IC/IP Design Process
in AI/ML-Big Data Era

中国科学院大学
University of Chinese Academy of Sciences

Machine Learning and AI-IC

Applications of AI-ICs

Chun-Zhang Chen, Ph.D.

June 28 - July 2, 2021




中国科学院大学2021年夏季

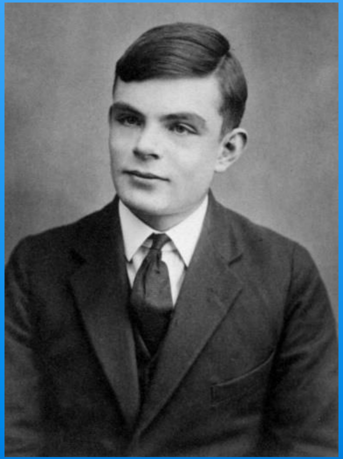
Definition/Classification of AI

- What is AI? Merriam-Webster Dictionary:
  - “An area of computer science that deals with giving machines the ability to seem like they have human intelligence”
  - “The power of a machine to copy intelligent human behavior”
- How AI is classified?
  - Artificial Weak/Narrow Intelligence (**ANI**)
    - Focuses on improvement of individual ability, *e.g.* Siri
  - Artificial General Intelligence (**AGI**)
    - On humankind, human’s brains, *e.g.* TrueNorth
  - Artificial Superintelligence (**ASI**)
    - Smarter than human brains, including innovation, recognition and social

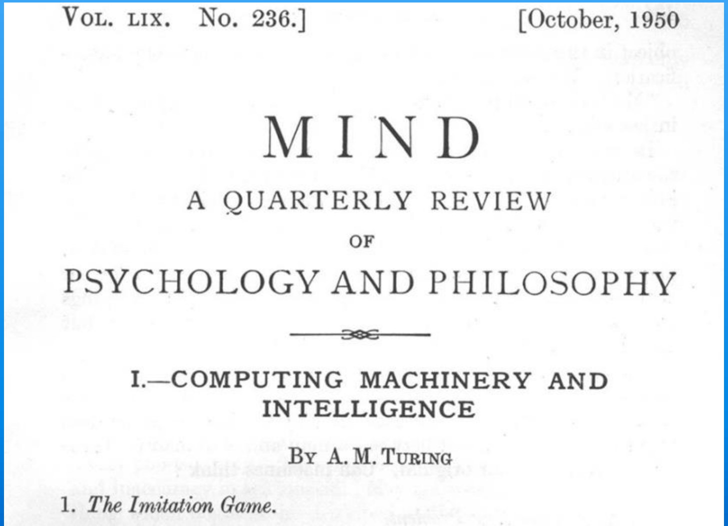
Alan Turing and AI

the "Nobel Prize of computing"

- ACM: A.M. Turing's Award By Year (since 1966)
- Turing is widely considered to be *the father of theoretical computer science and artificial intelligence*.



Turing aged 16



VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

The Birth of AI (1952-56)

University of Chinese Academy of Sciences

John McCarthy (Stanford)

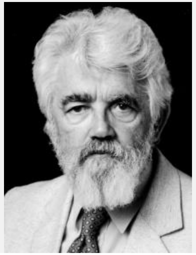
Marvin Minsky (MIT)

Trenchard More (IBM ret'd)

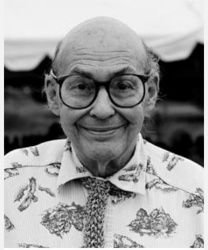
Ray Solomonoff (London)

Oliver Selfridge (MIT)

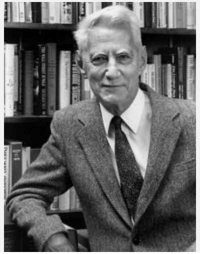
Dartmouth Summer Research Project on Artificial Intelligence 1956



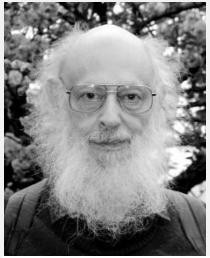
John McCarthy,
"AI" 1955



Marvin Minsky,
MIT AI Lab



Claude Shannon,
MIT Boolean alg.



Ray Solomonoff,
Inductive Inference

Allen Newell,
Turing 1975

Herbert Simon,
Nobel78,Turing75

Arthur Samuel,
"ML" 1959

Oliver Selfridge,
Machine Perc.

Nat Rochester *(IBM 701);* **Trenchard More**

Julian Bigelow,
IAS/MANIAC

Source: https://en.wikipedia.org/wiki/Dartmouth_workshop

The Past 60+ Years of AI

University of Chinese Academy of Sciences

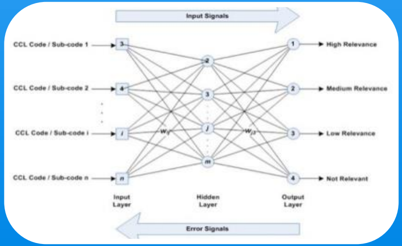
"The First Wave of AI (1956-76)"

The Golden Years of AI (1956-74);
H. Simon & A. Newell 1975 Turing

"Winter Seasons of AI (1976-05)"

PC Market ; IBM-Deep Blue
1997 & Jeopardy 2011

DNN Algorithm, Backpropagation *(1986)*;

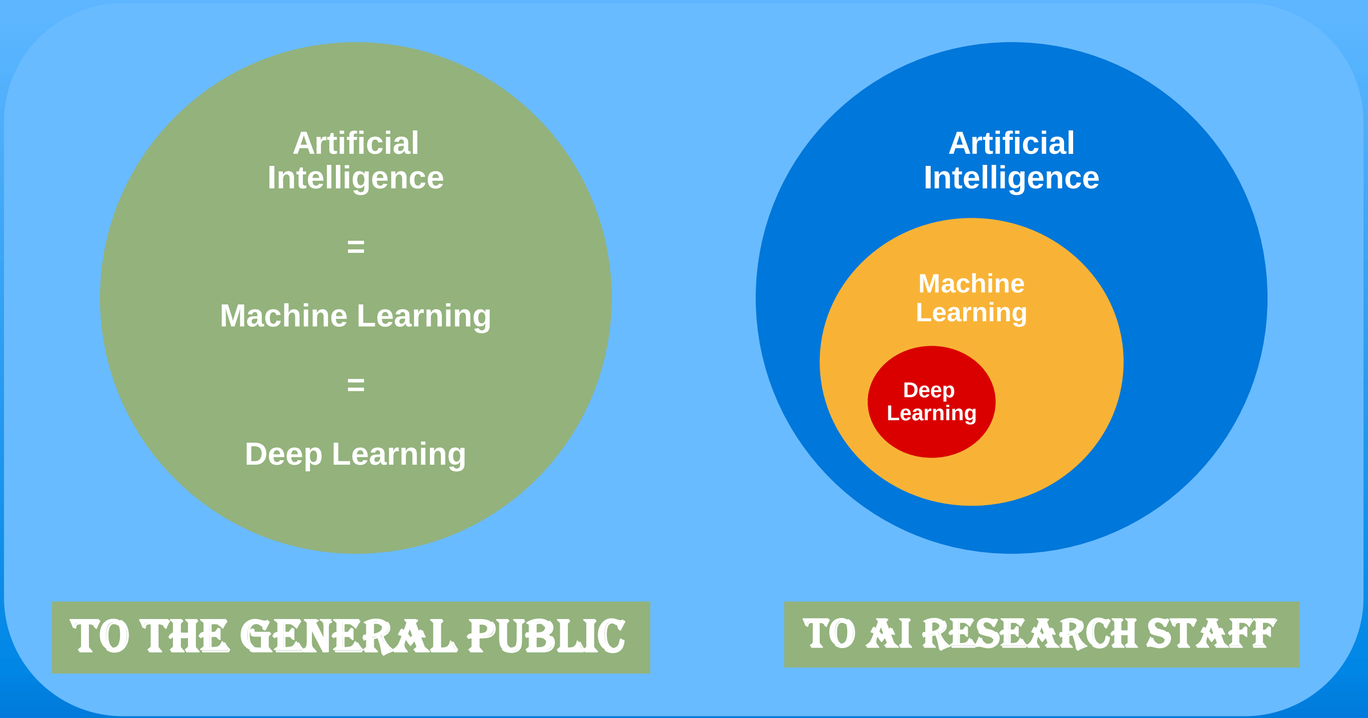


Next Step:
BP? Capsule?

