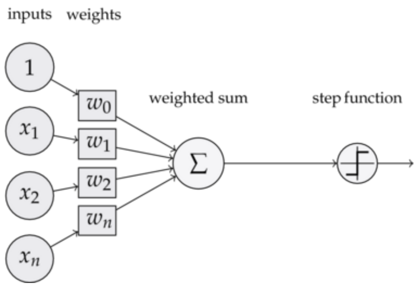
x_i: Input signal
w_i: Weight
u: Internal state
f(u): Activation function
(Sigmoid, ReLU, etc.)
y: Output signal

Deep Neural Network (DNN)



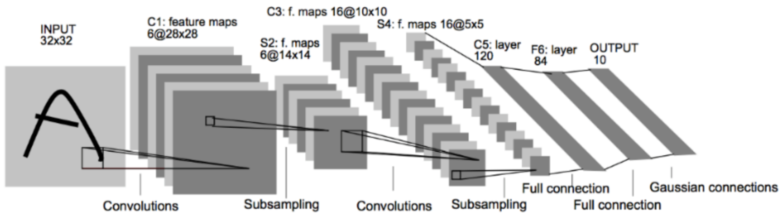
出典: imotionsglobal.com

Brief History: DNNs



Perceptron

Back-
propagation



Convolutional
Neural Network

Google
Brain Project

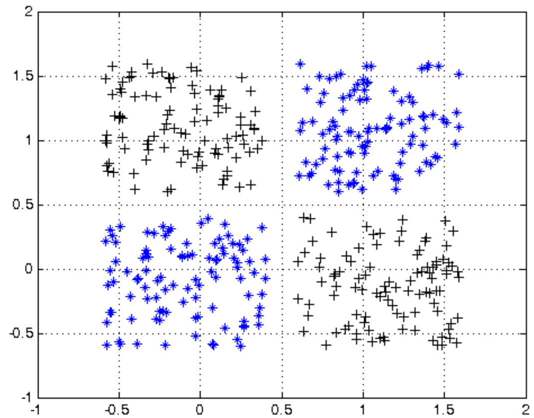
58 — 69 — 74 — 95 — 98 — 06 — 12

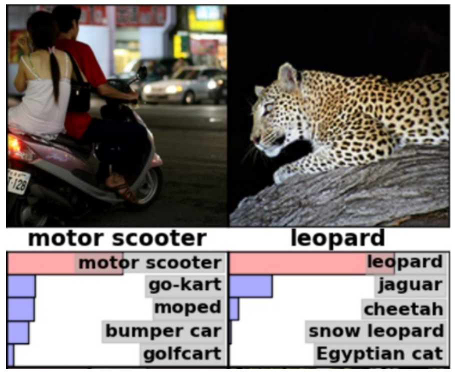
XOR Problem

SVM

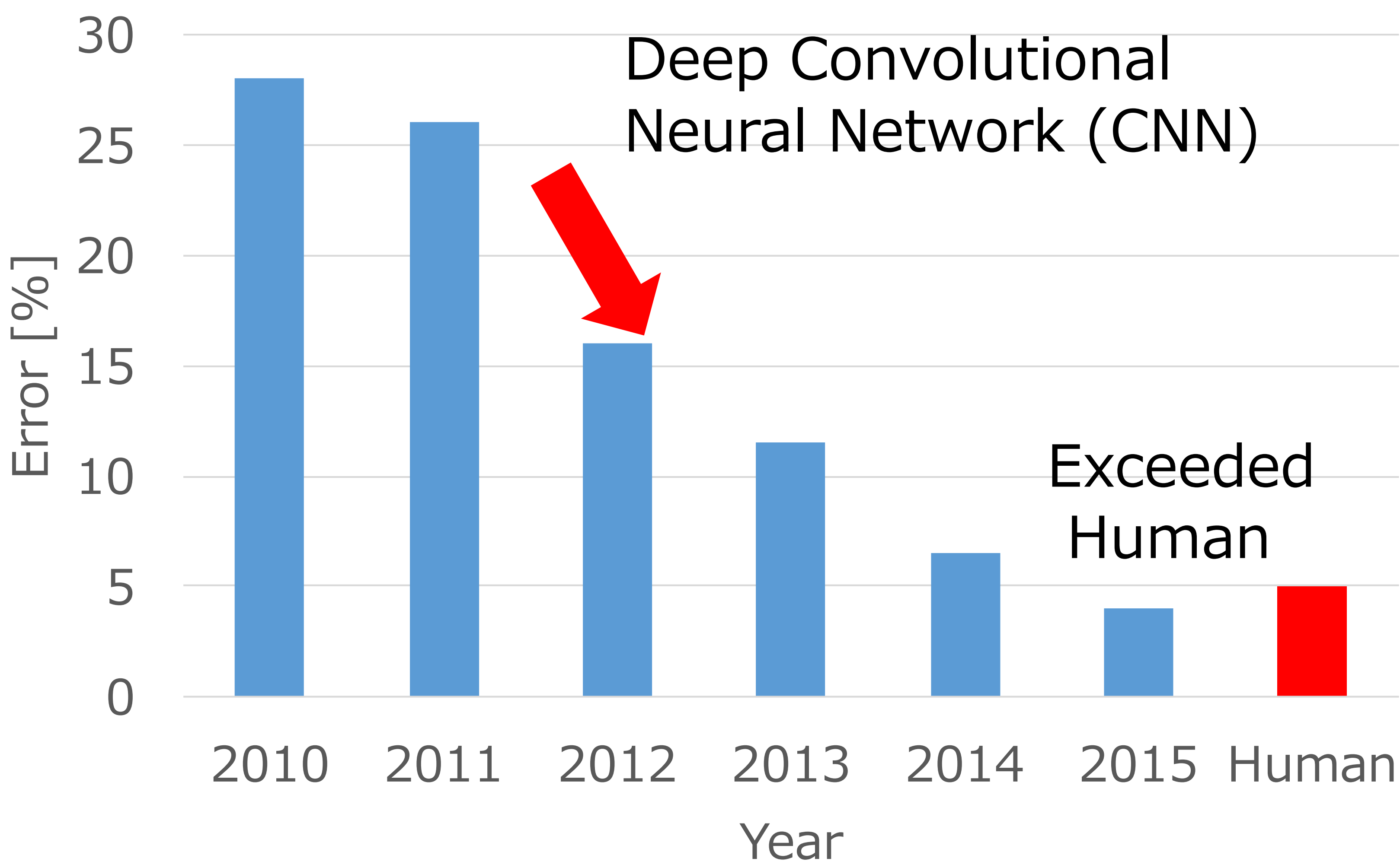
Restricted
Boltzmann
Machine

AlexNet
Wins





Accuracy of a DNN



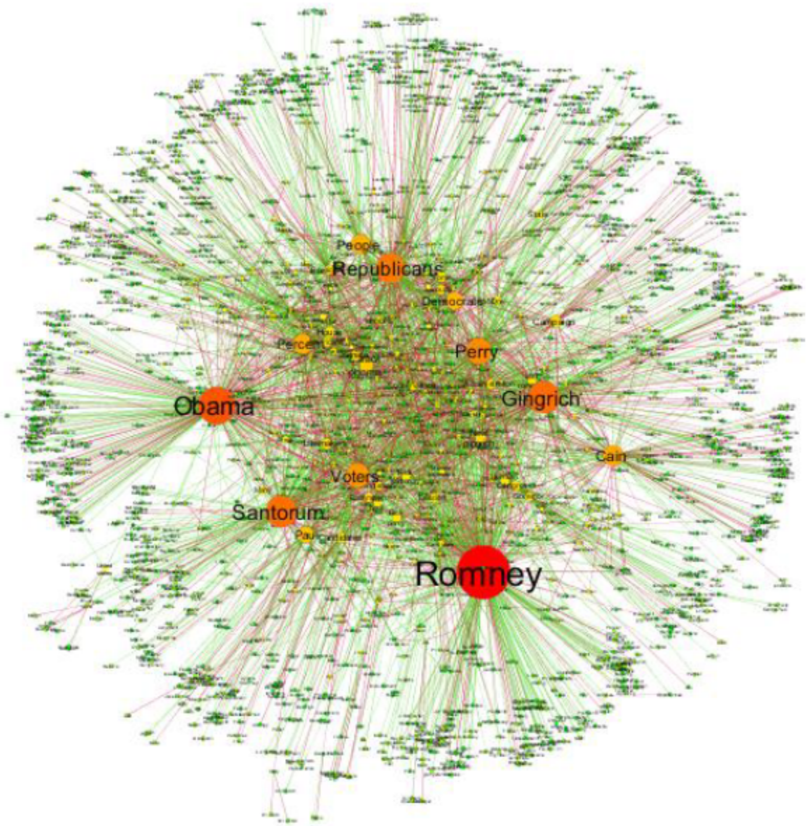
O. Russakovsky et al. “ImageNet Top 5 Classification Error (%),” IJCV 2015.

Technological singularity

- The **technological singularity** (also, simply, the **singularity**)[1] is the hypothesis that the invention of artificial superintelligence will abruptly trigger runaway technological growth, resulting in unfathomable changes to human civilization[3]
- Ray Kurzweil predicts the singularity to occur around 2045[7]

[1] M. John, "When Is the Singularity? Probably Not in Your Lifetime." The New York Times. The New York Times Company, 2016.
[2] Singularity hypotheses: A Scientific and Philosophical Assessment. Dordrecht: Springer. 2012. pp. 1–2.ISBN 9783642325601.
[3] R. Kurzweil, “The Singularity is Near”, pp. 135–136. Penguin Group, 2005.

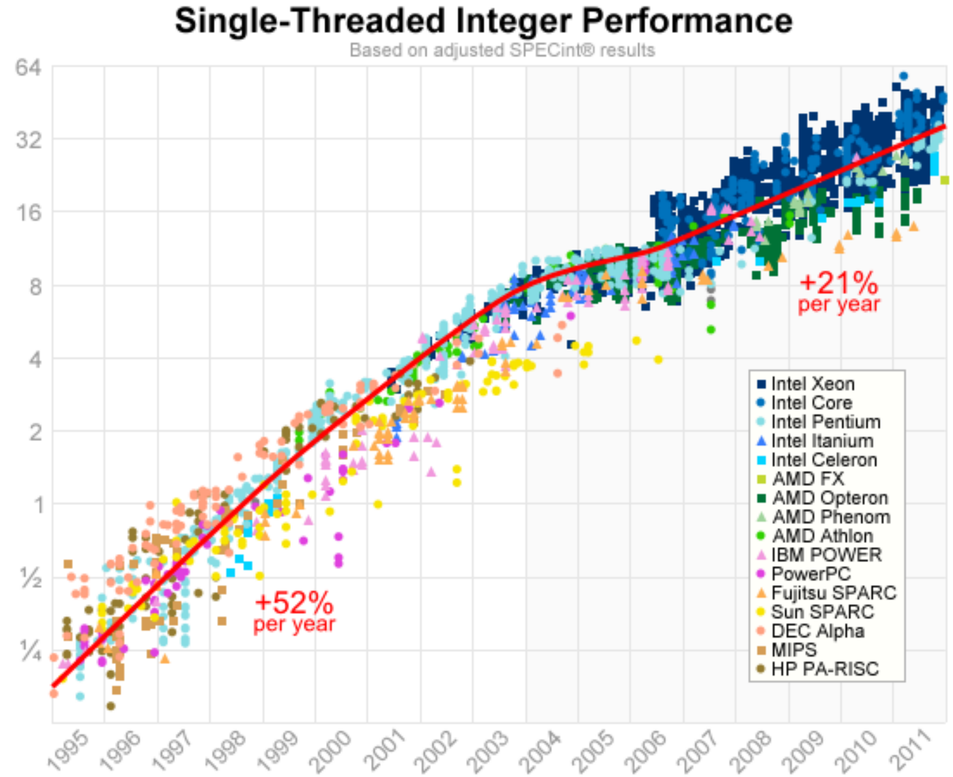
Why Deep Neural Networks?

