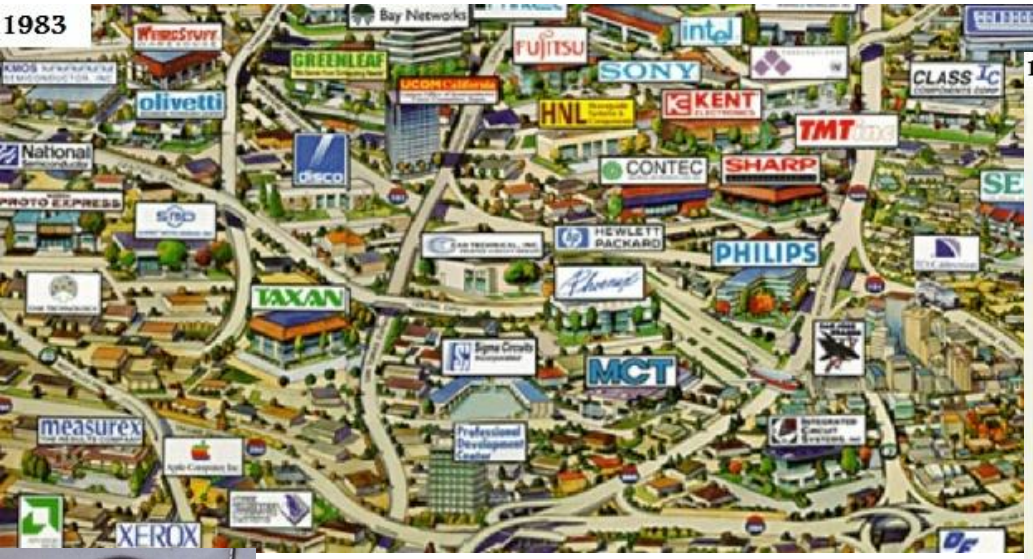
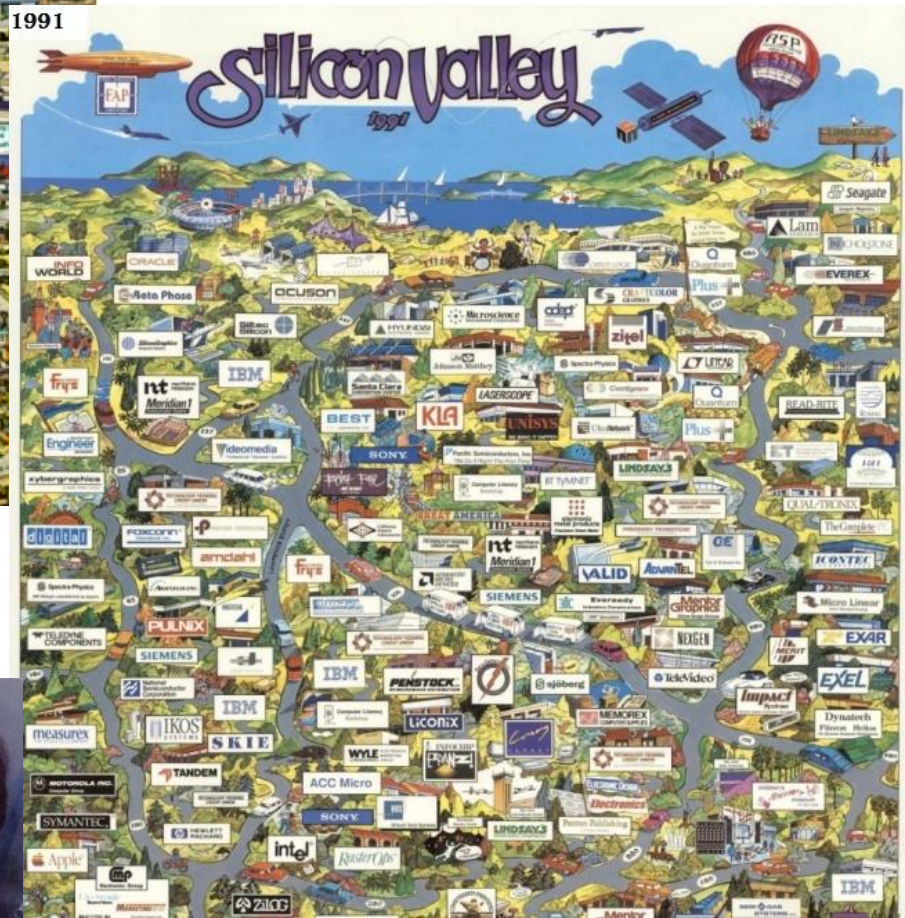
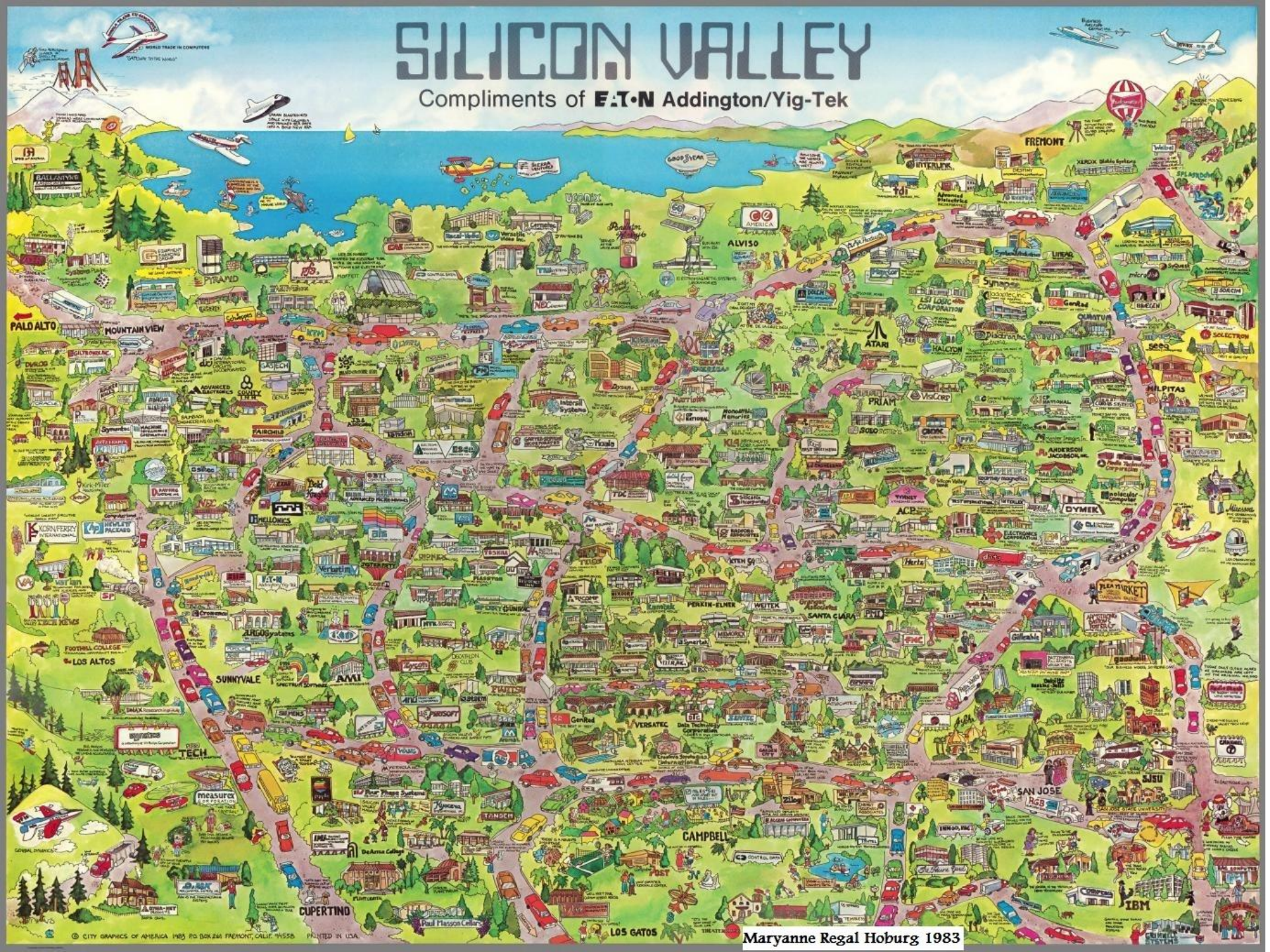
Hardware 2.0