Machine Learning and AI-IC

Views of AI/ML/DL

Can AI Replace Human Beings?



ML: Schools? Algorithms?

Schools of Machine Learning

- Symbolic *(e.g. Frank Rosenblatt, 1957)*
  - *logicism, aka logicism, psychology, or computerism*
- **Connectionism** *(e.g. Geoffrey Hinton, 1986 Nature)*
  - *Aka bionicsism or physiologism*
- Actionism
  - *Aka evolutionism or cybernetícism*
- **Probabilistic Graphical Models**
  - *E.g. Bayes network, Random mean forest*

Key Aspects of ML, Algorithms

- **Types of ML Algorithm**
  - Supervised learning
  - Semi-supervised learning
  - Unsupervised learning
  - Reinforcement learning

- Regression, KNN, SVM, Boosting (Ada, X-G), Decision Tree, Random Forest etc.
- Clustering (e.g. K-means, GMM), Dimensionality Reduction, PCA, ICA, etc.

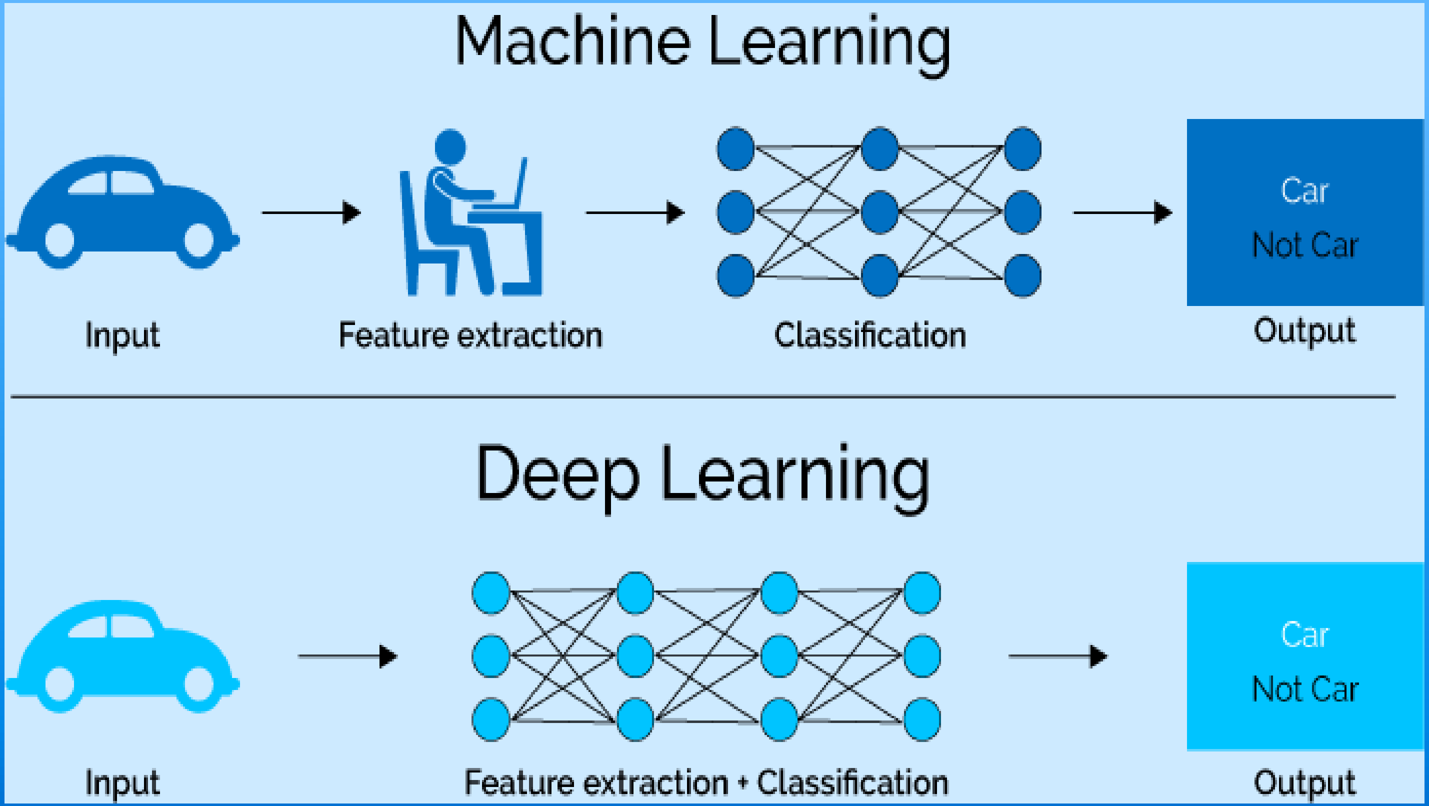
Machine Learning and AI-IC

University of Chinese Academy of Sciences

- Machine Learning Methods
- Machine Learning and Deep Learning
- CPU and GPU in AI-IC
- Applications of AI-IC
- Discussion

Comparison of Learning Flow in ML and DL

University of Chinese Academy of Sciences



Types of Deep Learning Algorithms

University of Chinese Academy of Sciences

- Artificial Neural Network
  - *ANN,*
    - *Artificial Neural Network*
  - FNN,
    - Feedforward Neural Network
  - CNN,
    - Convolutional Neural Network
    - *Cellular Neural/Nonlinear Network*
  - RNN,
    - Recurrent Neural Network
    - LSTM, modified RNN
    - Transformer(s) from Google
  - GAN,
    - Generative Adversarial Network

2018 ACM A.M. Turing Award

University of Chinese Academy of Sciences

- Yoshua Bengio,
  - Professor at the University of Montreal and Scientific Director at Mila, Quebec's Artificial Intelligence Institute
- Geoffrey Hinton,
  - VP & Engineering Fellow of Google, Chief Scientific Adviser of The Vector Institute, and Univ. Prof. Emeritus at Univ. Toronto
- Yann LeCun,
  - Professor at New York University and VP and Chief AI Scientist at Facebook

ACM DL DIGITAL LIBRARY

ML/DL Applications

University of Chinese Academy of Sciences

- **ML Applications**
  - Image Recognition. One of the most common uses of machine learning
  - Speech Recognition. SR the translation of spoken words into text.
  - Medical Diagnosis. ML provides methods, techniques, and tools that can help solving diagnostic...
  - Statistical Arbitrage.

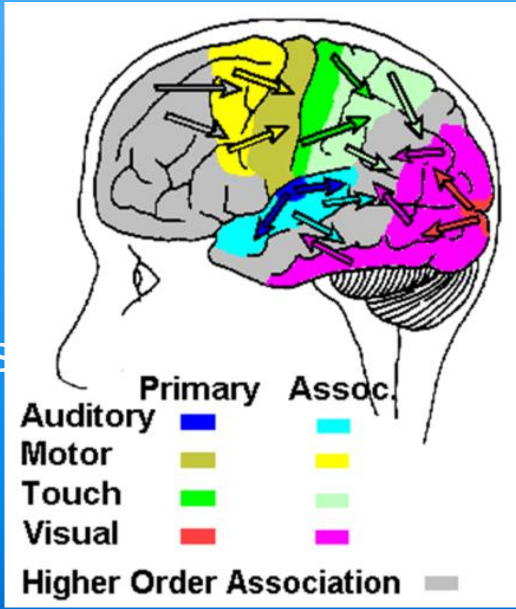
- **DL Applications**
  - Self-driving cars
  - Deep Learning in Healthcare
  - Voice Search & Voice-Activated Assistants
  - Automatic Colorization of Black and White Images.
  - Automatically Adding Sounds To Silent Movies
  - Automatic Machine Translation

In Human Cerebrum

- Functional areas
  - Frontal lobe
    - Thinking/plan/short mem.
  - Parietal lobe
    - Touch/smell/taste
  - Occipital lobe
    - Visual activity
  - Temporal lobe
    - Memories