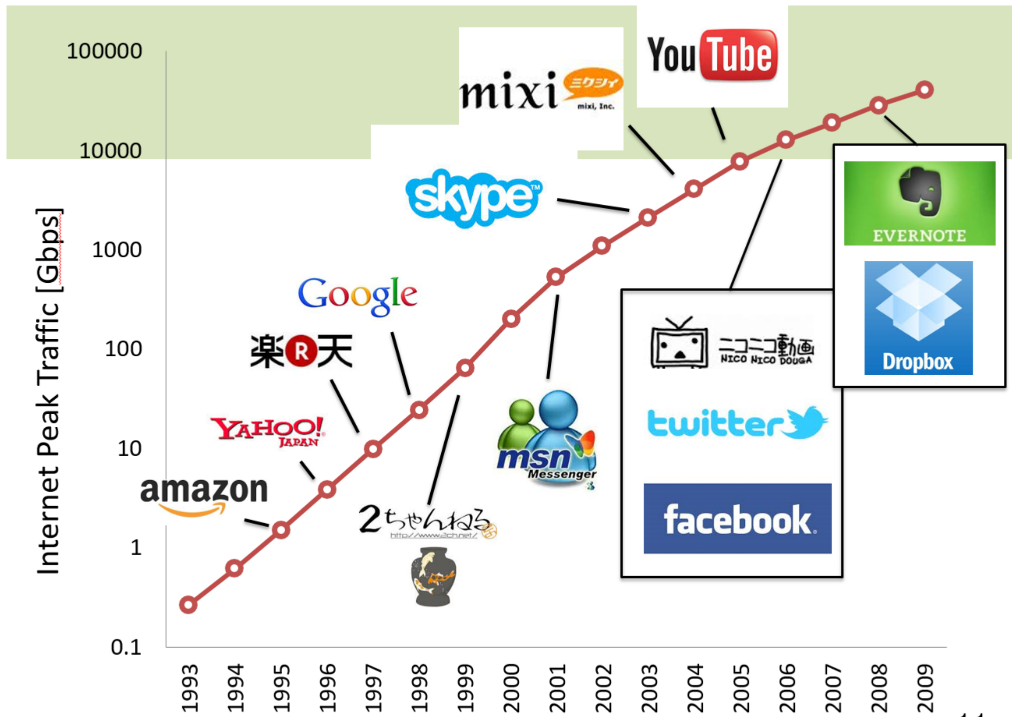


Big Data

Computational Power

Computational Power and Big Data





High performance computation, big data, and a progress of Algorithms

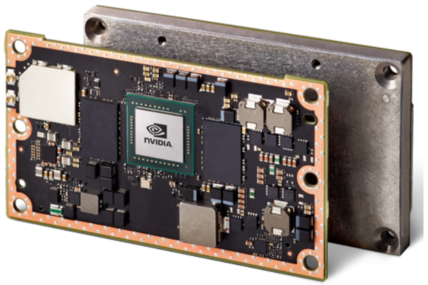
(Left): "Single-Threaded Integer Performance," 2016

(Right): Nakahara, "インターネットにおける検索エンジンの技術動向(In Japanese)," 2014

Inference Device

- Flexibility: R&S const, especially for new commoner Algs.
- Power performance efficiency
- FPGA→Better flexibility and power efficiency

CPU
(Raspberry Pi3)



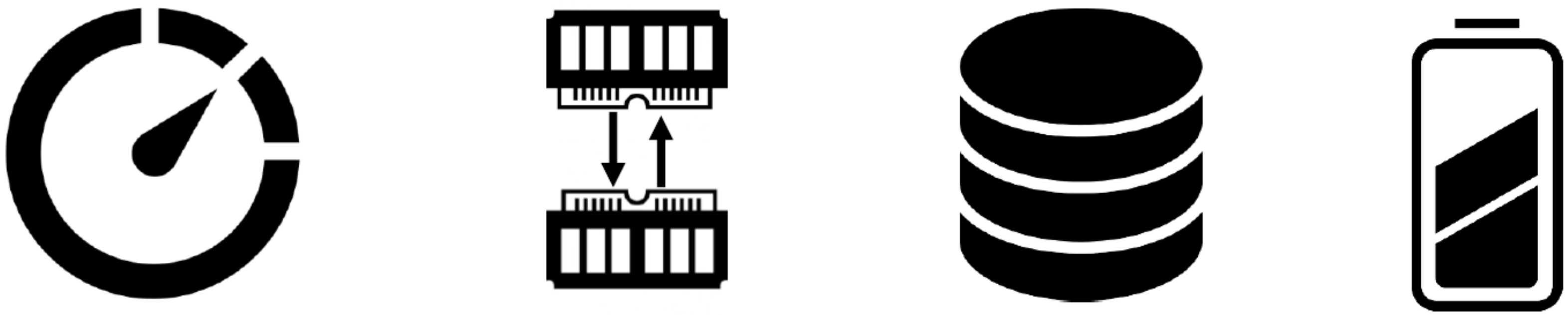
GPU
(Jetson TX2)

FPGA
(UltraZed)

ASIC
(Movidius)

Flexibility ⟷ **Power Performance Efficiency**

Requirements for DNNs



Performance	Memory Bandwidth	Storage	Power
Teraflops	100s of GB/s	10s of GBs	100s of Watts

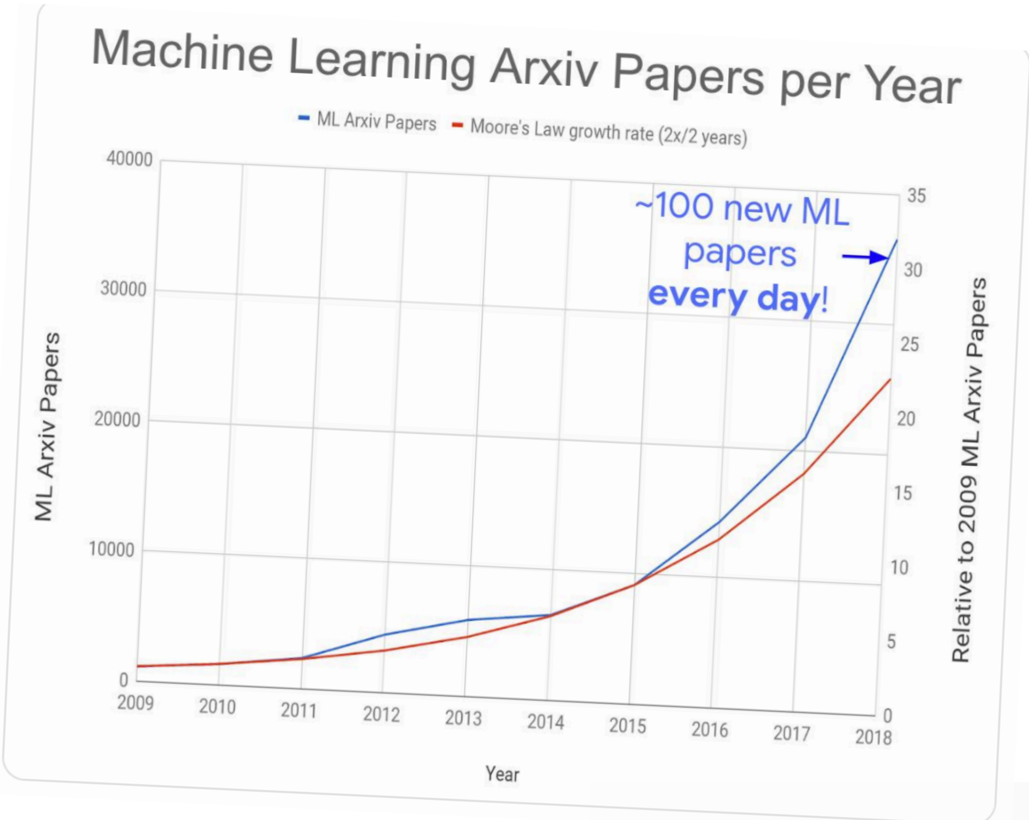
- 20 Billion MACs (Multiply ACcumulation operation)/image

