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

Impact of AI-Big Data on SoC Design

AI Study and AI-Chip Design

Chun-Zhang Chen, Ph.D.

June 25-29, 2018

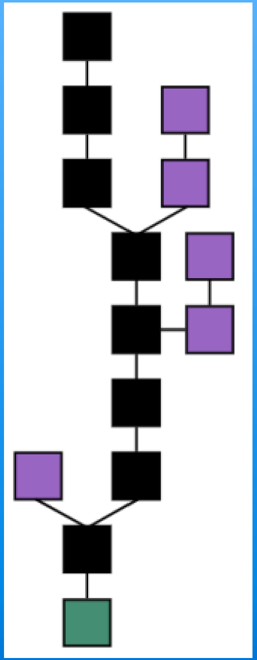
中国科学院大学2018年夏季

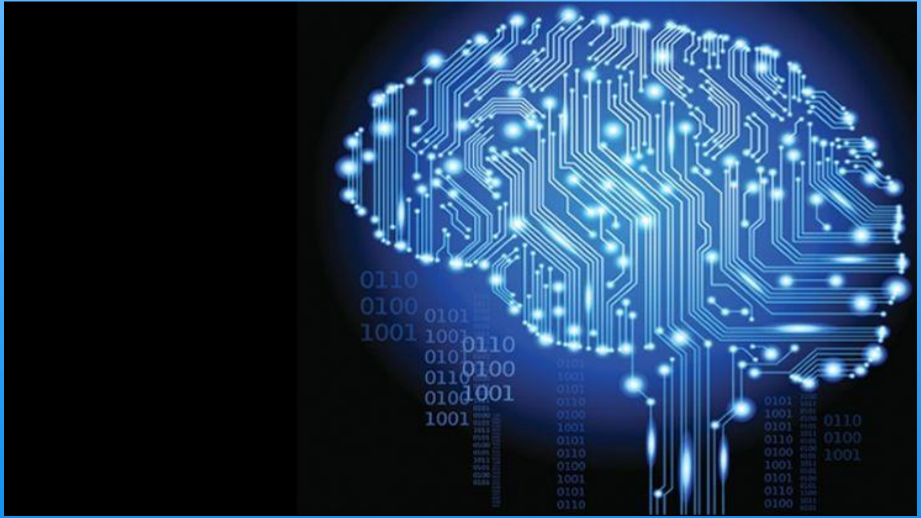
Agenda

- Deep Learning & AI-Chip ▷
- Architecture in AI-Chip ▷
- Designs of AI-Chip ▷
- Energy Efficiency of CPUs ▷
- Discussion ▷

Environment of AI

Big Data, Blockchain, IoT and Cloud Computing





Why deep learning is suddenly changing your life?

- Google,
- Baidu,
- Facebook,
- Microsoft,
- Apple,

Source: Fortune Magazine, 2017

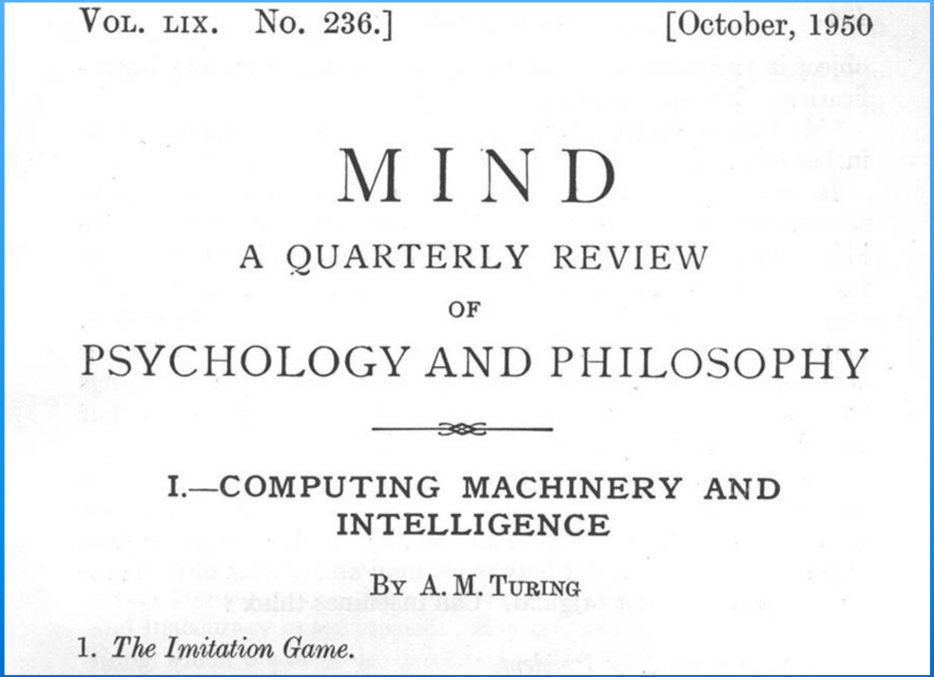
AI Engineering Force (Effects) in USA and China

Type	USA	China	Comparison
Lang./voice	20200	6600	3
xPU/Chips	**17900**	**1300**	**~14**
ML	17600	9800	~2
UAV	9220	4660	~2
Visual/image	4335	1510	~3
Robotics	2100	6400	~0.3

Source: Tencent Institute, 2017

Background of AI

- Alan Turing’s Paper



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)

John McCarthy (Stanford)

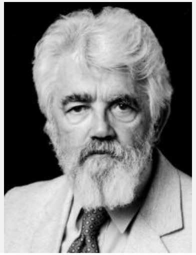
Marvin Minsky (MIT)

Trenchard More (IBM ret'd)

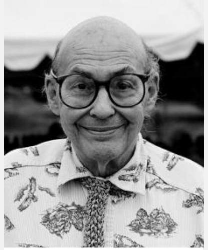
Ray Solomonoff (London)

Oliver Selfridge (MIT)

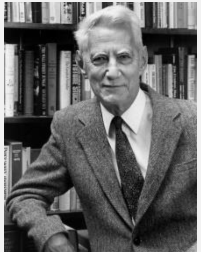
Dartmouth Summer Research Project on Artificial Intelligence 1956



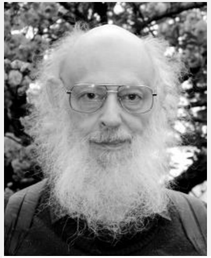
John McCarthy,
"AI" 1955



Marvin Minsky,
MIT AI Lab



Claude Shannon,
MIT Boolean alg.



Ray Solomonoff,
Inductive Inteference

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

“The First Wave of AI (1956-76)”

AI at Universities Too Optimistically

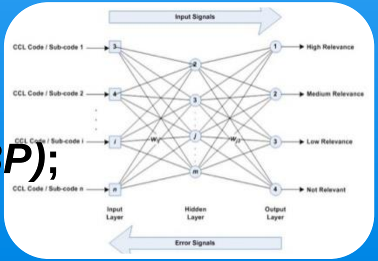
The Golden Years of AI (1956-74);
H. Simon & A Newell 1975 Turing

The 1st Winter (1974-80)
The 2nd Winter (1987-93)

PC Market ; IBM-Deep Blue
1997 & Jeopardy 2011

“Winter Seasons of AI (1976-06)”

ML/DNN Algorithm, Backpropagation *(BP)*;