- Under cerebral cortex

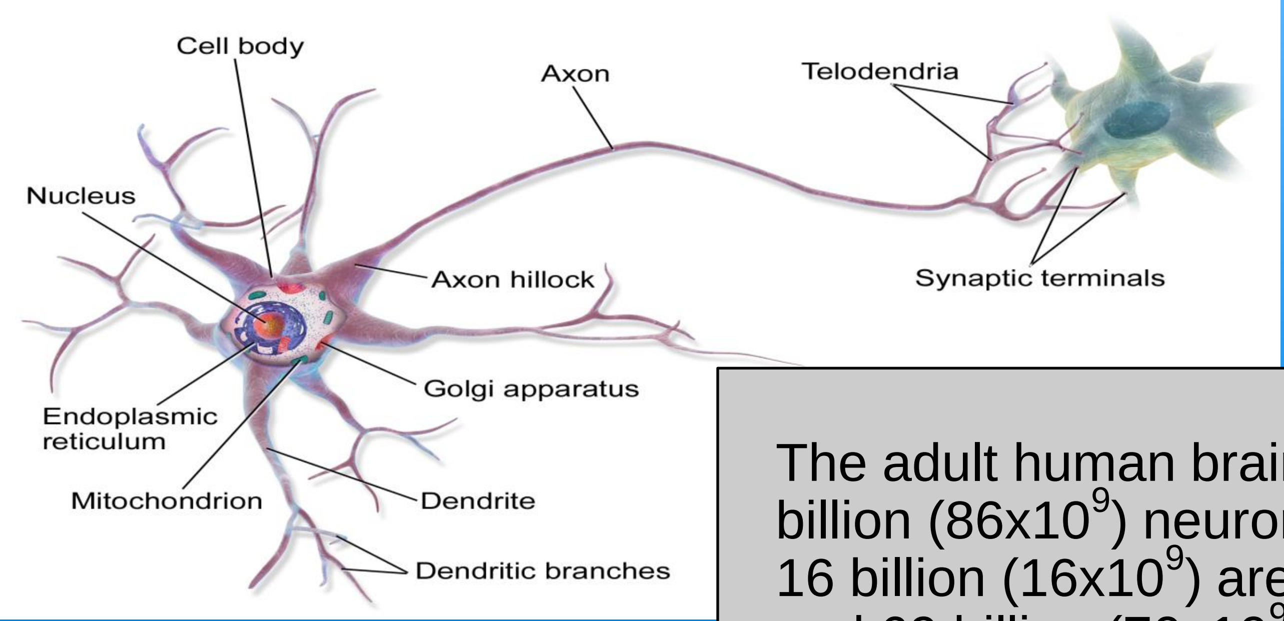
  (Primary, Association)
  - Auditory,
  - Motor,
  - Touch,
  - Visual,
  - High order as

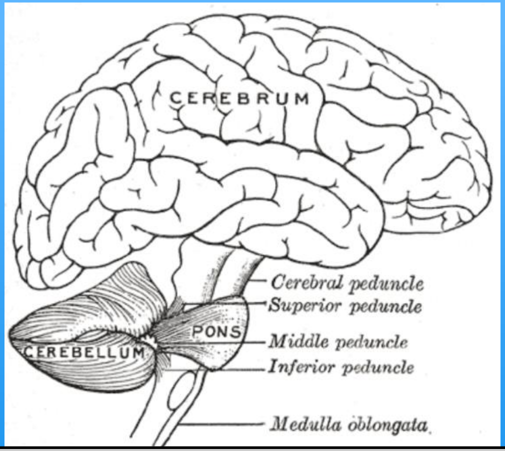


Neurons and Synapses (Neural Network)

University of Chinese Academy of Sciences

- Human brain: Cerebrum, Cerebellum





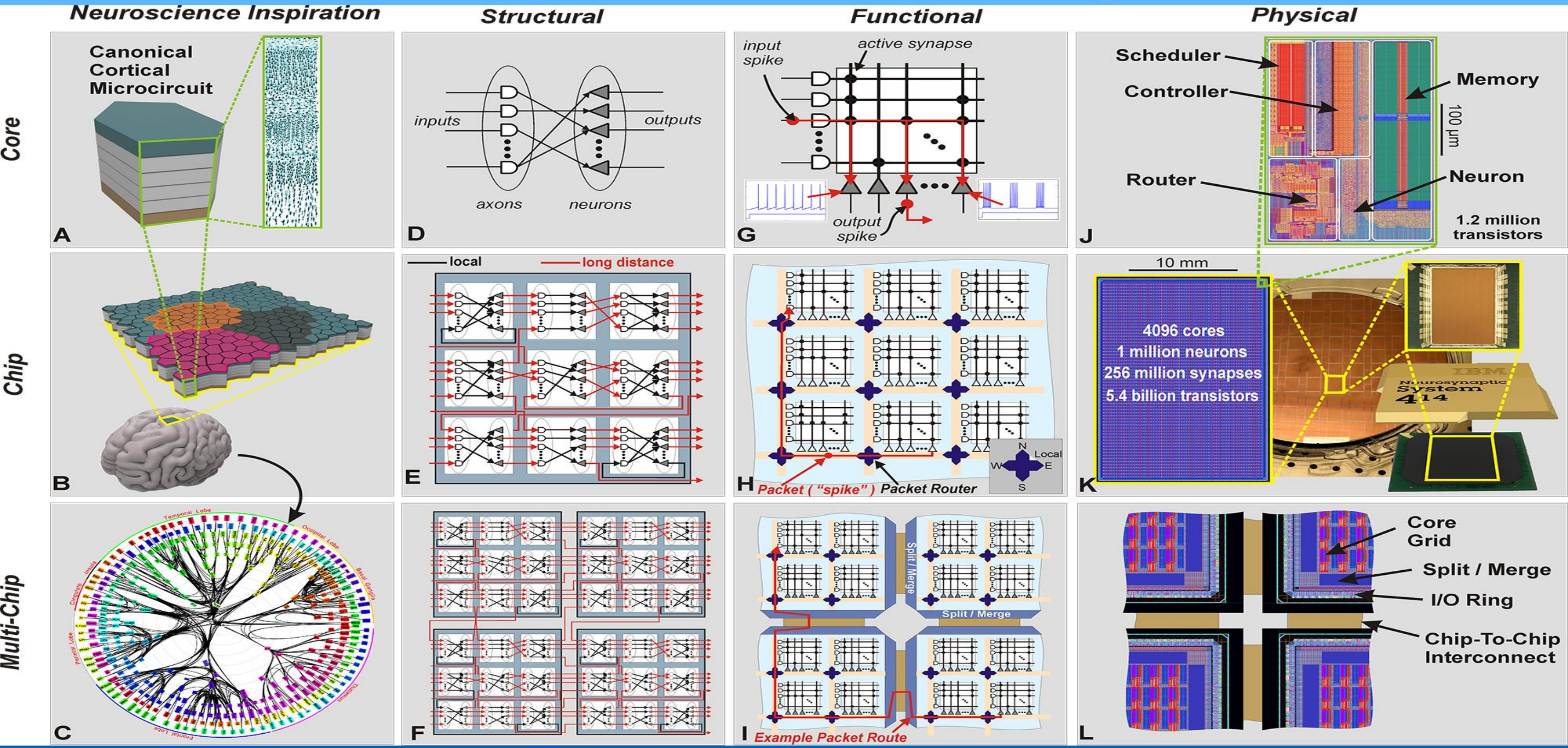
The adult human brain contains about 85-86 billion ($86x10^9$) neurons,[38][39] of which 16 billion ($16x10^9$) are in the cerebral cortex and 69 billion ($70x10^9$) in the cerebellum.[39]