J. Park, "Deep Neural Network SoC: Bringing deep learning to mobile devices," Deep Neural Network SoC Workshop, 2016.
J. Cong and B. Xiao, "Minimizing computation in convolutional neural networks," ***Artificial Neural Networks and Machine Learning (ICANN2014)*****, 2014, pp. 281-290.**

AI Platform

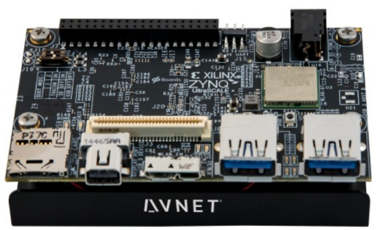
- Flexibility → R&D costs

  100 ML papers/day !!
- Power performance



CPU
(Raspberry Pi3)

GPU
(Jetson Nano)

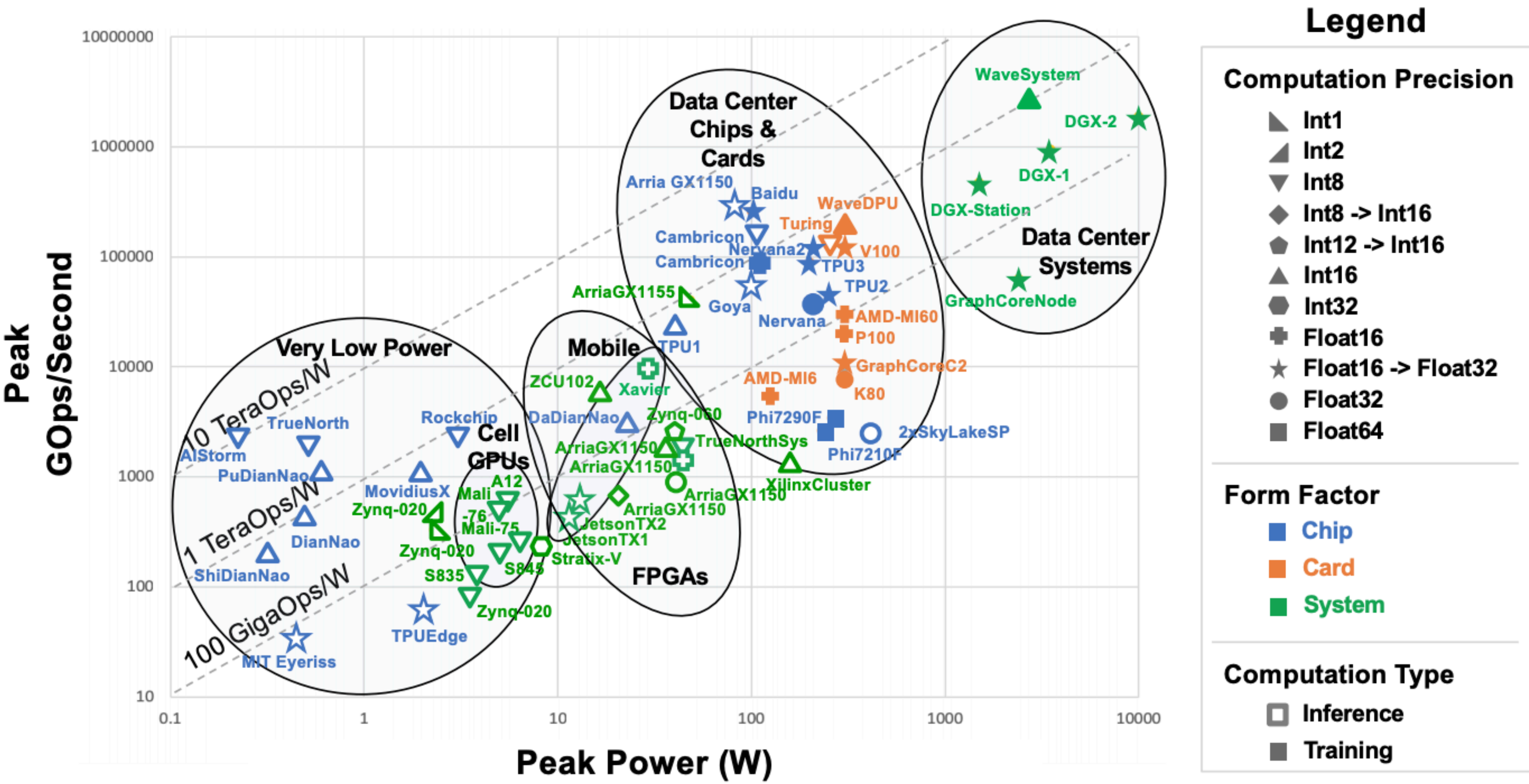


FPGA
(Ultra96)

ASIC
(Edge TPU)