Next Step: *BP? Capsule？*

Definition/Classification of AI

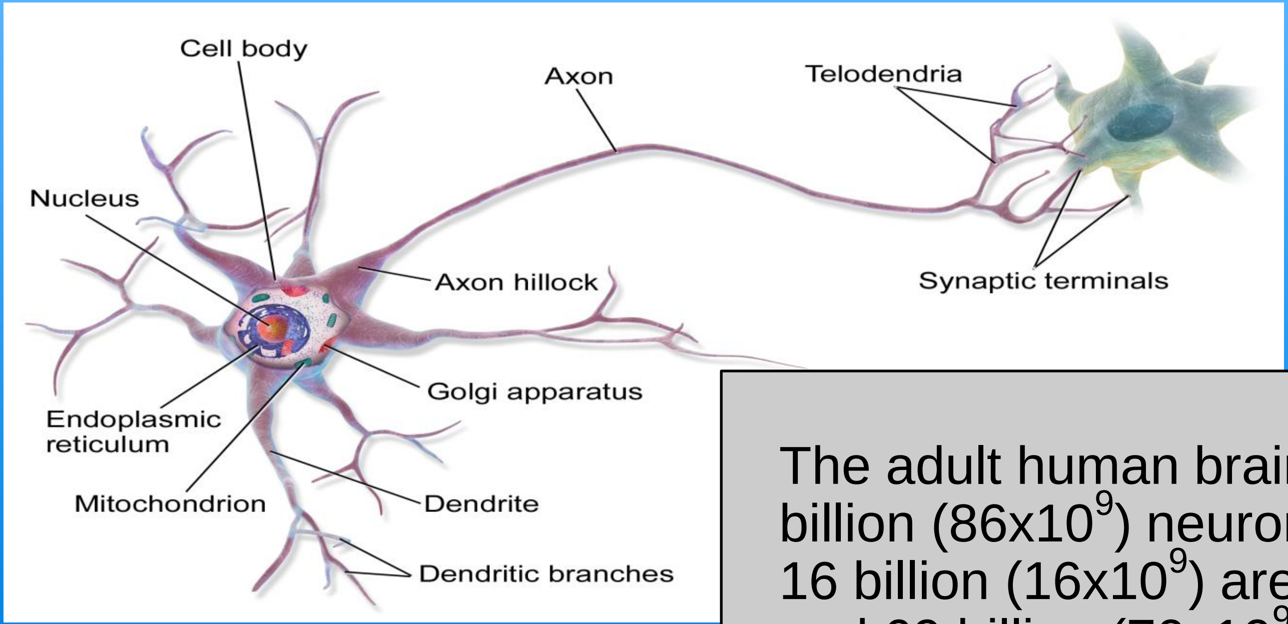
- What is AI? Merriam-Webster Dictionary:
  - "An area of computer science that deals with giving machines the ability to seem like they have human intelligence"
  - "The power of a machine to copy intelligent human behavior"
- How AI is classified?
  - Artificial Weak/Narrow Intelligence (ANI)
    -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

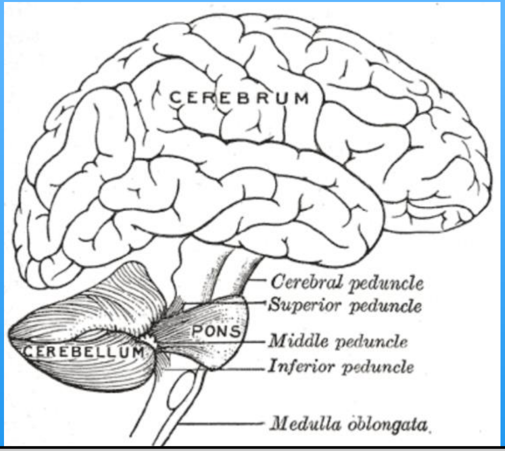
Schools of Machine Learning

- **Symbolism** *(Frank Rosenblatt, 1957)*
  - Bayes nets, Judea Pearl, ACM Turing Award 2011
  - Knowledge Graph, Google
- **Connectionism (neuron study)**
  - Marvin Minsky, ACM Turing Award 1969
  - Geoffrey Hinton (CNN), Backpropogation (1986, Nature)
- **Actionism**
  - AlphaGo (DeepMind), AlphaGo Zero (2017) and AlphaZero

Neurons and Synapses (Neural Network)

- Human brain: Cerebrum, Cerebellum





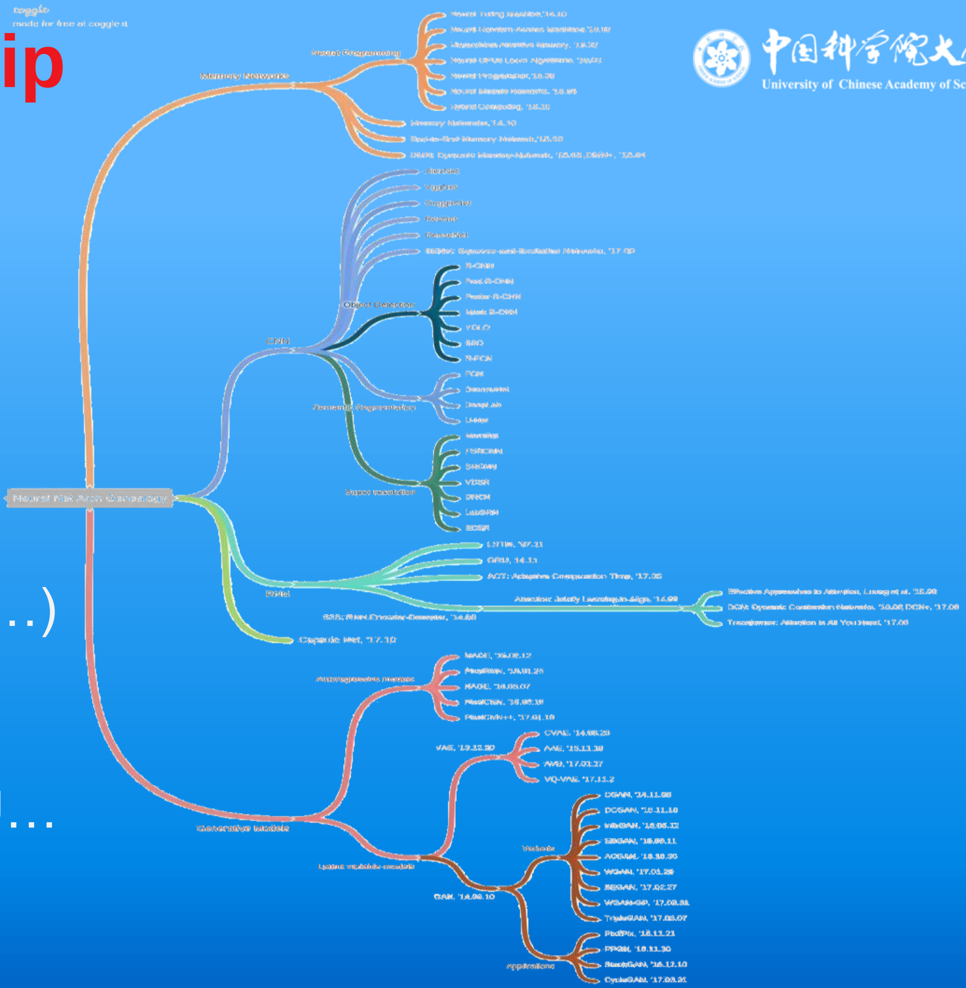
The adult human brain contains about 85-86 billion ($86x10^9$) neurons,[38][39] of which 16 billion ($16x10^9$) are in the cerebral cortex and 69 billion ($70x10^9$) in the cerebellum.[39]

Background of CNN/DNN

- **AI Terms:** CNN, ANN and **DNN**
  - AI, 1956, **John McCarthy**
  - ANN (Artificial Neural Network)
  - CNN (*Cellular*[1]/Cognitive[2]/**Convoluted[3]** Neural Network)
  - **DNN (Deep Learning**, Deep CNN or DCNN**)**
- **Key People/Team**
  - Marvin Minsky (8/9/27-1/24/16), CNN[2], Co-Founder, MIT AI Lab
  - **Geoffrey Hinton**, CNN[2], Deep Belief Networks, U. Toronto/Google
    - **Back-Propgate 80s, 2006 DL, 12/11/13 Science**
  - **Yann LeCun**, **CNN[3]** (1989), AI Facebook/NYU prof.
  - Yoshua Bengio, ANN/DNN (2006), experiments for on DBN, UdM
  - IBM, CNN[2] (Cognitive Computer, 2012)

AI – From Object to Chip

- Data
  - Big Data, massive,
- Algorithm
  - Fast evolving
  - xNN (CNN,RNN,DNN,SNN…)
- HW,
  - GPU,FPGA,DSP,ASIC,TPU…



AI Teaching at CMU and Others

- AI Boot Camp at CMU (1956) and Starting in the Fall:
  - Launched (05/10/18) Undergraduate Degree in AI at SCS
- HIT: Institute of AI Research (May 5, 2018)
- NJU: AI College (March 2018) and Under Program:
  - ML and Data Mining; AI System & Application
- BUAA: AI for Under Program (Sept 2017)