Machine Learning and AI-IC

Neuromorphic AI-IC, IBM 2014

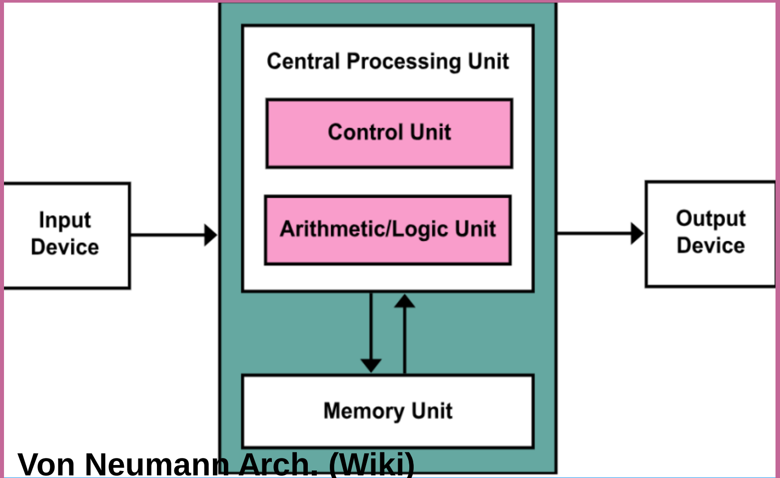
- **A neuromorphic CMOS IC, TrueNorth chip**
  - Many cores, 4096 cores, simulating a total $>10^6$ neurons
  - The programmable synapses is $>268 \times 10^6$ (2^{28})
- **Contains 5.4×10^9 transistors (Sg28nm)**
  - At low T, 70 mW, about 1/10,000th of conventional MPU
- **Application**
  - SyNAPSE 16 chips for DARPA

IBM SyNAPSE Project and TrueNorth Chip

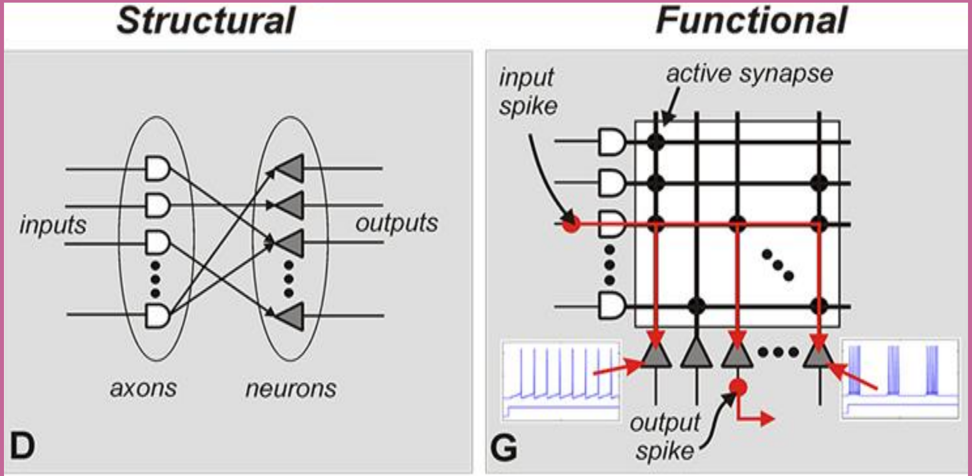


CPU and von Neumann Bottleneck

University of Chinese Academy of Sciences



Von Neumann Arch. (Wiki)

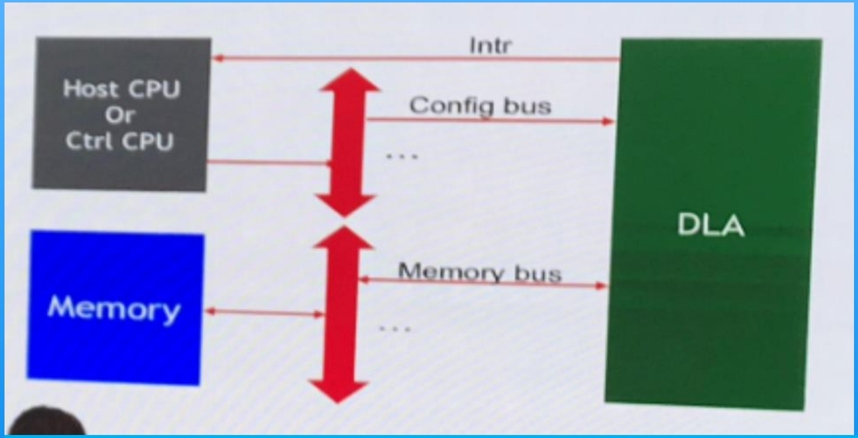


Von Neumann Computing
Serial Computing
Separate Memory
High Precision

Neuromorphic Computing
Highly Parallel
In Memory Computing
Tolerance to Low Precision

GPU for Deep Learning

- Nvidia, 1999 GeForce 256: "*the world's first GPU*"
- a "single-chip processor with integrated transform, lighting, triangle setup/clipping, & rendering engines"
- GPU for DL and EC



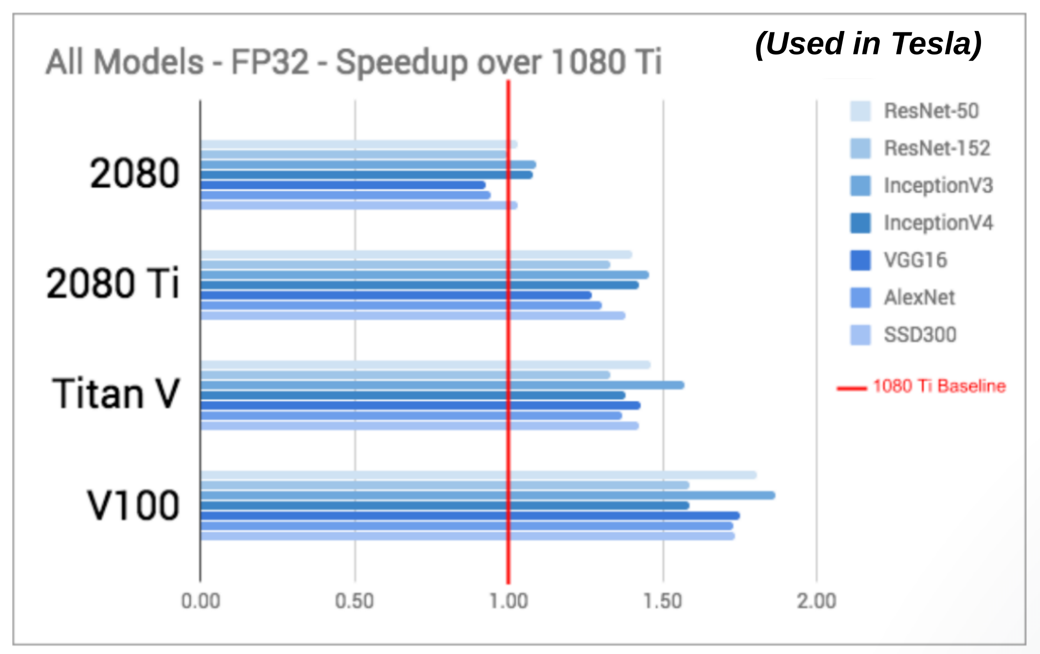
Nvidia DLA (Deep Learning Accelerator)
"DL Solution for Edge Computing"
(SiFive-RISC-V, 03/29/19 Shanghai)

GPU Benchmarks for DL

University of Chinese Academy of Sciences

- ResNet-50
- ResNet-152
- Inception-V3
- Inception-V4
- VGG16
- AlexNet
- SSD300