Flexibility ⟷ **Power Performance Efficiency**

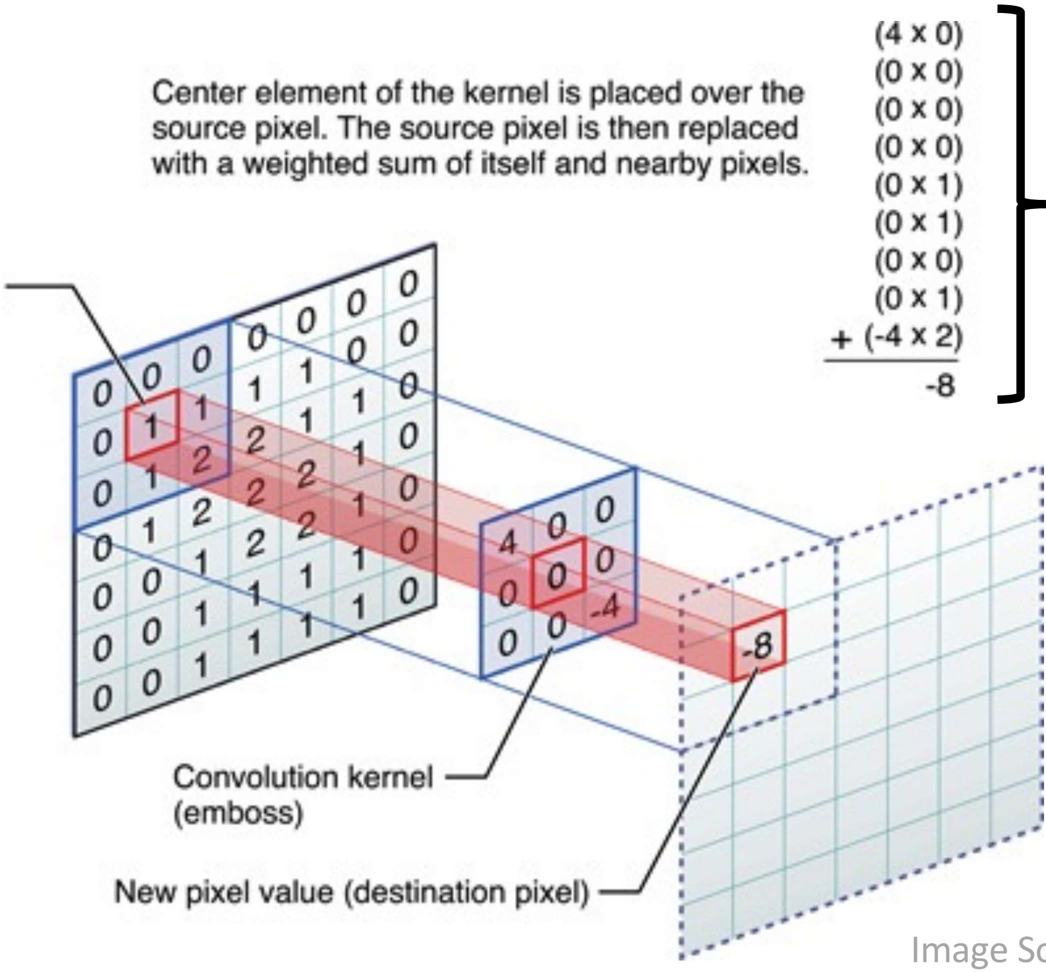
Hardware Platform Trend



A. Reuther et al., "Survey and Benchmarking of Machine Learning Accelerators," arXiv:1908.11348, Aug., 2019. https://arxiv.org/abs/1908.11348

Convolution Operation

- Applying multiple-accumulation (MAC) operations
- Occupies more than 90% of computation