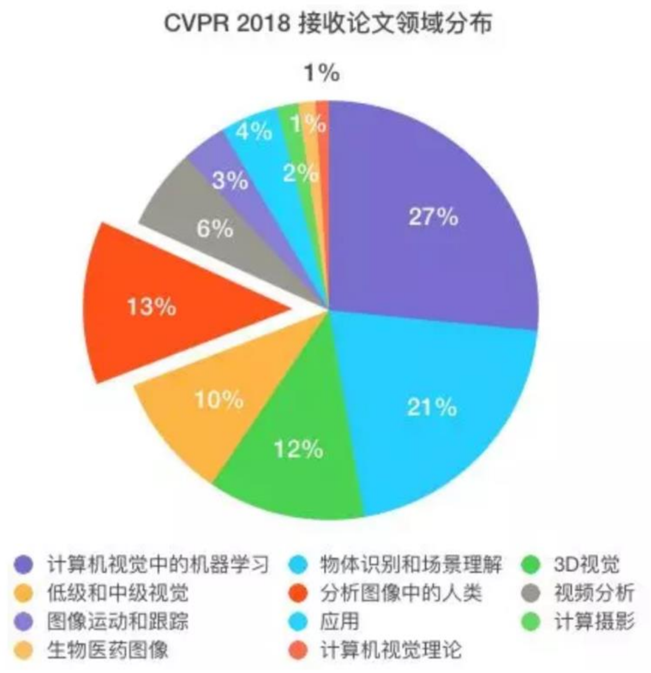
CVPR 2018 (IEEE, 1985-)

2018 CVPR
SALT LAKE CITY • JUNE 18-22

- computer vision and pattern recognition
- Geoffrey Hinton, DL is limited
- DL/CNN → GAN
  - Generative Adversarial Networks

We need to start over …

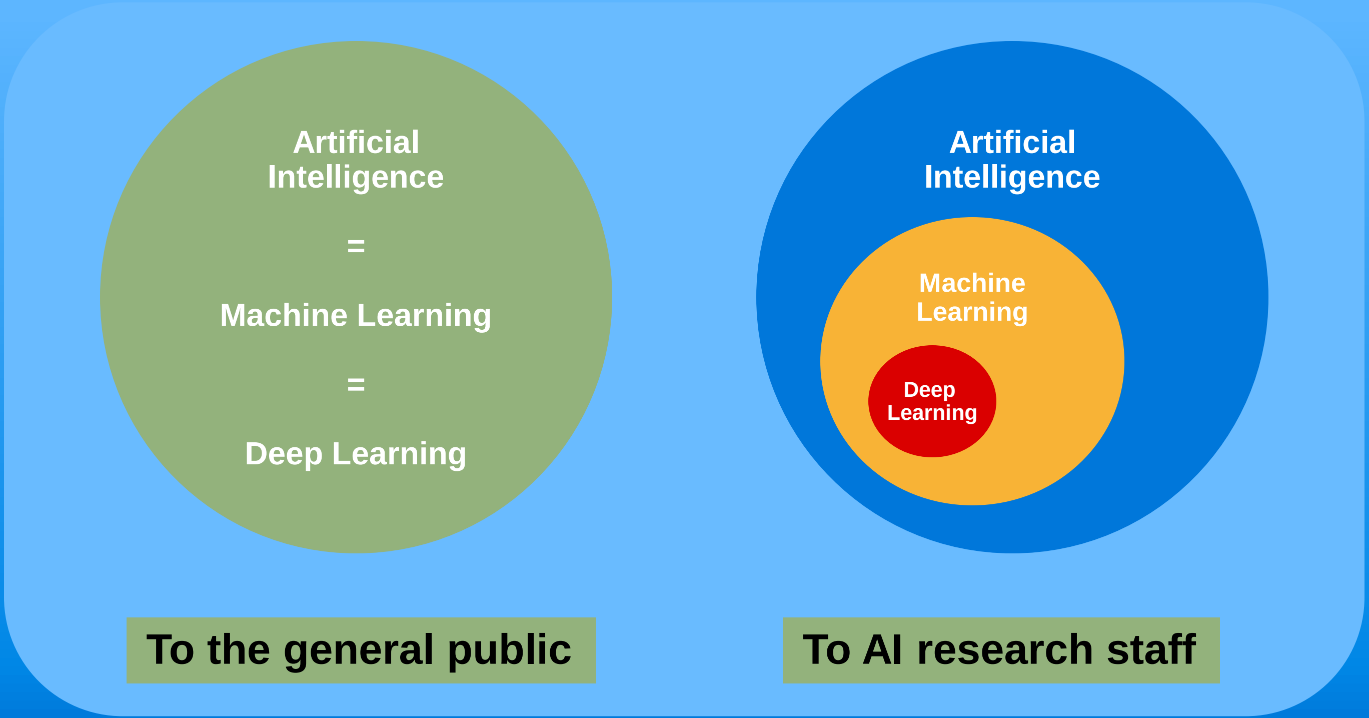
What is wrong with convolutional neural nets?
Fields Institute, 2017 | Geoffrey Hinton, U of Toronto



CVPR 2018 接收论文领域分布

Views of AI/ML/DL

Can AI Replace Human Beings?

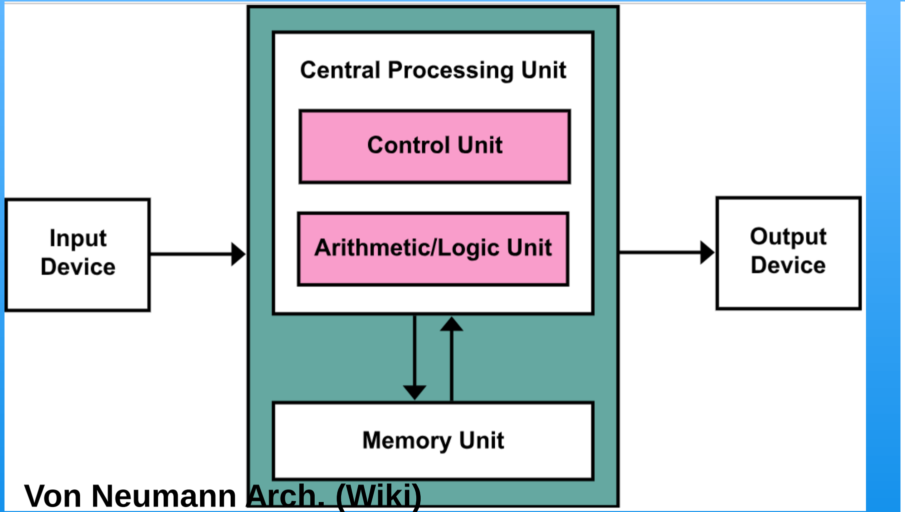


Brain-alike? Or Brain-inspired?

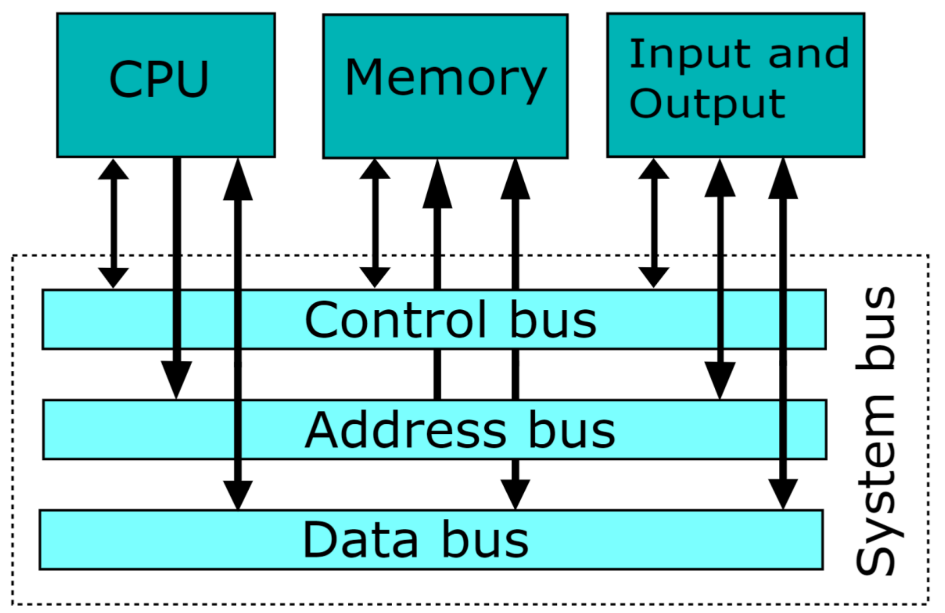
AI-Chip and SoC Design

- **Deep Learning & AI-Chip:** XNN & Algorithms
- **Architecture in AI-Chip:** Von Neumann, GPU & HSA
- **Designs of AI-Chip:** TrueNorth, TPU, Cambricon
- **Energy Efficiency of CPUs:** Area/Power/Computation Efficiency
- **Discussion:** w.r.t BD-Cloud/IoT/ADAS