Silicon Valley used to be about hardware…





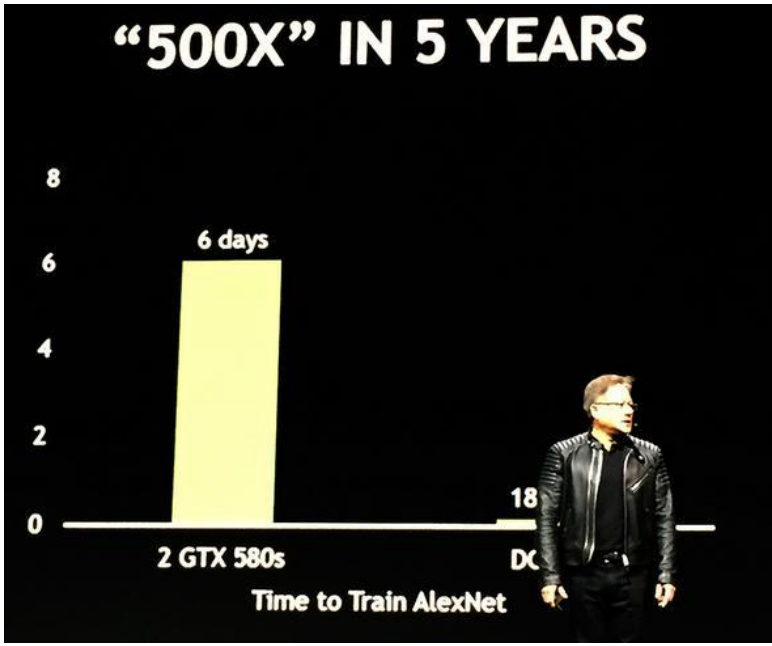

SILICON VALLEY
Compliments of EATON Addington/Yig-Tek
FREMONT
ALVISO
PALO ALTO
MOUNTAIN VIEW
ATARI
MILPITAS
SUNNYVALE
SANTA CLARA
CAMPBELL
SAN JOSE
CUPERTINO
LOS GATOS
LOS ALTOS
FOOTHILL COLLEGE
TECH
IBM
SJSU
© CITY GRAPHICS OF AMERICA 1983 PO BOX 261 FREMONT, CALIF. 94538 PRINTED IN USA.
Maryanne Regal Hoburg 1983

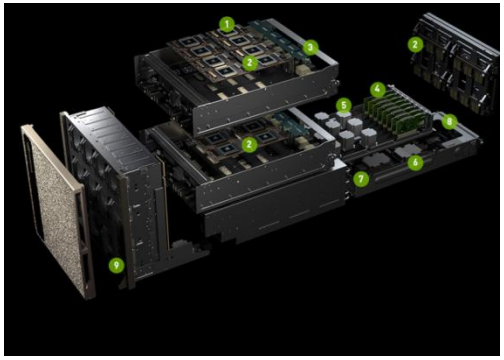
Huang's Law

"Nvidia's GPUs today are 25 times faster than 5 years ago" (Nvidia CEO Jensen Huang , April 2018)

2012: It took six days on two Nvidia GTX 580s to train AlexNet

2018: It takes 18 minutes with the Nvidia DGX-2





IEEE SPECTRUM 2 Apr 2018

Move Over, Moore's Law: Make Way for Huang's Law

Graphics processors are on a supercharged development path that eclipses Moore's Law, says Nvidia's Jensen Huang

Hardware 2.0

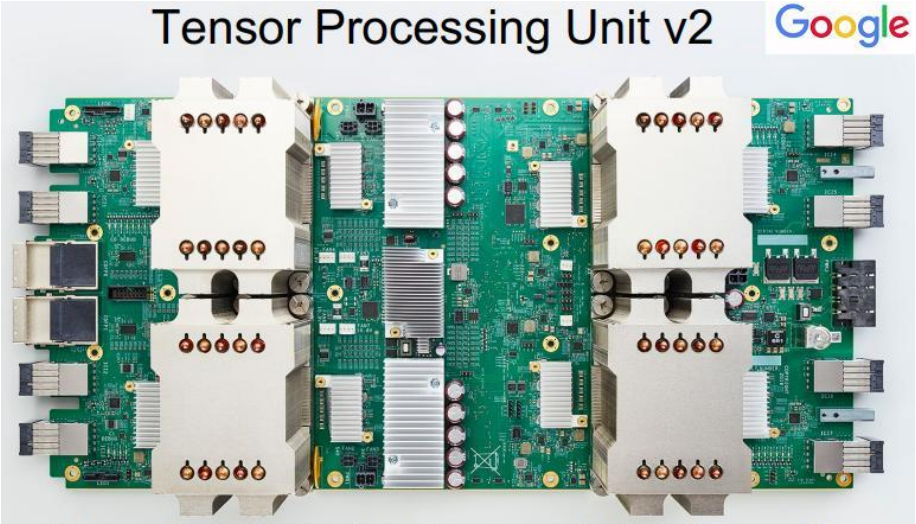
Google's TPU (programmed via TensorFlow and offered via Google Cloud)