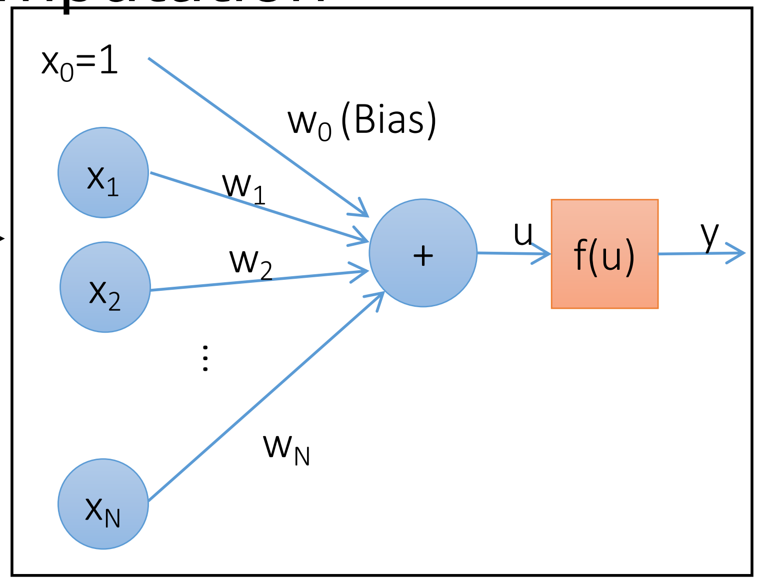


Image Source: http://i.stack.imgur.com/GvsBA.jpg

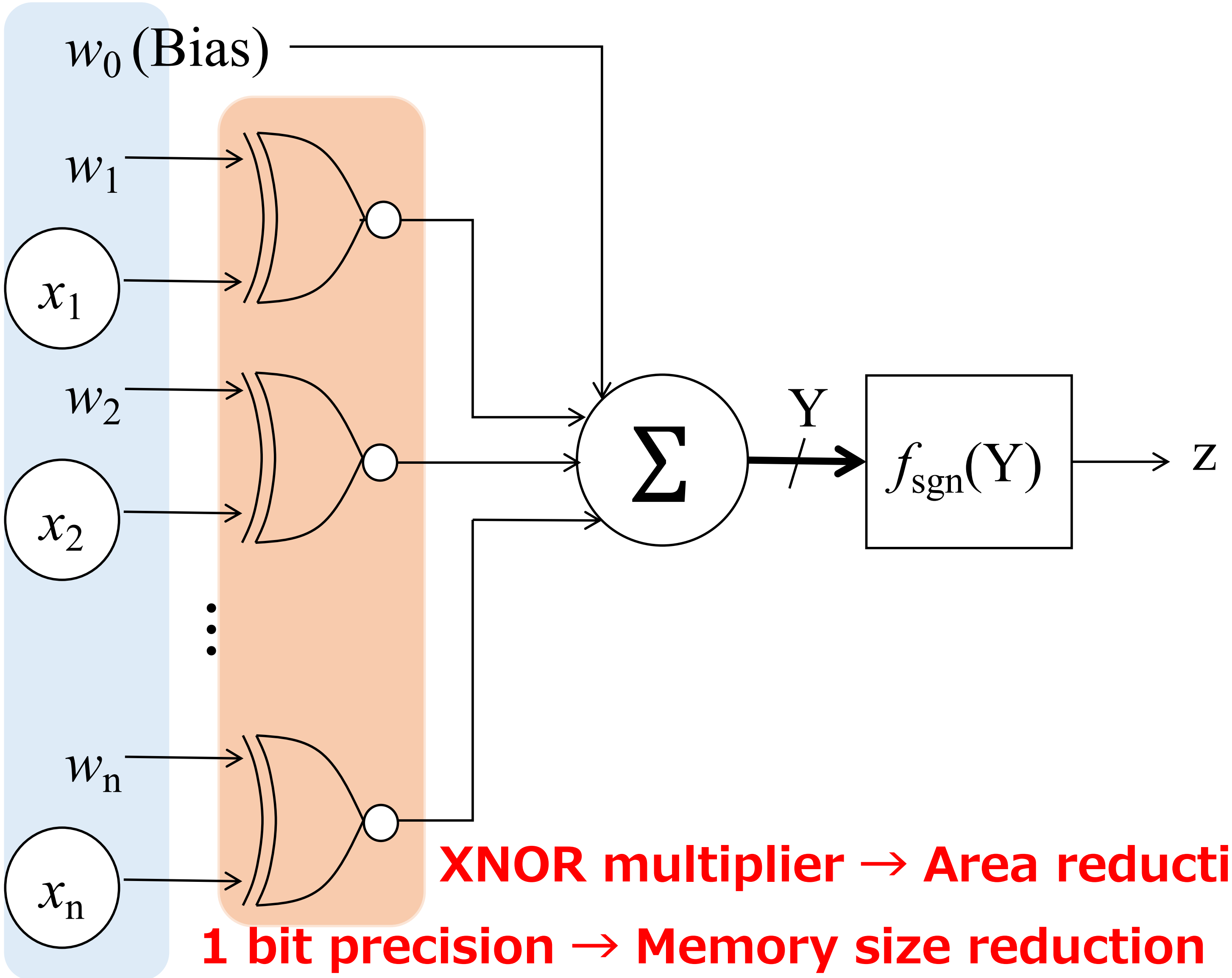
Binarized Neural Network

- 2-valued (-1/+1) multiplication
- Realized by an XNOR gate

x1	x2	Y
-1	-1	1
-1	+1	-1
+1	-1	-1
+1	+1	1

x1	x2	Y
0	0	1
0	1	0
1	0	0
1	1	1

Binarized CNN by XNORs



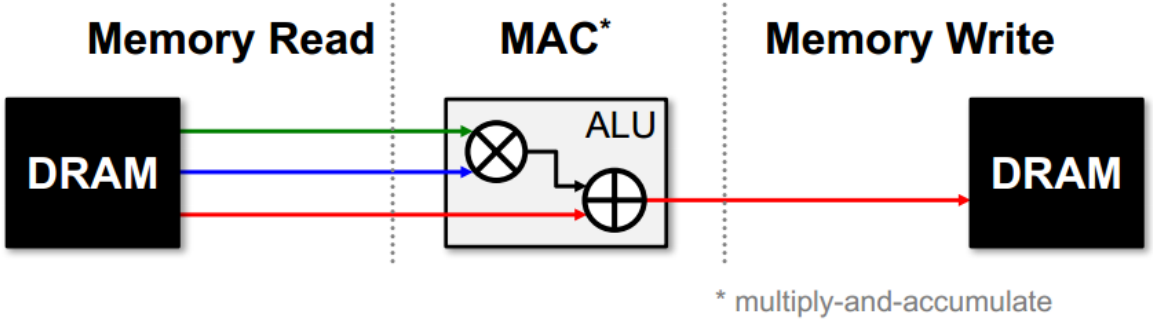
XNOR multiplier → Area reduction

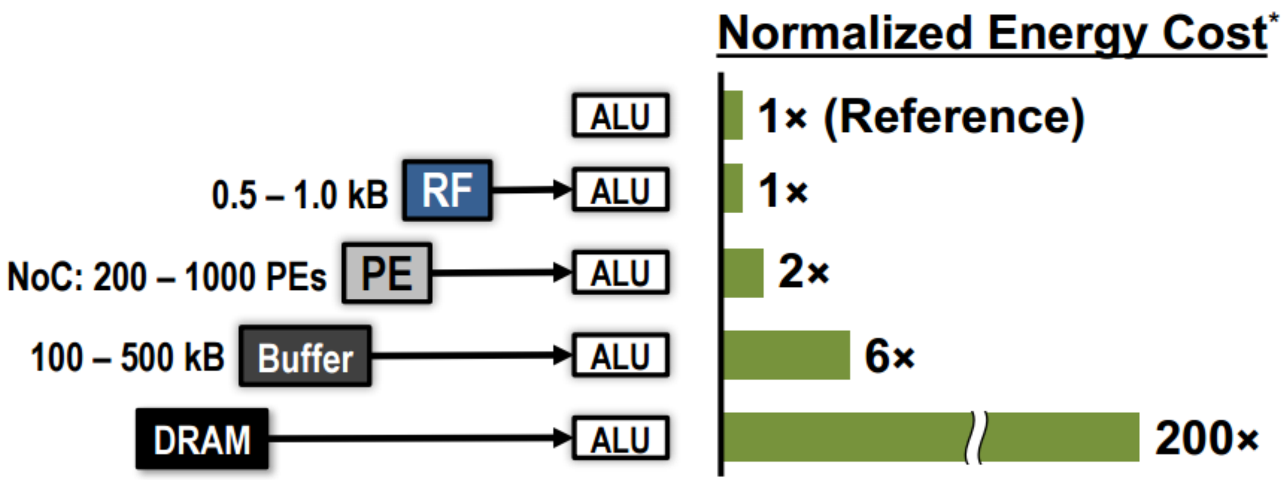