- GPU tested/used in *Tesla*

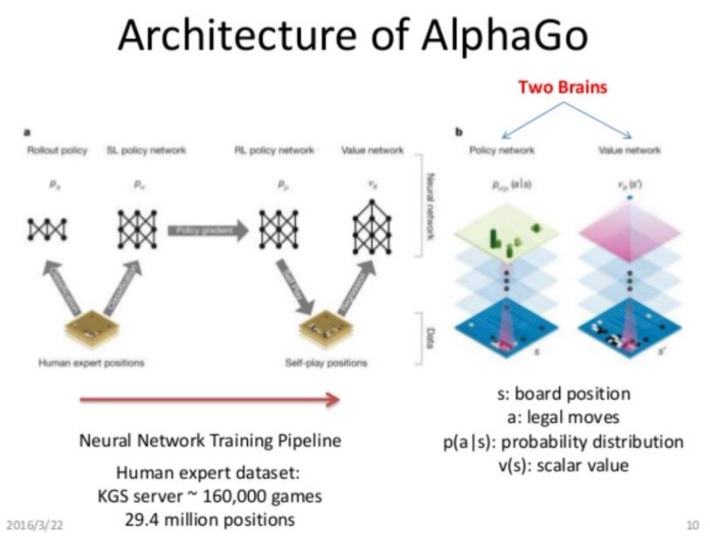


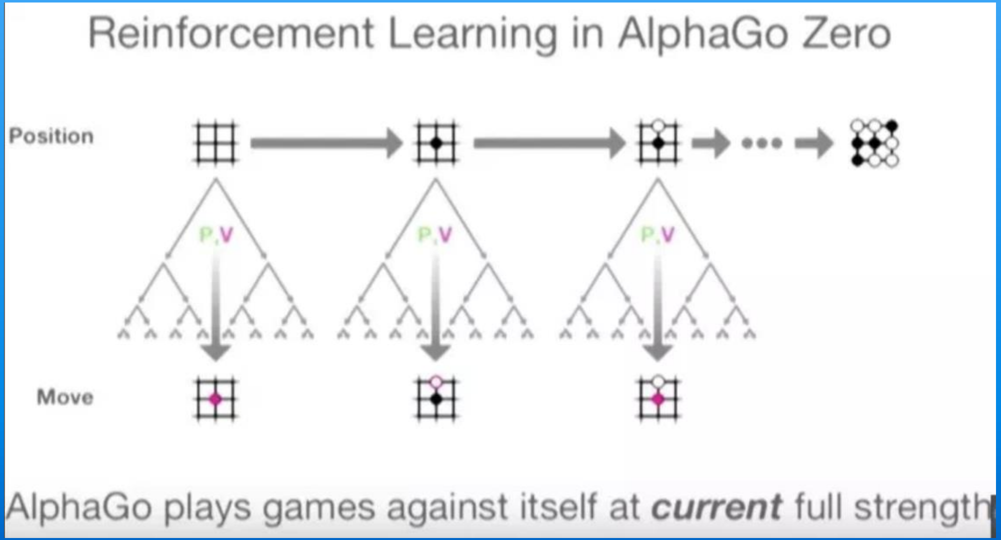
GPU from AMD

- Acquisition of ATI in 2006
  - CPU chips: *A4, A6, A8*
  - *Fusion* chip (CPU+GPU)
  - *Llano* (Accelerated) APU

AI-IC in AlphaGo/AlphaGo Zero

- Architecture based on CPU + GPU
  - AlphaGo (Oct. 15; Mar. 16; Mar. 17)
  - AlphaGo Zero (10/19/17)

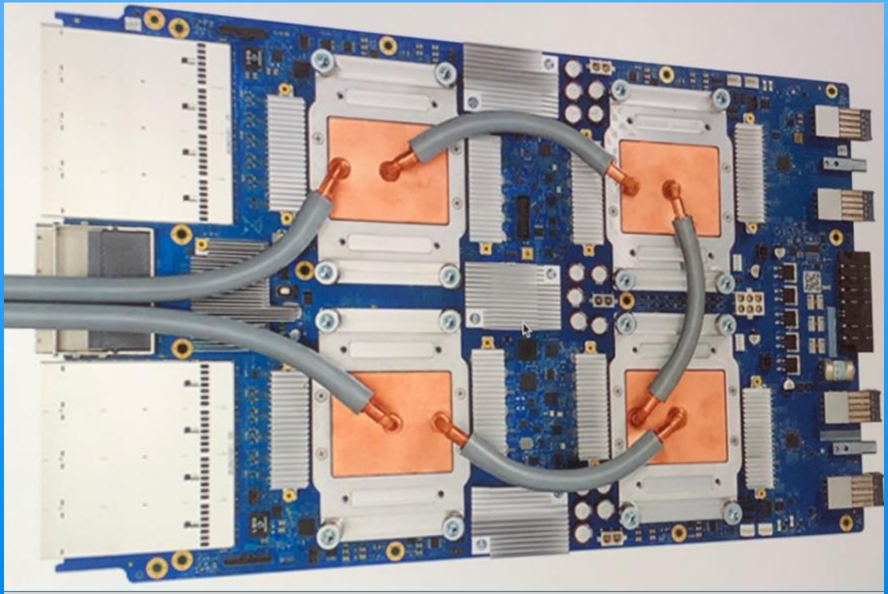




TPU used in AlphaGo for ML

University of Chinese Academy of Sciences

- Can be accessed from Cloud
- 2018 Edge TPU: IoT EC
- TPU in AlphaGo 2016 (Lee Sedol)
  - 2016 Gen1: CISC 8-bit, PCIe 3.0, 28nm, $\leq$331 mm^2,
  - 2017 Gen2: 4x16GB HBM 4x600GB/s, 180TFLOPS→ 11.5PFLOPS
  - 2018 Gen3: 2x of Gen2, 1024 chips per pod, 28nm, 700MHz, 40W

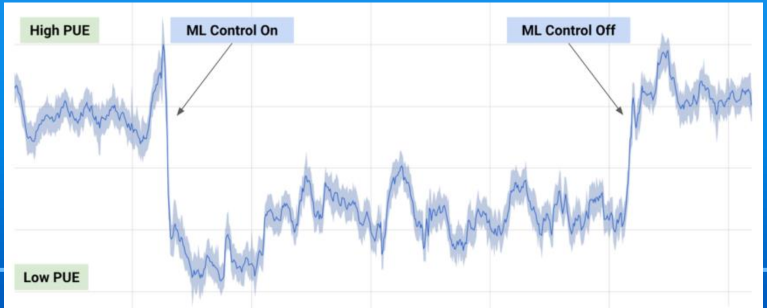


2018, TPU 3.0, 100 PFLOPS, Google

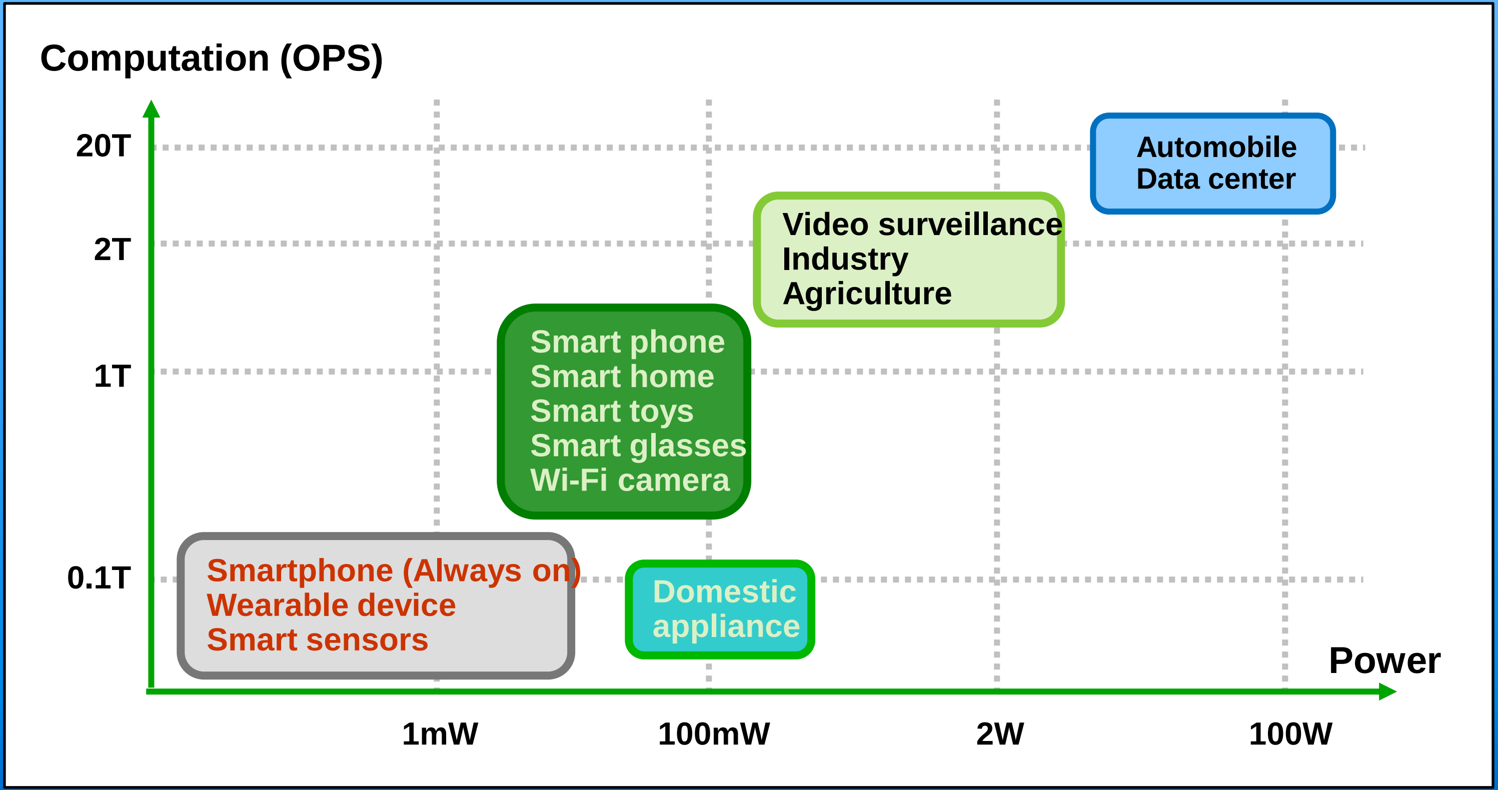
AlphaGo, AlphaStar, AlphaZero, AlphaFold