Internal search ranking model training:

~9 hours on 1/4 pod vs. ~132 hours on 275 high end CPU machines (14.2 times faster)

Internal image model training:

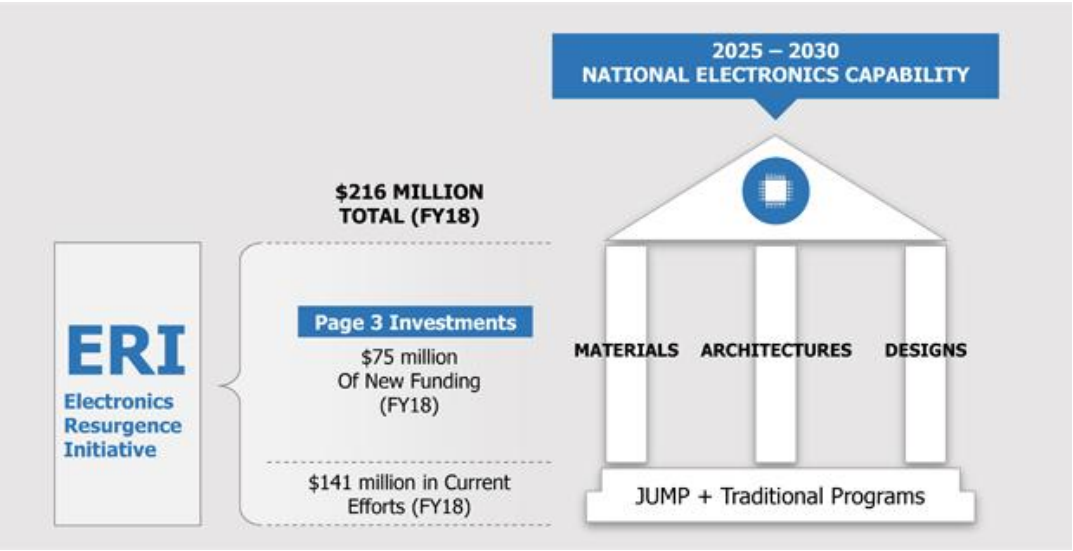
~22 hours on 1/4 pod vs. ~216 hours on previous production setup (9 times faster)



Hardware 2.0

DARPA Electronics Resurgence Initiative (ERI): $1.5 billion, 5-year program

Extend Moore's Law by using new materials such as carbon nanotubes



DARPA Announces First Annual Electronics Resurgence Initiative Summit

U.S. electronics community to convene at inaugural event around five year, $1.5B effort to create transformative advances in electronics

ERI ELECTRONICS RESURGENCE INITIATIVE
SUMMIT
2018 | SAN FRANCISCO, CA | JULY 23-25

Hardware 2.0

Neuromorphic chips

- SpiNNaker @ Manchester
- BrainScaleS @ Heidelberg

Human Brain Project

SEPT. 24, 2019

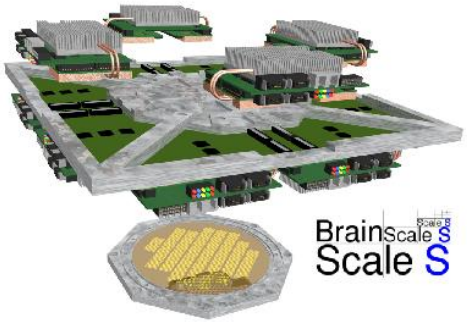
Second Generation SpiNNaker Neuromorphic Supercomputer to be Built at TU Dresden

Human Brain Project

Silicon Brains

PROJECTS STORY | 19 March 2018

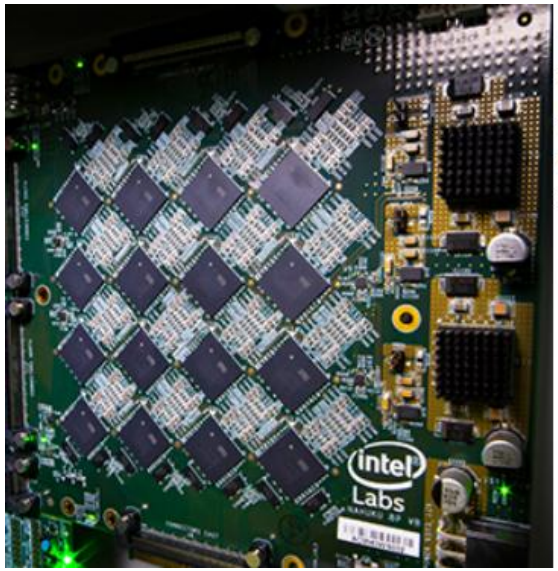
Two new neuromorphic chips developed in the Human Brain Project



Hardware 2.0

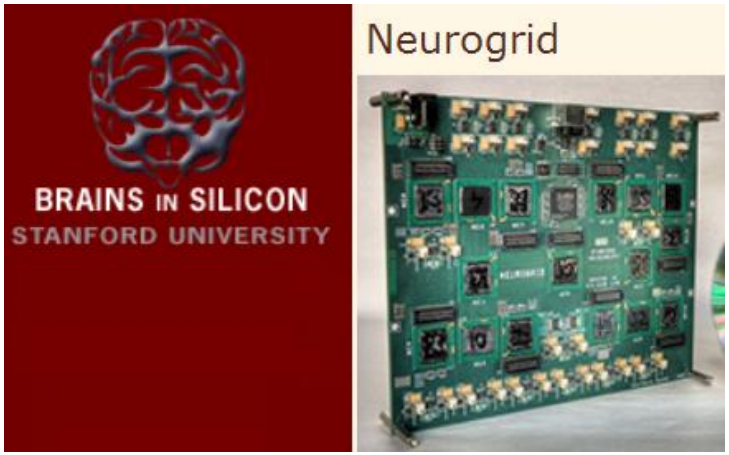
Neuromorphic chips

- Intel Loihi
- IBM TrueNorth
- Stanford Neurogrid



July 15, 2019

Pohoiki Beach, is made up of multiple Nahuku boards and contains 64 Loihi chips.