Von Neumann Architecture



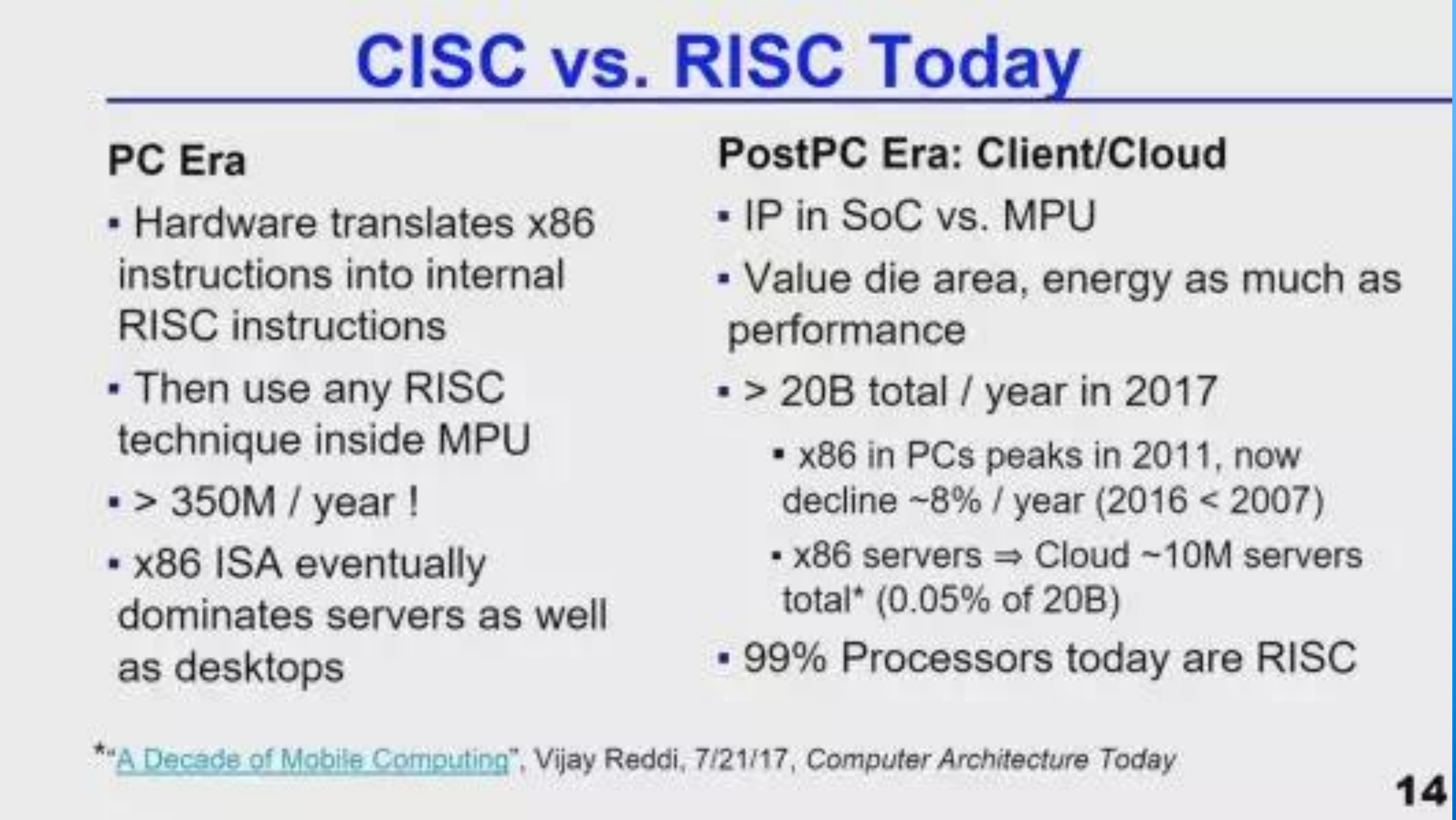
Von Neumann Arch. (Wiki)



Single system bus evolution of the architecture

Original Von Neumann Arch
has been developed to
Single System Bus Arch

CISC vs RISC



CISC vs. RISC Today

PC Era

- Hardware translates x86 instructions into internal RISC instructions
- Then use any RISC technique inside MPU
- > 350M / year !
- x86 ISA eventually dominates servers as well as desktops

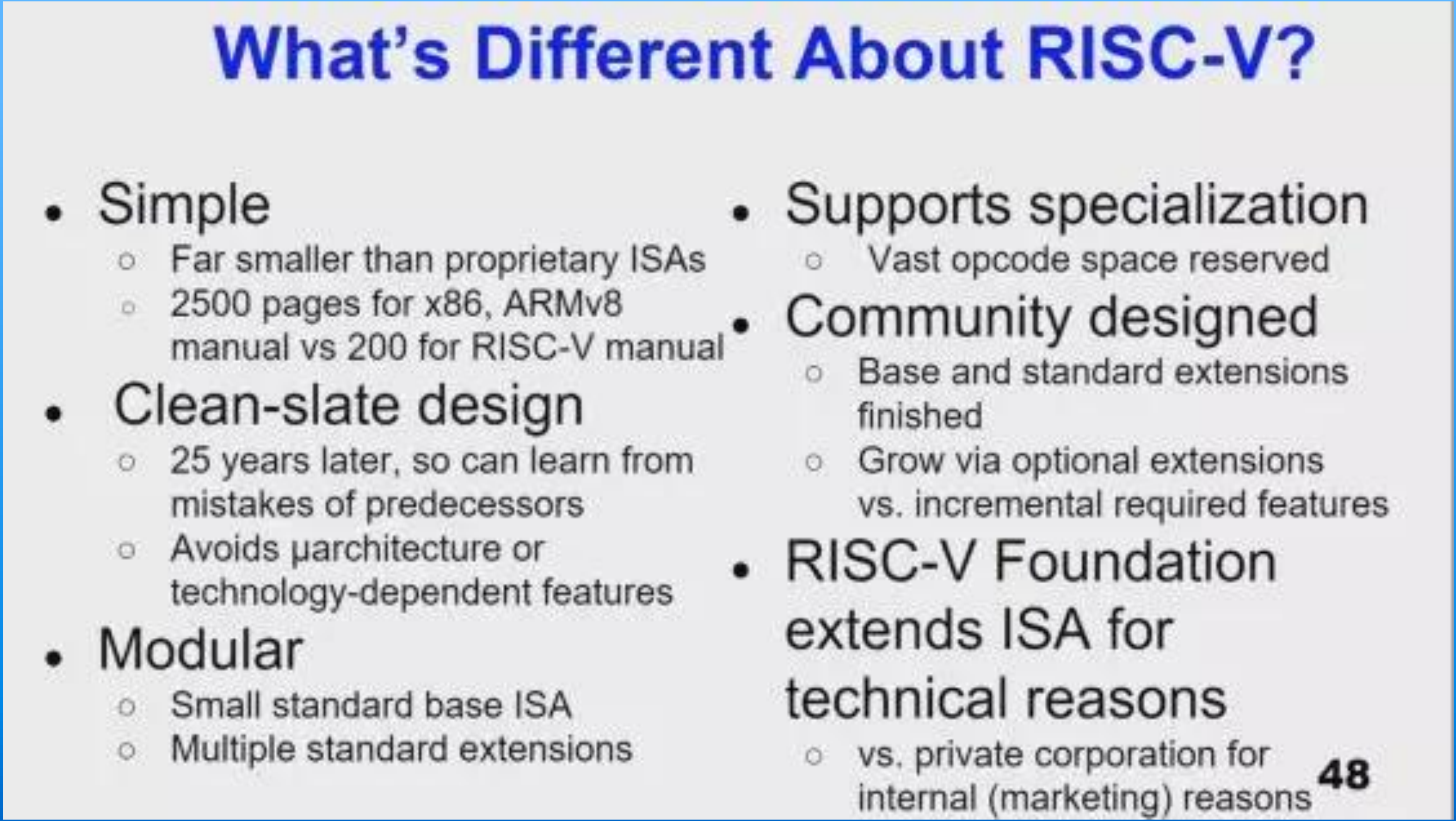
PostPC Era: Client/Cloud

- IP in SoC vs. MPU
- Value die area, energy as much as performance
- > 20B total / year in 2017
  - x86 in PCs peaks in 2011, now decline ~8% / year (2016 < 2007)
  - x86 servers ⇒ Cloud ~10M servers total* (0.05% of 20B)
- 99% Processors today are RISC