1 bit precision → Memory size reduction

Higher Power Efficiency

- Distance for the memory and ALU∝Power

→ On-chip memory realization





E. Joel et al., “Tutorial on Hardware Architectures for Deep Neural Networks,” MICRO-49, 2016.

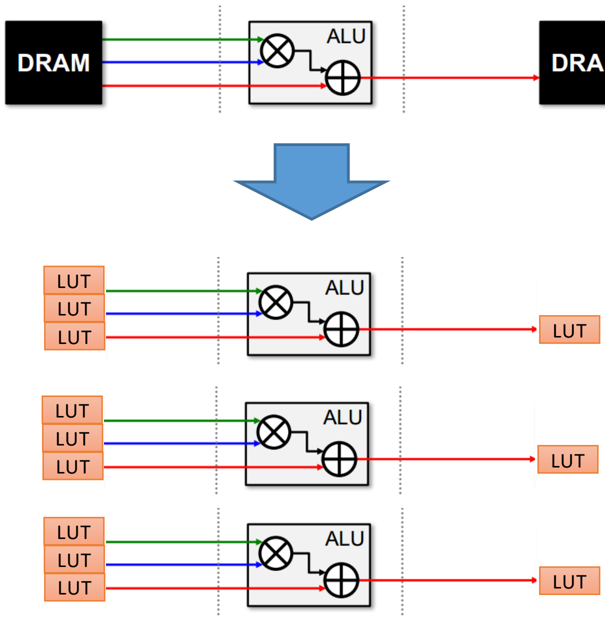
On-chip Memory Realization

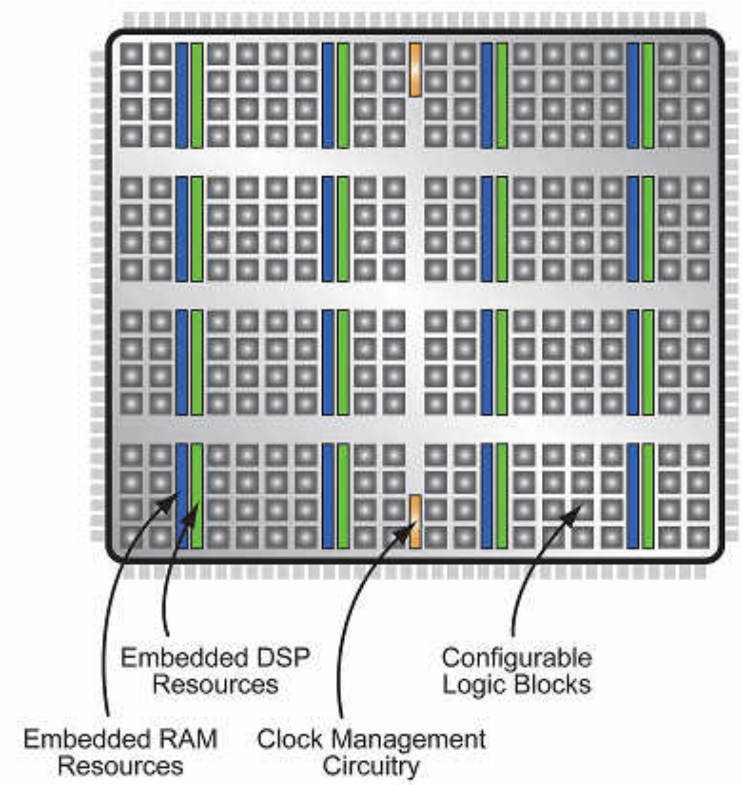
- FPGA on-chip memories
  - BRAM (Block RAM) → 100s～1,000s
  - Distributed RAM (LUT) → 10,000s～100,000s