Hardware 2.0

Neuromorphic chips

Intel Loihi

Intel Creates Neuromorphic Research Community to Advance 'Loihi' Test Chip

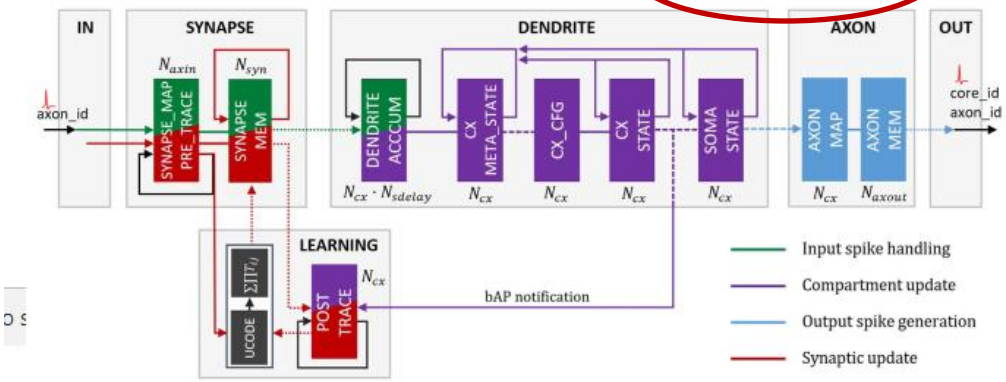
March 1, 2018

Loihi: A Neuromorphic Manycore Processor with On-Chip Learning

Mike Davies
Intel Labs, Intel Corporation

Narayan Srinivasa
Eta Compute

IEEE Micro
Volume: 38, Issue: 1,
January/February 2018



Hardware 2.0

Neuromorphic chips

IBM True North (2014)

Tayfun Gokmen and Yuri Vlasov (IBM, 2016): Resistive Processing Unit (analog units)

Introducing a Brain-inspired Computer

TrueNorth's neurons to revolutionize system architectur

Dharmendra Modha

By Dharmendra S. Modha

New IBM patent on neuro-chip of Sep 2018

(19) **United States**

(12) **Patent Application Publication**

Amir et al.

(10) **Pub. No.: US 2018/0260682 A1**

(43) **Pub. Date: Sep. 13, 2018**

(54) **GRAPH PARTITIONING AND PLACEMENT FOR MULTI-CHIP NEUROSYNAPTIC NETWORKS**

(71) Applicant: **INTERNATIONAL BUSINESS MACHINES CORPORATION**, Armonk, NY (US)

(72) Inventors: **Arnon Amir**, San Jose, CA (US); **Pallab Datta**, San Jose, CA (US); **Myron D. Flickner**, San Jose, CA (US); **Dharmendra S. Modha**, San Jose, CA (US); **Tapan K. Nayak**, San Jose, CA (US)

(21) Appl. No.: **15/457,658**

Publication Classification

(51) **Int. Cl.**
G06N 3/04 (2006.01)

(52) **U.S. Cl.**
CPC ***G06N 3/04*** (2013.01)

(57) **ABSTRACT**

Graph partitioning and placement for multi-chip neurosynaptic networks. According to various embodiments, a neural network description is read. The neural network description describes a plurality of neurons. The plurality of neurons has a mapping from an input domain of the neural network. The plurality of neurons is labeled based on the mapping from the input domain. The plurality of neurons is grouped into a plurality of groups according to the labeling. Each of the plurality of groups is continuous within the input domain. Each of the plurality of groups is assigned to at least one neurosynaptic core.

Hardware 2.0

- BRAIN Initiative - Iarpa's Microns: Allen Institute; Baylor College of Medicine; Harvard University; Princeton University (David Markovitz)

Office of the Director of National Intelligence
IARPA
BE THE FUTURE

Machine Intelligence from Cortical Networks (MICrONS)

NYU

Hardware 2.0

- Neuromorphic processors

ICONS Tennessee, July 23-25, 2019.
International Conference on Neuromorphic Systems

BICT Bio-inspired Information and Communications Technologies

BICT 2020 - 12th EAI International Conference on Bio-inspired Information and Communications Technologies