*“A Decade of Mobile Computing”, Vijay Reddi, 7/21/17, *Computer Architecture Today*

14

About RISC-V



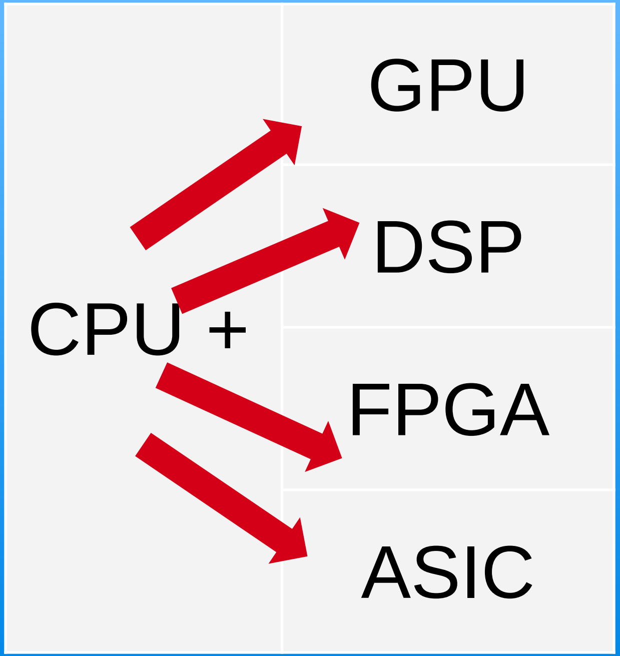
What's Different About RISC-V?

- Simple
  - Far smaller than proprietary ISAs
  - 2500 pages for x86, ARMv8 manual vs 200 for RISC-V manual
- Clean-slate design
  - 25 years later, so can learn from mistakes of predecessors
  - Avoids μarchitecture or technology-dependent features
- Modular
  - Small standard base ISA
  - Multiple standard extensions
- Supports specialization
  - Vast opcode space reserved
- Community designed
  - Base and standard extensions finished
  - Grow via optional extensions vs. incremental required features
- RISC-V Foundation extends ISA for technical reasons
  - vs. private corporation for internal (marketing) reasons

48

HSA for AI-Chip Architecture

- CPU + GPU
  - GPU by Nvidia (GeForce), AMD, Intel, ARM etc
  - SW: CUDA (Nvidia), OpenVX (Intel), OpenCL
- CPU+ DSP
  - DSP from Cadence; SW/OS
- CPU+FPGA
  - eFPGA, reconfig FPGA, FPGA/ASIC
- CPU+ASIC
  - Customized ASIC



AI-Chip and SoC Design

- **Deep Learning & AI-Chip:** XNN & Algorithms
- **Architecture in AI-Chip:** Von Neumann, GPU & HSA
- **Designs of AI-Chip:** TrueNorth, TPU, Cambricon
- **Energy Efficiency of CPUs:** Area/Power/Computation Efficiency
- **Discussion:** w.r.t BD-Cloud/IoT/ADAS

Top 10 AI-Chip Companies

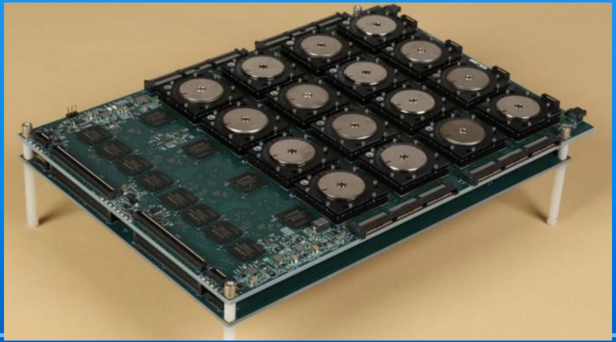
- Nvidia, Tesla P100 GPU
- ARM, Blue Sky Program
- Intel, Nervana
- IBM, syNAPSE
- Google , TPU
- ViMicro, NPU
- MS, Catapult
- KnuEdge, LambaFablic
- Horizon Robotics, Neuromorphic
- Krtkl, 430K LUT

AI Chips in AI Era

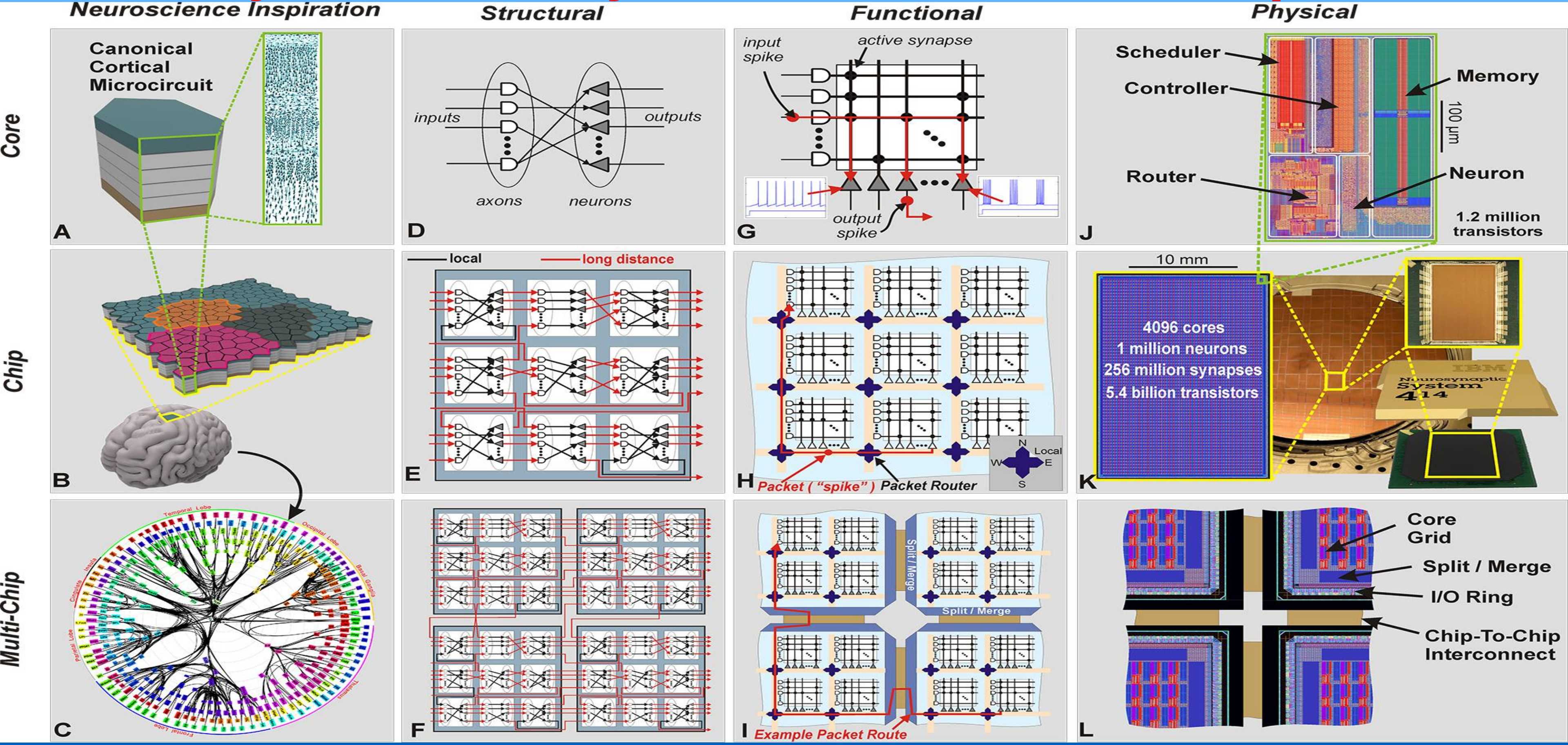
- Architectures are based on …(von Neumann?)
  - GPU, FPGA, ASIC
- Based on DL/CNN module…
  - Ex.
- Cloud and AI chips
- Applications
  - Automotive, Robotics, SmartHomes