→ Small size, however, wide band

Cf. Jetson TX1(GPU) LPDDR4, 25.6GB/s

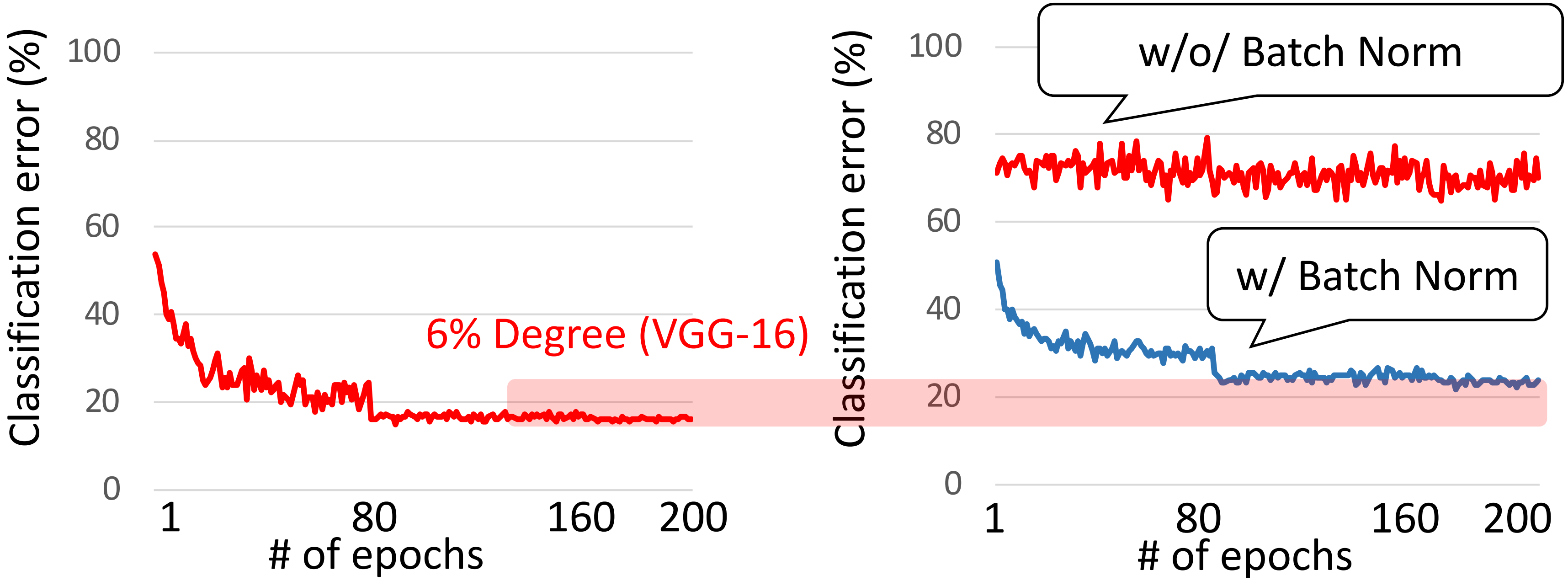
10,000@100MHz → 125GB/s





Error Rate Reduction

- Introduce a batch normalization

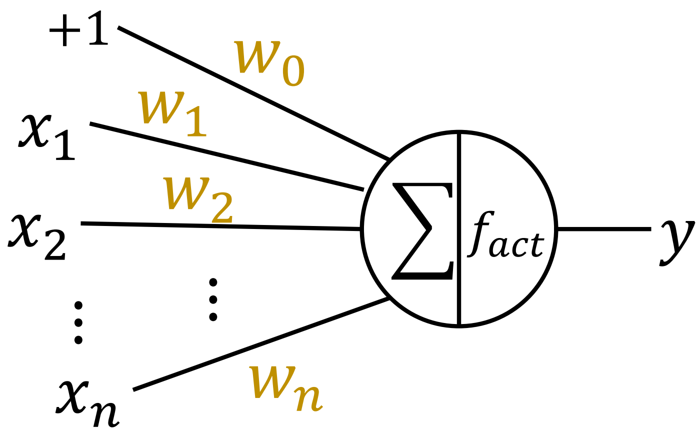


(a) float32 bit precision CNN

(b) Binarized CNN

H. Nakahara et al., “A memory-based binarized convolutional deep neural network,” FPT2016, pp285-288, 2016.

Ternary Weight Binary Activation Neuron



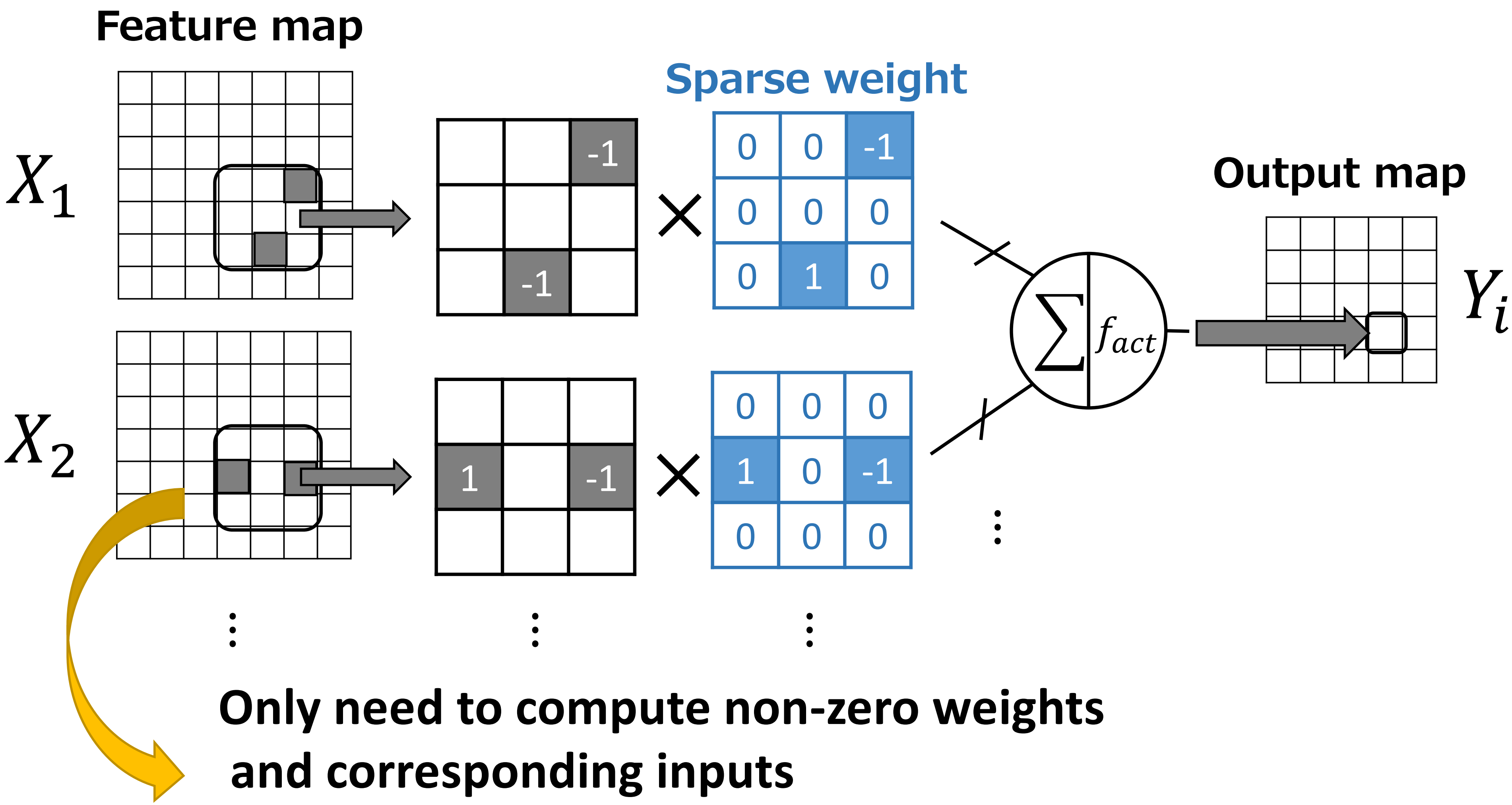
Neuron Model

$$\begin{bmatrix} -1 & 0 & \cdots & +1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & -1 & \cdots & 0 \end{bmatrix}$$

Sparse Matrix

- Define the weight to ternary one $w_i \in \{-1,0,+1\}$, x_i, $\boldsymbol{y} \in \{-1,+1\}$,
  - ➢Improve recognition accuracy by expression ability
- Since multiplication by zero weight is equal to skipping, the number of mult. can be reduced