July 7-8, 2020
Shanghai, People's Republic of China

OAK RIDGE National Laboratory

Catherine Schuman

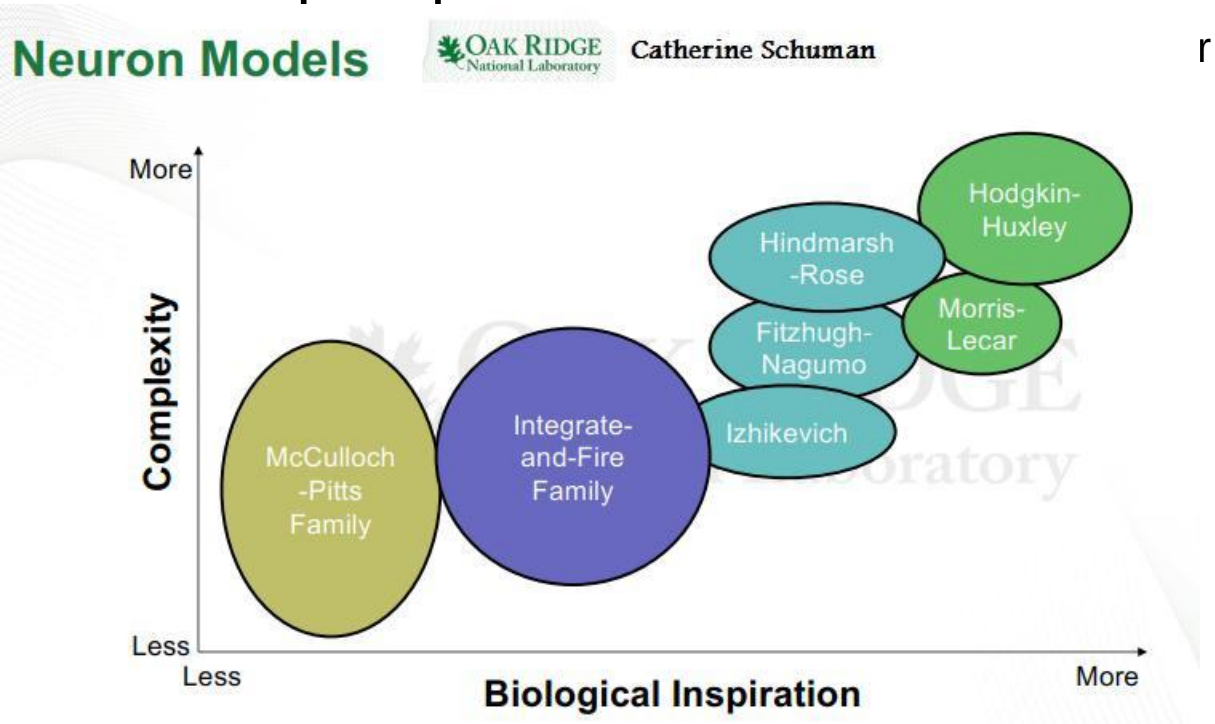
May 2017

A Survey of Neuromorphic Computing and Neural Networks in Hardware

Catherine D. Schuman, *Member, IEEE,* Thomas E. Potok, *Member, IEEE,* Robert M. Patton, *Member, IEEE,* J. Douglas Birdwell, *Fellow, IEEE,* Mark E. Dean, *Fellow, IEEE,* Garrett S. Rose, *Member, IEEE,* and James S. Plank, *Member, IEEE*

Hardware 2.0

- Neuromorphic processors



Hardware 2.0

- Open-source Hardware

Open-Source Processors: Bringing Silicon Back to Silicon Valley?

Tuesday, September 18, 2018
6:00 pm
Stanford Faculty Club, Stanford University, 439 Lagunita Drive, Stanford, CA 94305

Sponsored by:
US-Asia Technology Management center

Stanford University

Hardware 2.0

- Boom of semiconductor design:
  - unrelenting demand for new memory chips
  - digitization of entire industry sectors such as automotive and industrial
- Chip design spending in 2018: $412 billion

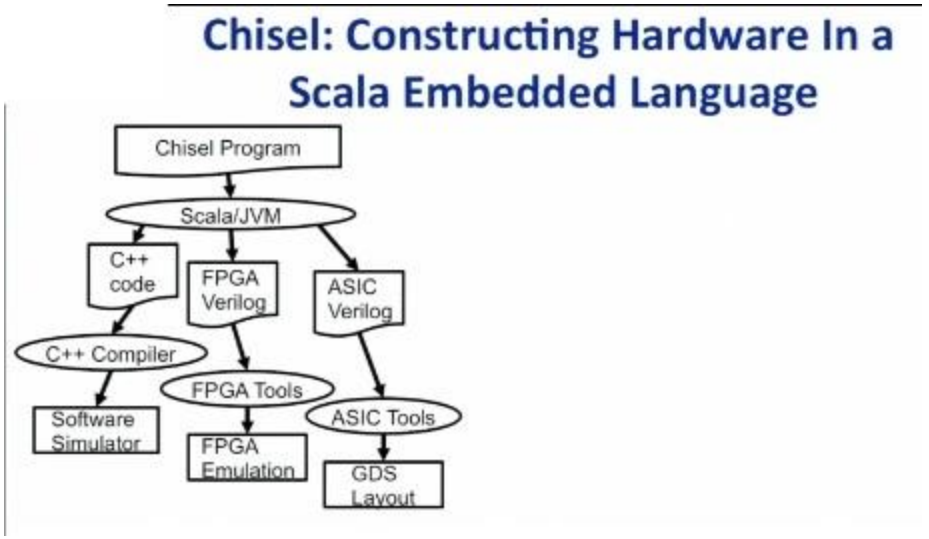
Hardware 2.0

- Open-source Hardware
  - RISC-V Foundation: open-source RISC-V instruction set
  - RISC-V architecture for Internet of Things devices

Hardware 2.0

- Open-source Hardware
  - RISC-V

Krste Asanovic

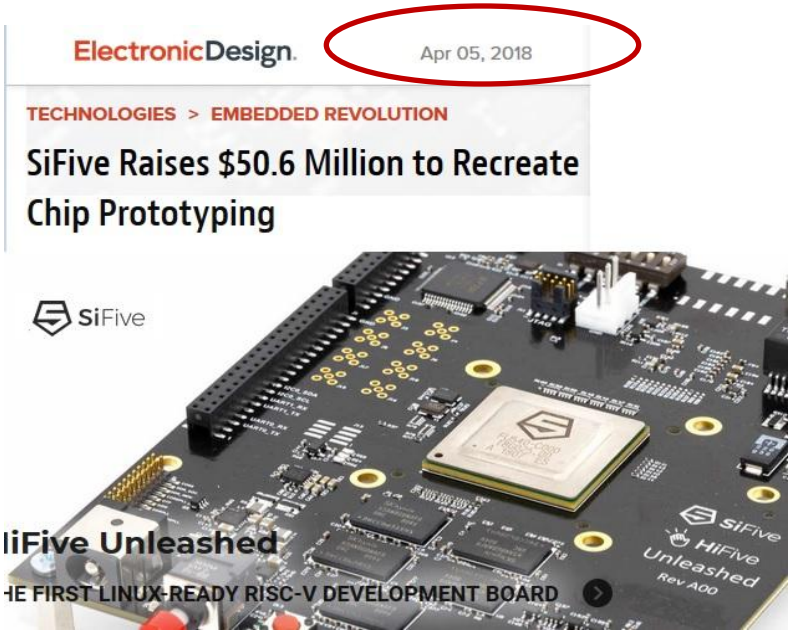


Hardware 2.0

- Open-source Hardware
  - Members of the RISC-V Foundation include Google, Microsoft, IBM and Oracle and Tesla