TPU used in TensorFlow

- TPU (announced at Google I/O, Mtn View))
  - TPU 1.0 (05/20/16); TPU 2.0 (05/23/17); TPU 3.0 (5/8-5/10/18)
- TensorFlow, ASIC (CPU+GPU)
  - https://www.tensorflow.org/
  - Feb 15, 2017, **TensorFlow 1.0** [09/27/16 → 11/06/16 → 02/15/17]
  - Nov 04, 2018, **TensorFlow 2.0**
- DeepMind,

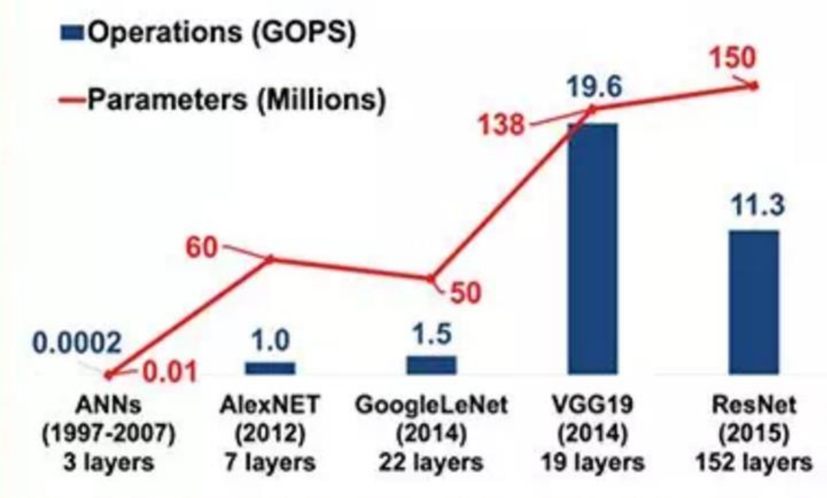


Computation Needs vs Power Limits

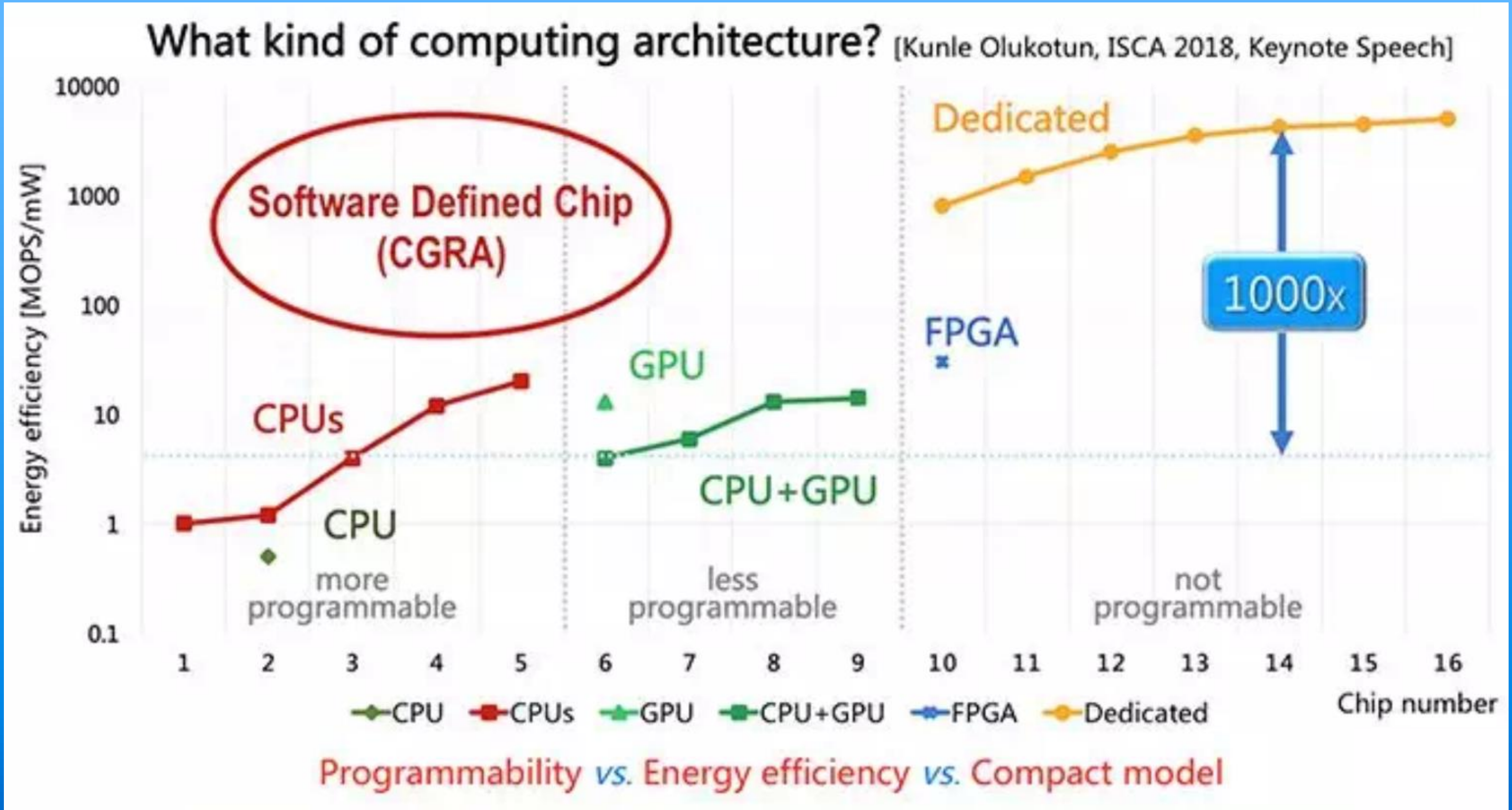


Features of AI-IC

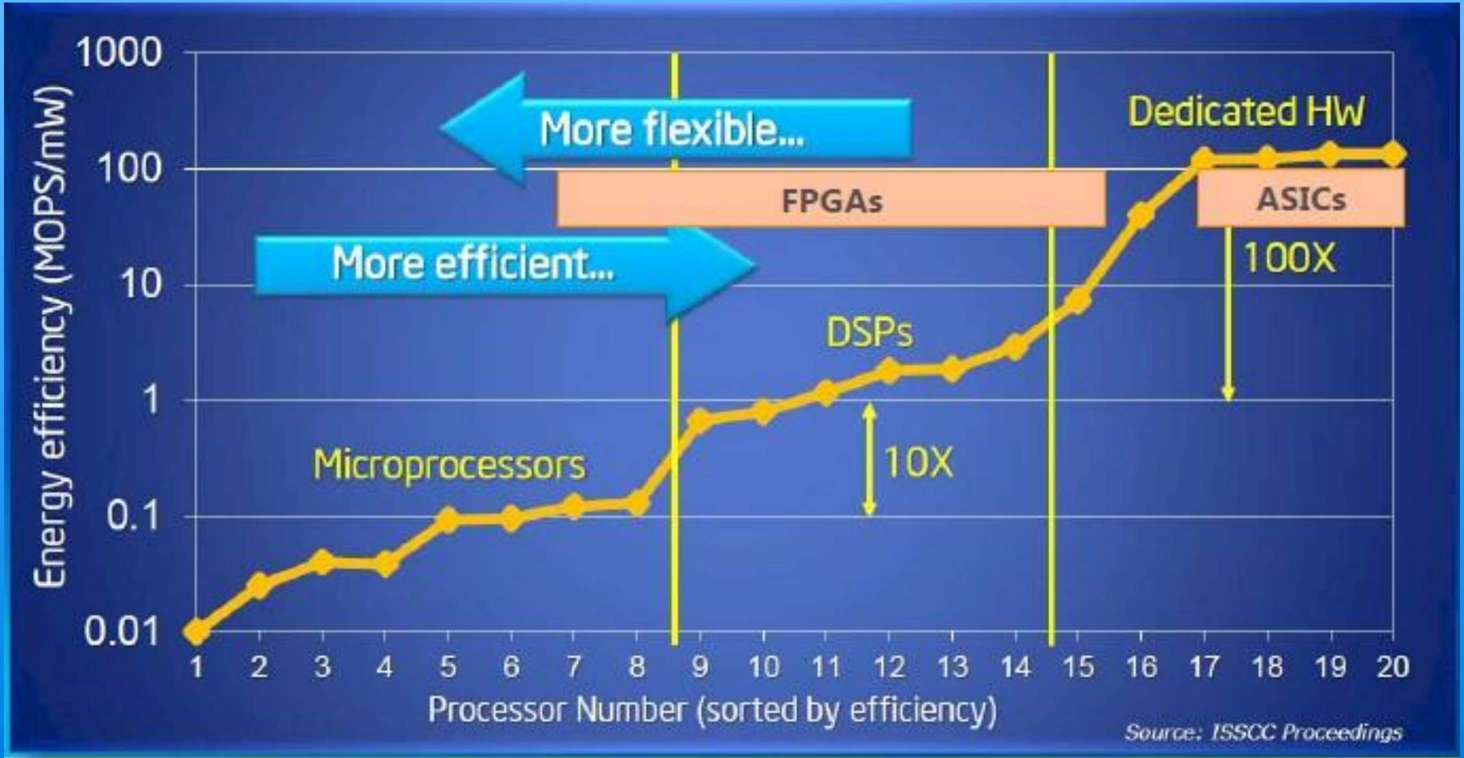
- Requirements for an AI-Chip
  - Programmability
  - **High Energy Efficiency**
  - Common use for DL