- Contributions
  - ➢Develop training method
  - ➢Evaluation of reduction (zero) ratio by using benchmark

Skip Operation for Sparse Convolution

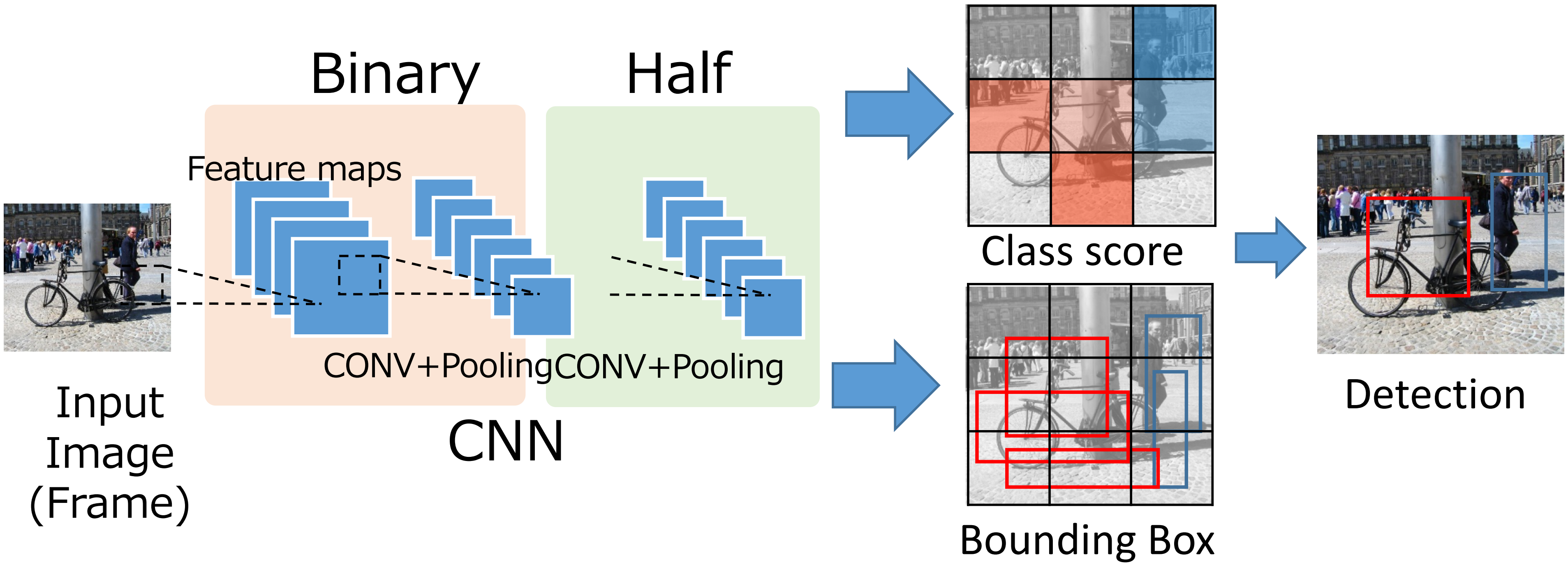


Only need to compute non-zero weights and corresponding inputs

- **Reduction of number of calculations**
- **Memory size reduction**

Mixed-Precision

- Mandatory for more complex detector
  - Former: Binary precision … Area and performance
  - Latter: Higher precision … Regression (Accuracy)



H. Nakahara et al., "A Lightweight YOLOv2: A Binarized CNN with A Parallel Support Vector Regression for an FPGA," Int'l Symp. on FPGA (ISFPGA), 2018.

Homework 2

1. (Mandatory) How do you think a “Technological singularity”? Near/Far/Never? why? and what’s happen?

Deadline is 25th, Nov., 2019

Send an E-mail to nakahara@ict.e.titech.ac.jp

with entitled “Homework 2 (your name)”