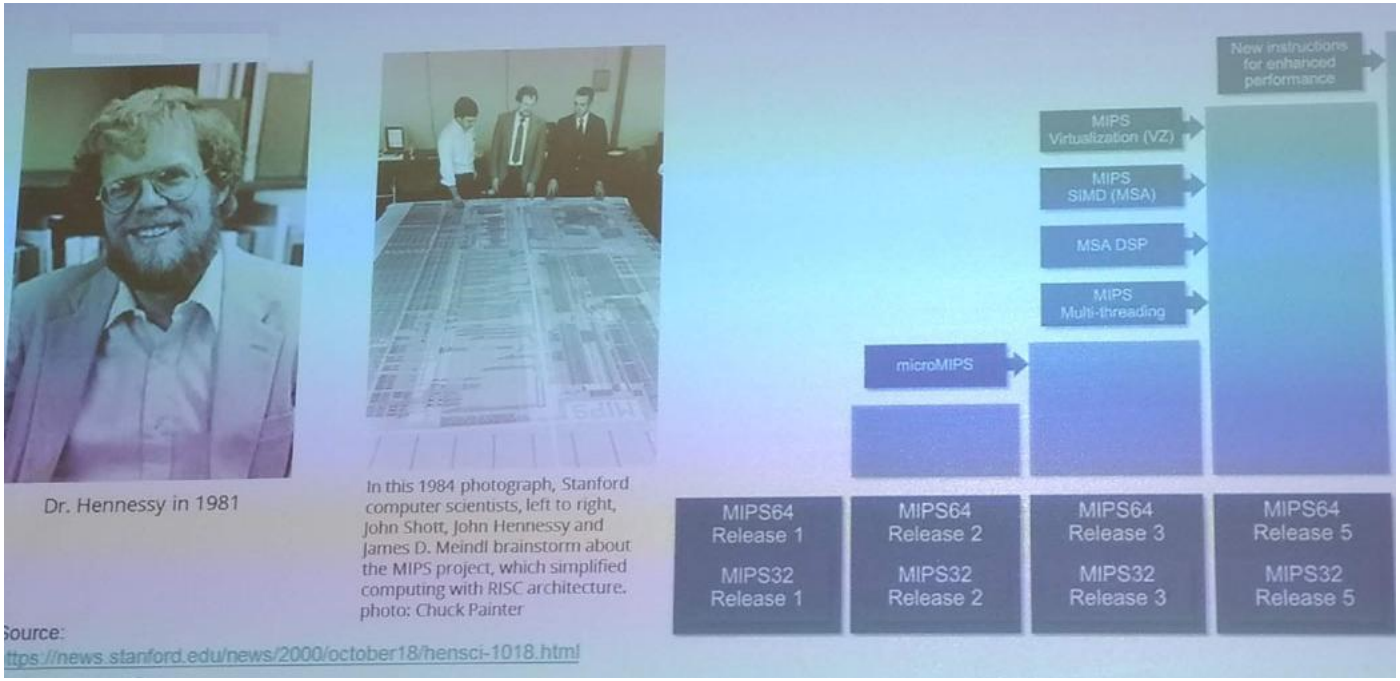
Hardware 2.0

- Open-source Hardware
  - Chip-design factories in the cloud (TSMC's Cloud Alliance, 2018)
  - SiFive (Krste Asanovic, Yunsup Lee and Andrew Waterman, the researchers that created the RISC-V architecture)
  - Esperanto (founded by Dave Ditzel)



Hardware 2.0

- Open-source Hardware
  - MIPS:

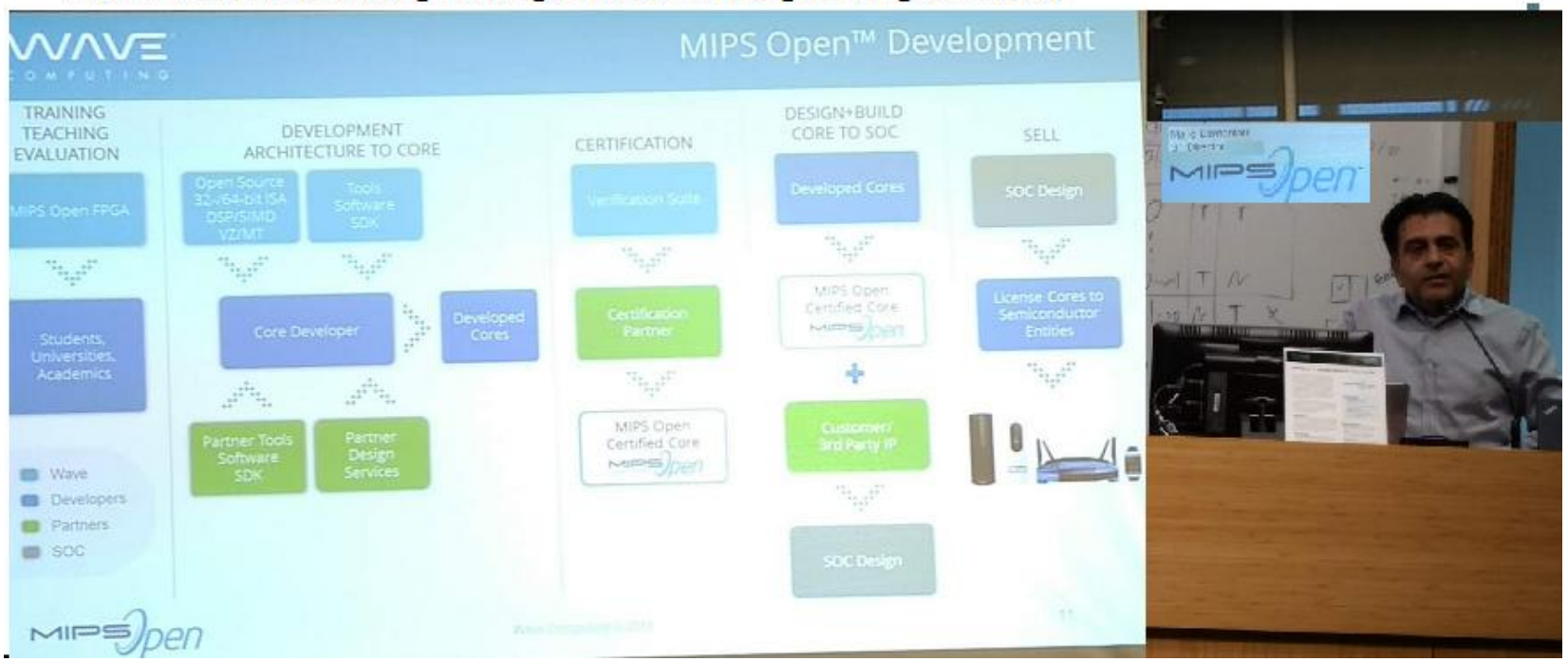


Hardware 2.0

- Open-source Hardware
  - MIPS: Wave MIPS Open (May 2019)