- CNN becomes most popular
  - Pattern (LeNet)
  - Image (AlexNet)
  - Vision (LRCN net)
- To meet **Edge Computing** with high Energy Efficiency
- OPS and Parameters



Progress of Energy Efficiency (OPS/W)

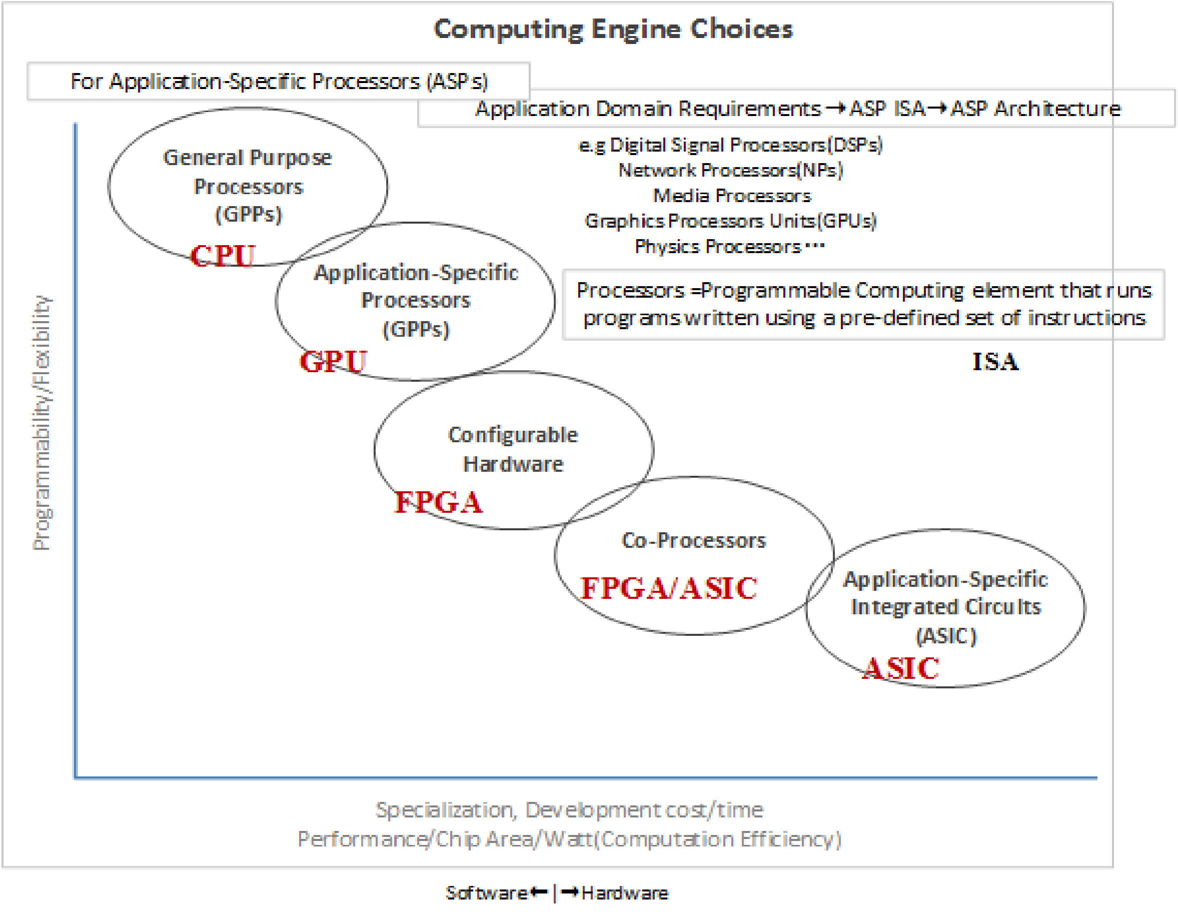


Comparison of Energy Efficiency (Processors)



Source: Bob Broderson, Berkeley Wireless group

Comparison of Computing Engines' Flexibility



Machine Learning and AI-IC

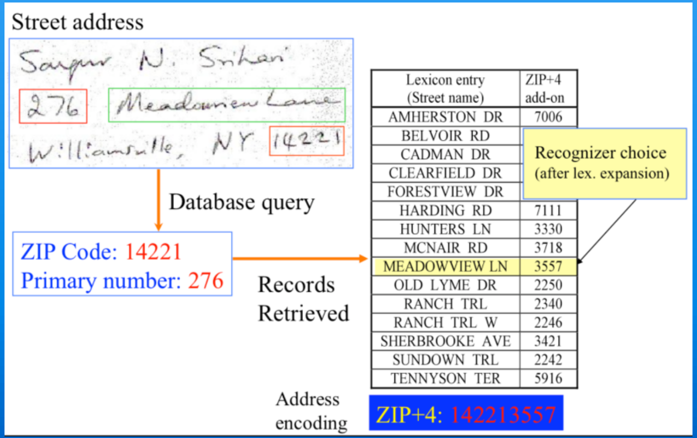
- Machine Learning Methods
- Machine Learning and Deep Learning
- CPU and GPU in AI-IC
- **Applications of AI-IC**
- Discussion