Neuromorphic AI Chip by IBM

- **A neuromorphic CMOS IC, TrueNorth chip in 2014**
  - Many cores, 4096 cores, simulating a total >10^6 neurons
  - The programmable synapses is >268×10^6 (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

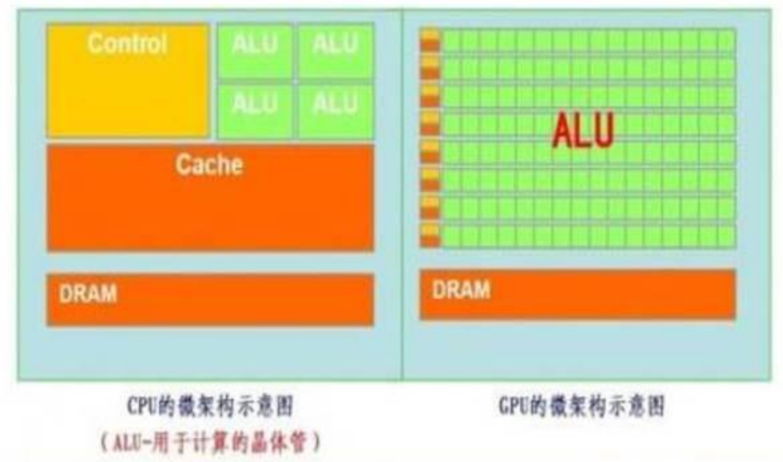


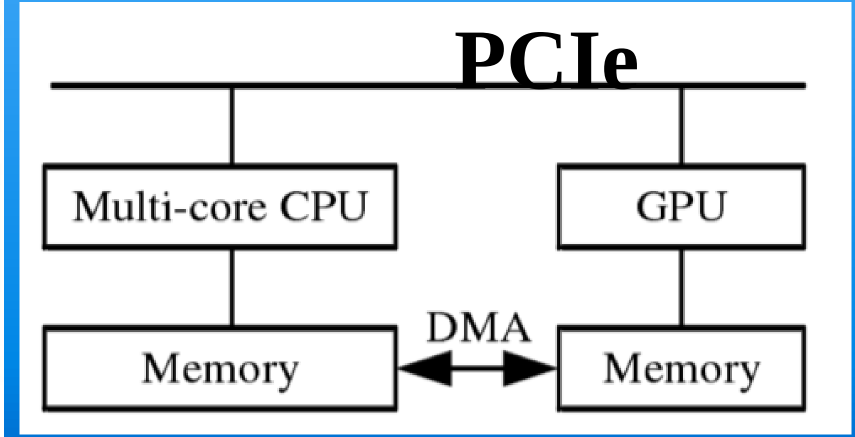
IBM SyNAPSE Project and TrueNorth Chip



AI Chip AlphaGo by Google

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

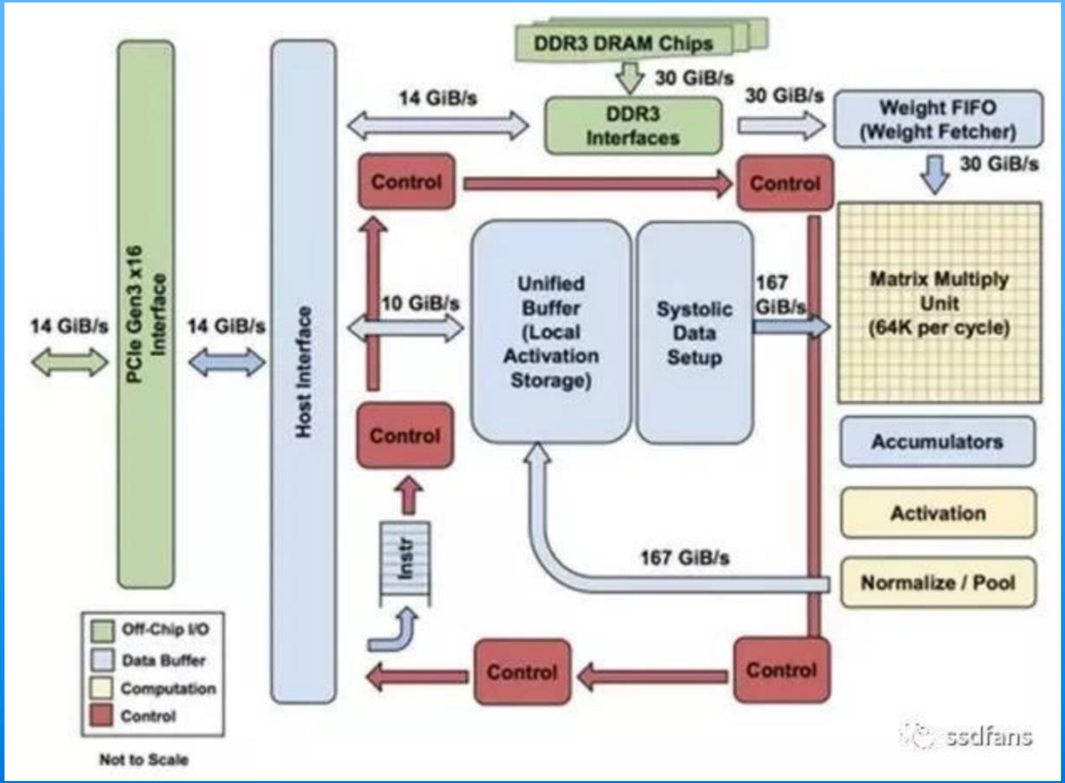


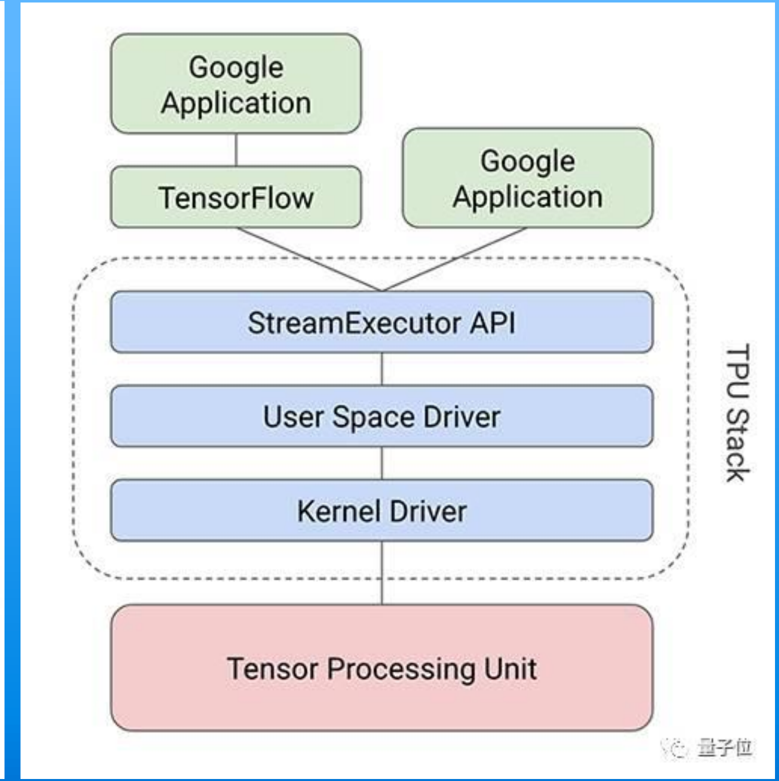


TPU, TensorFlow and Google I/O

- TPU3.0 (5/8-5/10/18, Mtn View)
- TensorFlow 1.0, https://www.tensorflow.org/
  - 09/27/16 → 11/06/16 → 02/15/17
  - Google, DeepMind, DropBox, Qualcomm, mi
- TensorFlow, ASIC (CPU+GPU)