Stanford EE Computer Systems Colloquium

4:30 PM, Wednesday, May 1, 2019
Shriram Center for Bioengineering and Chemical Engineering Room 104



Hardware 2.0

- Traditional memory/storage hierarchy: SRAM is integrated into the processor for cache, DRAM, used for main memory, is separate and located in a dual in-line memory module (DIMM), disk drives and NAND-based solid-state storage drives (SSDs) are used for storage
- Not optimal for AI
- An alternative: Memory-centric chip technologies, in-memory and near-memory computing

Hardware 2.0

- Memory-centric chip technologies: in-memory and near-memory computing
- Peter Kogge (IBM, 1994): EXECUBE, the first Processing-in-Memory chip
- Christos Kozyrakis & Mingyu Gao (Stanford, 2015): Near-Data Processing
- Mythic (Redwood City, CA)
- Syntiant (Irvine, CA)
- Untether AI (Toronto, Canada)

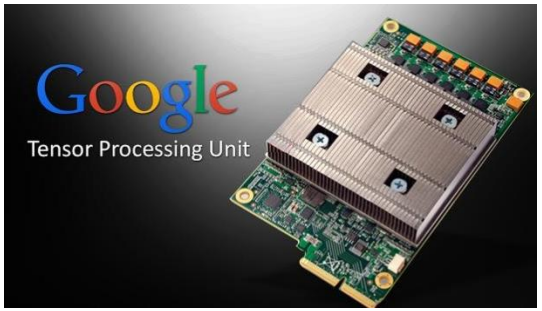
MYTHIC

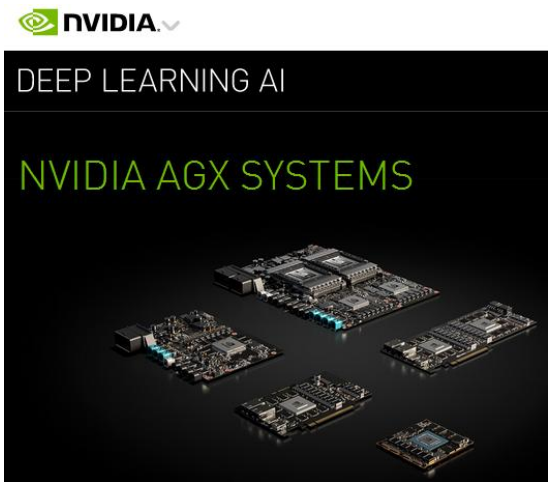
SYNTIANT

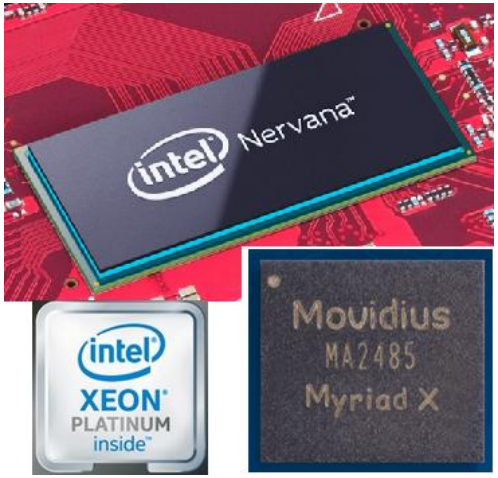
UNTETHER AI

Hardware 2.0

- A.I. Hardware
  - Google: TensorFlow
  - Nvidia: AGX etc
  - Intel: Nervana Neural Network Processor for Inference NNP-I (2019), Movidius Myriad Vision Processing Units (2018), acquired Nervana, Movidius, Untether (2016-19)







Hardware 2.0

- A.I. Hardware
  - Qualcomm (2020)
  - ARM: Trillium (2019)

Project Trillium, Arm's Machine Learning (ML) platform, enables a new era of advanced, ultra-efficient inference at the edge.

REUTERS

TECHNOLOGY NEWS

APRIL 9, 2019 / 6:25 PM

Qualcomm aims to take on Nvidia, Intel with new AI chips

Hardware 2.0

- A.I. Hardware
  - Wave Computing (Campbell, 2010)
  - Cerebras Systems (Los Altos, 2016)
  - Cambricon (China, 2016)
  - SambaNova (Palo Alto, 2017)
  - Groq (Palo Alto, 2017)
  - Untether (Toronto, 2017)

WAVE COMPUTING

cerebras

SambaNova SYSTEMS

crunchbase news

Groq, A Stealthy Startup Founded By Google's TPU Team, Is Raising $60M

September 6, 2018

UNTETHER AI

Hardware 2.0

- A.I. Hardware
  - Graphcore (Britain, 2016)

Hardware 2.0

- ASIC
  - Light (Palo Alto, 2013)

Hardware 2.0

- Jim Keller (Intel): "Moore's Law is not Dead" (UC Berkeley, 2019)
- https://www.youtube.com/watch?v=oIG9ztQw2Gc

Hardware 2.0

- New materials
  - Max Shulaker (MIT, 2019) builds a microprocessor from carbon nanotube transistors
    - 16-bit microprocessor with more than 14,000 CNFETs
    - RISC-V compliant

Hardware 2.0

- Intel
  - Kaby Lake-G processor: CPU + GPU + High Bandwidth Memory (HBM) + EMIB (Embedded Multi-die Interconnect Bridge)
  - Foveros (2019): 3D die stacking
  - Trailing behind TSMC: 45nm process in 2007, 32nm in 2009, 22nm in 2011, 14nm in 2015, 10nm in 2018
  - Extreme ultraviolet lithography (EUVL) not announced
  - Intel acquired Nervana

Hardware 2.0

- Global Foundries:
  - Skips 7nm logic technology

Hardware 2.0

- TSMC
  - CLN7FF (2017): 7nm Fin Field-Effect Transistor (FinFET) process technology that enables chip designers to shrink die sizes by 70% at the same transistor count and to lower power consumption by 60%
  - Built with deep ultraviolet (DUV) lithography
  - More than 50 chip designs of the 7nm process technology underway in 2018 (Apple, AMD, Huawei, and Qualcomm
  - Announced EUVL for 2019