Applications of ML and HW

- Biomedical informatics
- **Computer vision**
- Customer relationship management
- Data mining
- Email filtering
- Inverted pendulum
- **Natural language processing (NLP)**
  - Automatic ...
  - translation ...
- **Pattern recognition**
  - Facial recognition system

AlexNet
VGG
ResNet

- Handwriting recognition
- Image recognition
- Optical character recognition
- Speech recognition
- Recommendation system
- Search engine
- Social engineering

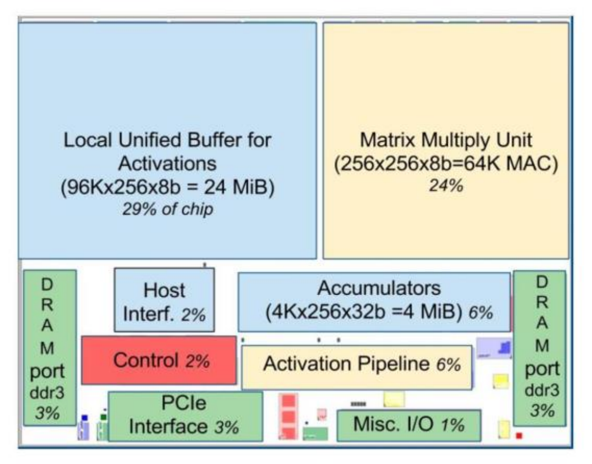


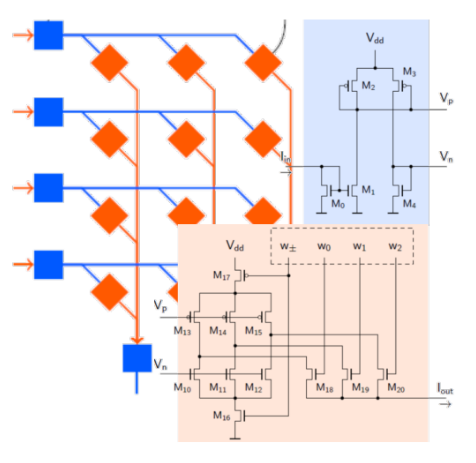
New Hardware for Inference and Training

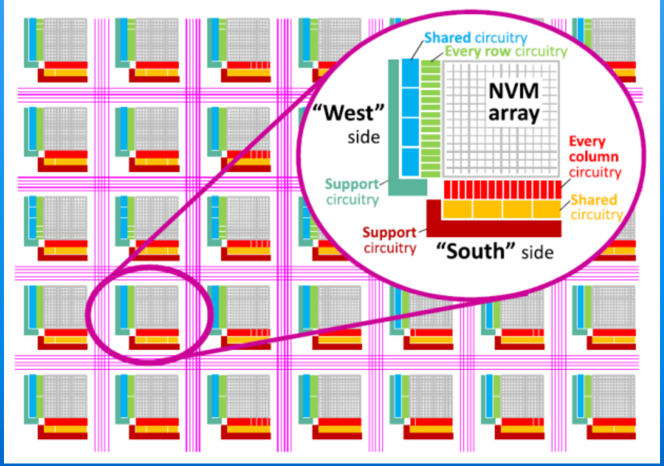
University of Chinese Academy of Sciences

- Digital
  - GPU,TPU,FPGA etc.
  - 10-10^3 speedup
  - Power hungry, large
- Analog
  - Neuromorphic
- Beyond-Si
  - RRAM, STT etc.
  - >10^3 speedup
  - Small footprint, E efficient

Ref. Mohanty et al. IEDM 2017







Apple Chip A Series

Year	2017	2018	2019	2020	2021
Generation	A11	A12	A13	A14	A15?
Process, nm	10	7	N7	5	
Chip area, mm^2	1.83	5.8	1.16		
TOPS	**0.6**	**5**	**6**	**11**	

WSE by Cerebras

	Cerebras WSE	Largest GPU	Cerebras Advantage
Chip size	46,225 mm^2	815 mm^2	56.7 X
Cores	400,000	5,120	78 X
On chip memory	18 Gigabytes	6 Megabytes	3,000 X
Memory bandwidth	9 Petabytes/S	900 Gigabytes/S	10,000 X
Fabric bandwidth	100 Petabits/S	300 Gigabits/S	33,000 X

WSE by Cerebras, CS-1

Cerebras CS-1: A 15 RU System for Training and Inference in the Data Center

cerebras

- Accelerates all deep learning models: CNN, RNN, LSTM, etc.
- Powered by an array of 400,000 Cerebras' AI optimized processor cores
- 18 GB on chip memory —> 3,000 times more than a graphics processing unit
- > 9PB/s on-die memory bandwidth-> 10,000 more than a graphics processing unit
- 100 Pb/s total interconnect bandwidth-->33,000 times more than a graphics processing unit
- System IO: 12 x 100 GbE
- System power: 20 KW;
- Programed with TensorFlow, PyTorch, Mxnet, Caffé2, Theano, CNTK

More Compute than Up To 1,000 GPUs

1/40th the space, 1/50th the power

cerebras

Cerebras WSE-2

- WSE Gen1 16nm, Aug 2019
- WSE Gen2 7nm, Q3 2021
  - 850,000 AI Cores
  - 2,600B transistors, 56 mTr/mm^2
  - Memory 40GB, 20 PB/s
  - Fabric 220 Pb/s

WSE-2 vs Largest GPU

IC	**WSE**	***WSE-2***	**Nvidia A100**
Area	46,255mm^2	*46,255mm^2*	826mm^2
Transistors	1.2 trillion	*2.6 trillion*	54.2 billion
Cores	400,000	*850,000*	6,912 + 432
On-chip memory	18GB	*40GB*	40MB
Memory bandwidth	9PB/s	*20PB/s*	1,555GB/s
Fabric bandwidth	100Pb/s	*220Pb/s*	600GB/s
Fabrication process	16nm	*7nm*	7nm

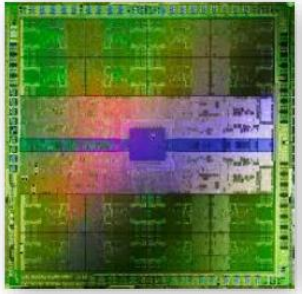
Machine Learning and AI-IC

University of Chinese Academy of Sciences

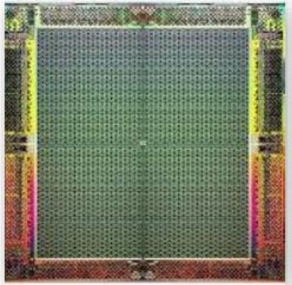
Hardware Acceleration Platforms

中国科学院大学
University of Chinese Academy of Sciences

- CPU, 1X
- GPU, FPGA, ASIC, Beyond CMOS

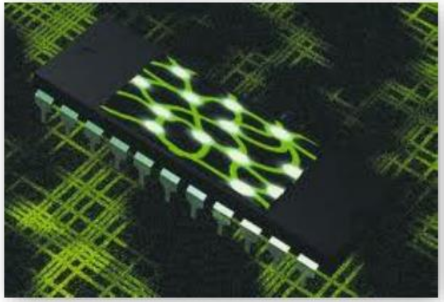


GPU
10 – 30 X



FPGA
10 – 50 X

CMOS ASIC
10^2 – 10^3 X



Beyond CMOS
>10^3 X

Comparison of AI-Chip Features

GPU, FPGA, ASIC, Neuromorphic

Type / Feature	GPU	FPGA	ASIC	Brain-inspired
Customization	General	Semi-custom	Customized	Neuromorphic
Programma-bility	No	Easy	Difficult	No
App scenario	Cloud train.	Acc., D Ctr, Infer.	Widely used	Comp. recog. Envir.
Vendor	Nvidia	Xilinx, Altera	Google, Cambricon	IBM
Advantages	Peak comp., mature	Perf. Power, prog., fast	Av. Perf., power, size	Power, comm., recog
Disadv.	Effic., prog., power	Cost, peak comp.	NRE, R&D cycle, risk	Immature