TPU Architecture and TensorFlow

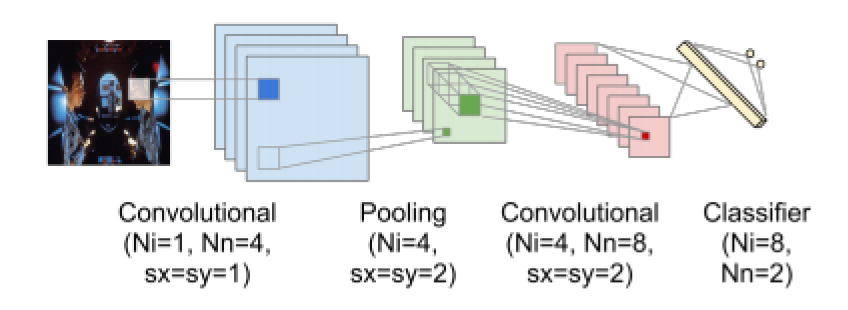




Cambricon (2016-) Chips

Cambricon
寒武纪科技

- 2016, Cambricon-1A chip – ***The First DL Chip***
  - $16x10^9$ virtual neurons/s, peak at $2x10^{12}$ synapses/s, Huawei 970 Karin chip in Mate 10
- 2018, 3 processor cores (2TOPS/4TOPS/8TOPS), w/ 5 TOPS/W
- 5/10/2018, Cambricon MLUv01 (Smart Cloud Processor Card), TSMC 16nm

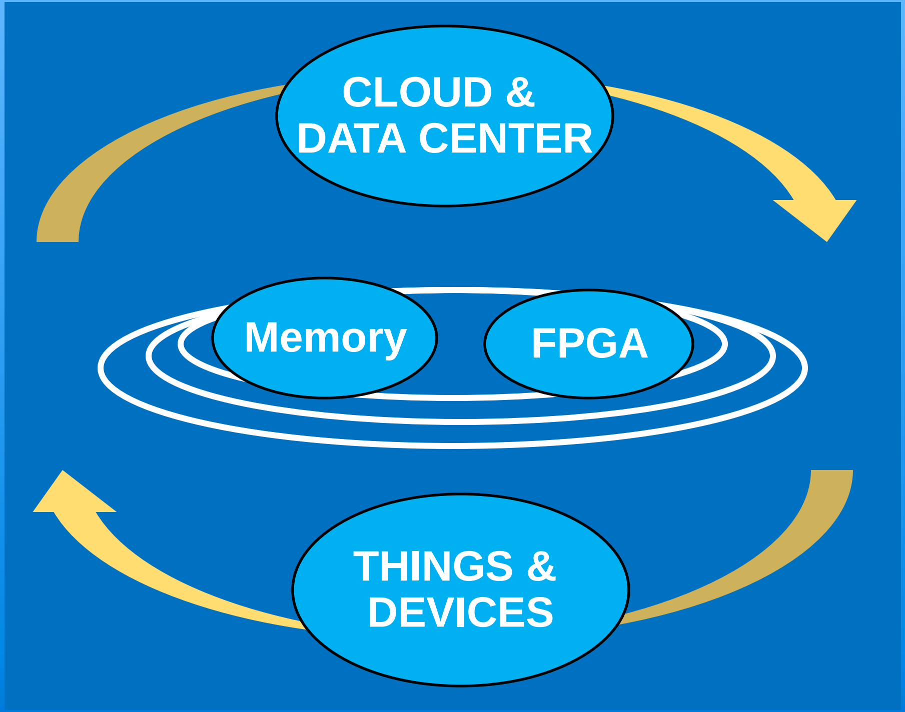

Convolutional (Ni=1, Nn=4, sx=sy=1)
Pooling (Ni=4, sx=sy=2)
Convolutional (Ni=4, Nn=8, sx=sy=2)
Classifier (Ni=8, Nn=2)

Cambricon AI-Chips

- 2016, Cambricon-1A chip
  - 16×10^9 virtual neurons/s, peak at 2×10^{12} synapses/s, Huawei 970 Karin chip in Mate 10
- 2018, 3 processor cores (2TOPS/4TOPS/8TOPS), w/ 5 TOPS/W
- 5/10/2018, Cambricon MLUv01 (Smart Cloud Processor Card), TSMC 16nm

AI Strategy at Intel

- IoT, Cloud, 2014
  - Storage and FPGA
- Nervana, ML, 2016
- Loihi chip, self-learning
  - 128 neurons + 3 x86 CPU
- Spring Crest, 2018



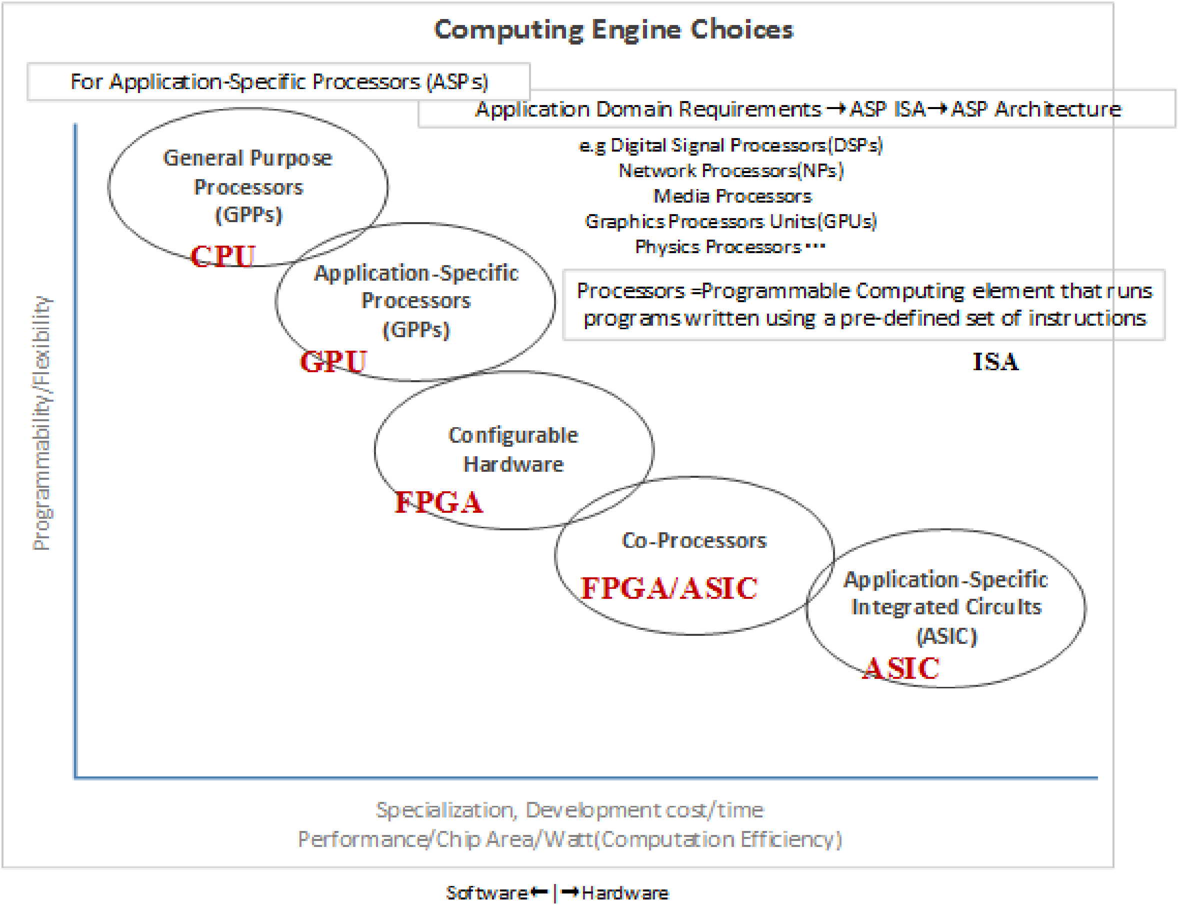
Qualcomm and IC

- 5G, transforming the way we interact w/ our world & each other.
- AR/VR
- SmartVideo
- Smart Home
- Drone
- Robotics

AI-Chip and SoC Design

- **Deep Learning & AI-Chip:** XNN & Algorithms
- **Architecture in AI-Chip:** Von Neumann, GPU & HSA
- **Designs of AI-Chip:** TrueNorth, TPU, Cambricon
- **Energy Efficiency of CPUs:** Area/Power/Computation Efficiency
- **Discussion:** w.r.t BD-Cloud/IoT/ADAS

Performance of Computing Engines



Computing Engine Choices

For Application-Specific Processors (ASPs)

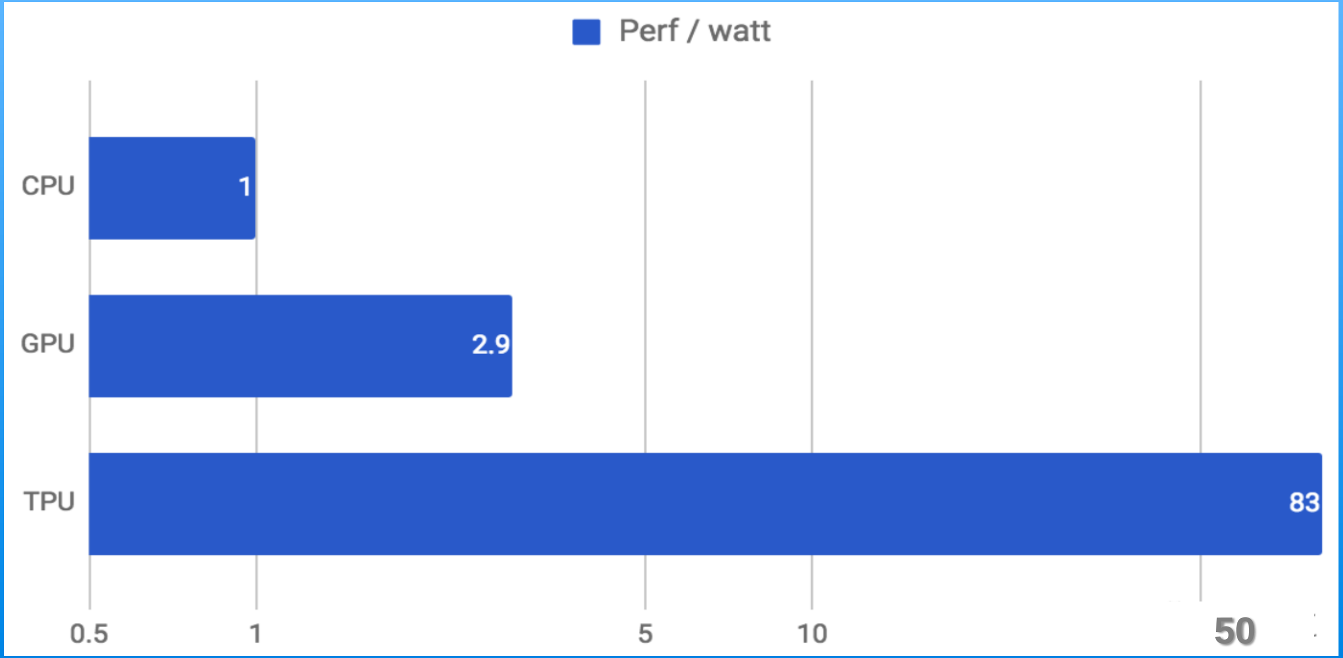
Application Domain Requirements → ASP ISA → ASP Architecture

e.g Digital Signal Processors(DSPs)
Network Processors(NPs)
Media Processors
Graphics Processors Units(GPUs)
Physics Processors ···

Processors =Programmable Computing element that runs programs written using a pre-defined set of instructions

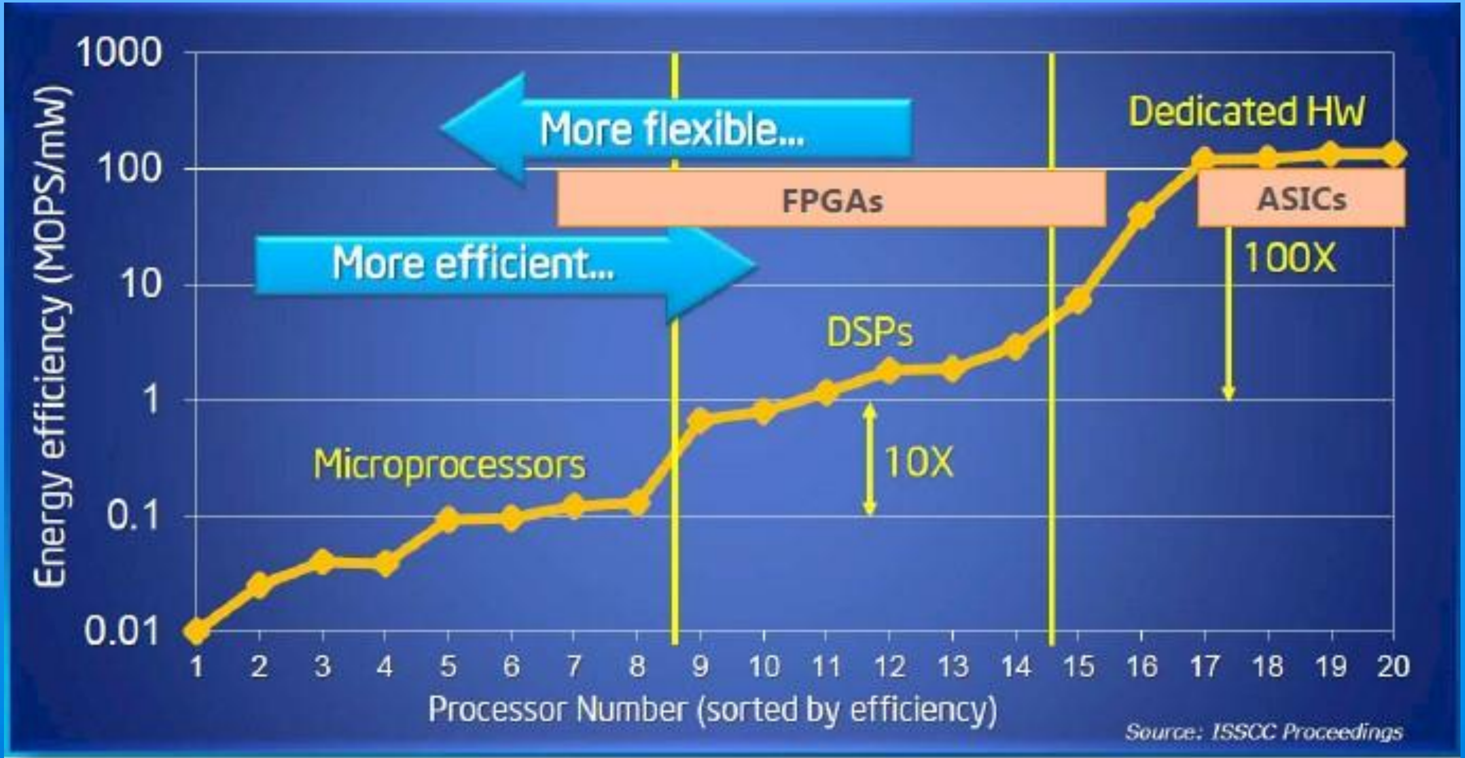
ISA
Programmability/Flexibility
General Purpose Processors (GPPs)
CPU
Application-Specific Processors (GPPs)
GPU
Configurable Hardware
FPGA
Co-Processors
FPGA/ASIC
Application-Specific Integrated Circuits (ASIC)
ASIC
Specialization, Development cost/time
Performance/Chip Area/Watt(Computation Efficiency)
Software ← | → Hardware

Energy Efficiency of CPUs



Energy Efficiency Comparison

CPU vs FPGA and ASIC



Source: Bob Broderson, Berkeley Wireless group

AI-Chip and SoC Design

- **Deep Learning & AI-Chip:** XNN & Algorithms
- **Architecture in AI-Chip:** Von Neumann, GPU & HSA
- **Designs of AI-Chip:** TrueNorth, TPU, Cambricon
- **Energy Efficiency of CPUs:** Area/Power/Computation Efficiency
- **Discussion:** w.r.t BD-Cloud/IoT/ADAS

AI-Era Chip Design and HW-SW Co-Design

- HPC Chip
- HBM Chip
- MCU Chip
- ECU Chip

- CPU Architecture: DSA
- SW: DSL
- RISC-V
- Security
- SW-HW Co-Design

Do we need to start over on a new ML?

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

- Background: ML schools are Symb., Conn., Action.

- DL: CNN/DNN → ML: GAN (CVPR 2018)?
  - Ian Goodfellow (PhD of Yoshua Benjio)
- What is the next step of TrueNorth Chip?
- How does Cambricon-MLU work with AI-ISA?