Hardware 2.0

- The cost-per-transistor is going down, but the cost of design with finFETs is in the $100-million range.
- Small companies can’t afford to design with FinFETs
- GlobalFoundries and Samsung offer FD-SOI based platforms that use planar transistors and reduce design cost (Fully Depleted Silicon On Insulator)
- SOI more suitable for IoT chips

Hardware 2.0

- Silicon Catalyst, the first incubator for semiconductor start-ups

ALUMNUS

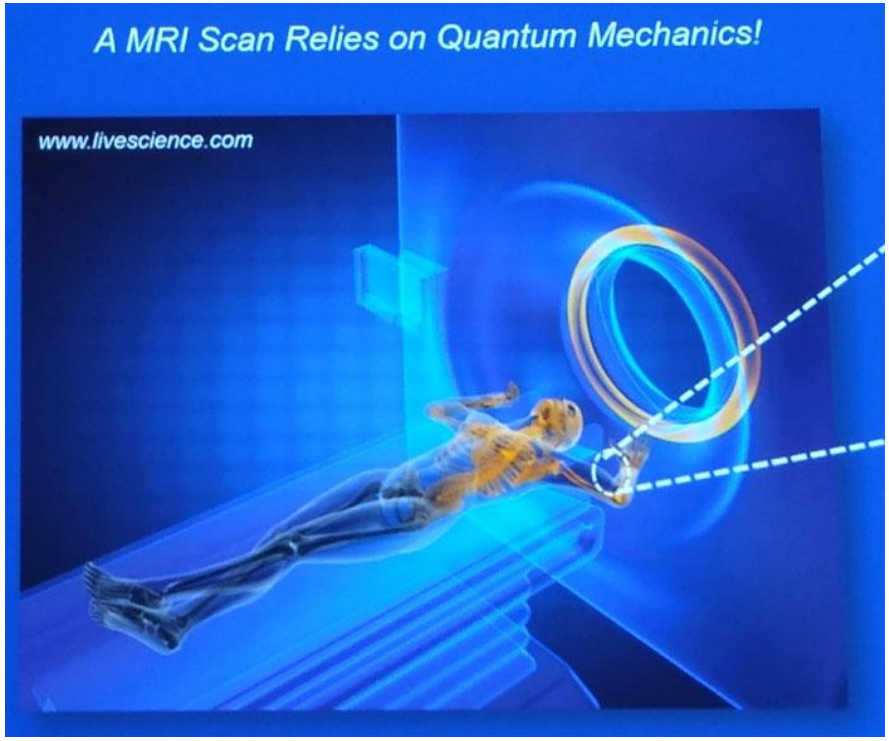
ALUMNUS

ALUMNUS

ALUMNUS

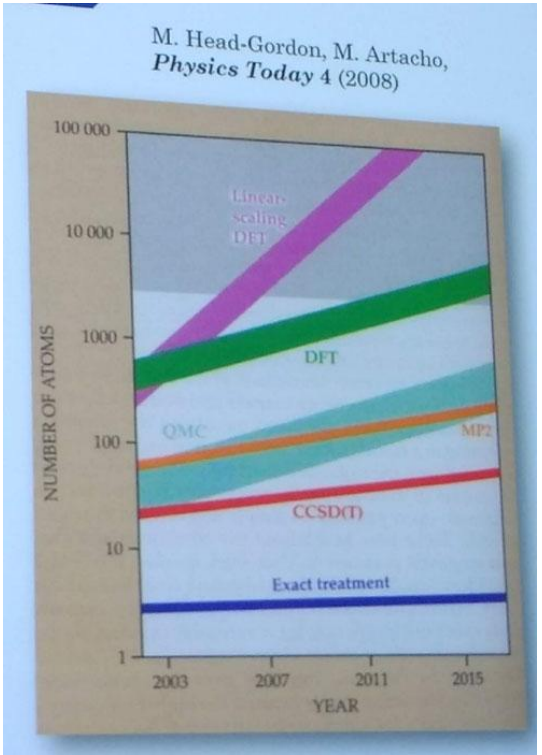
Hardware 2.0

- Quantum Processors
  - Does quantum mechanics work? The MRI!
  - Difference between the MRI and quantum computing: the MRI measures average properties, we need to measure each spin individually



Hardware 2.0

- Quantum Processors
  - How to make quantum materials: engineers “encourage” nature to assemble stable structures, and nature does the rest
  - Problem of entanglement: in order to describe 300 entangled qubits, we need more numbers than particles in the universe
  - Many different techniques… see graph of when we expect to be able to describe n entangled atoms
  - Superconducting circuits



Hardware 2.0

- Centers of quantum computing
  - D-Wave (Canada)
  - QuTech (Netherlands)
  - Nokia/Bell Labs (New Jersey)
  - IBM (upstate New York)
  - UC Santa Barbara

Hardware 2.0

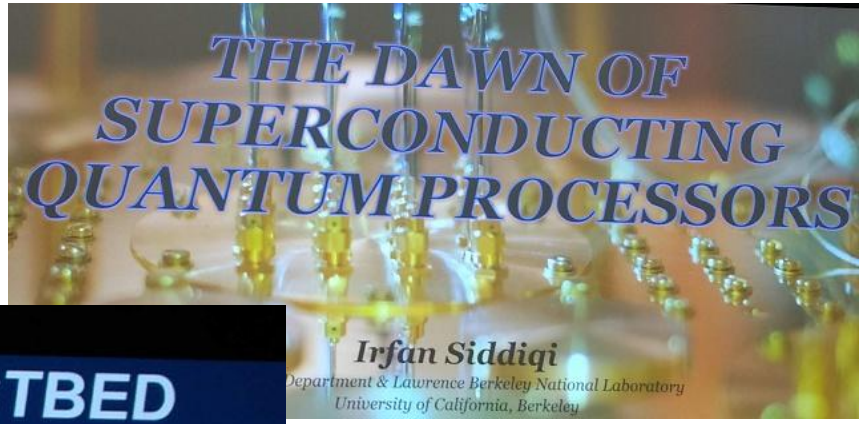
- Quantum-computing startups
  - QA Ware (Palo Alto, 2014): Forge cloud platform
  - Rigetti (Berkeley, 2013): quantum cloud computing
  - PsiQ (Palo Alto, 2019): will encode information in photons instead of electrons

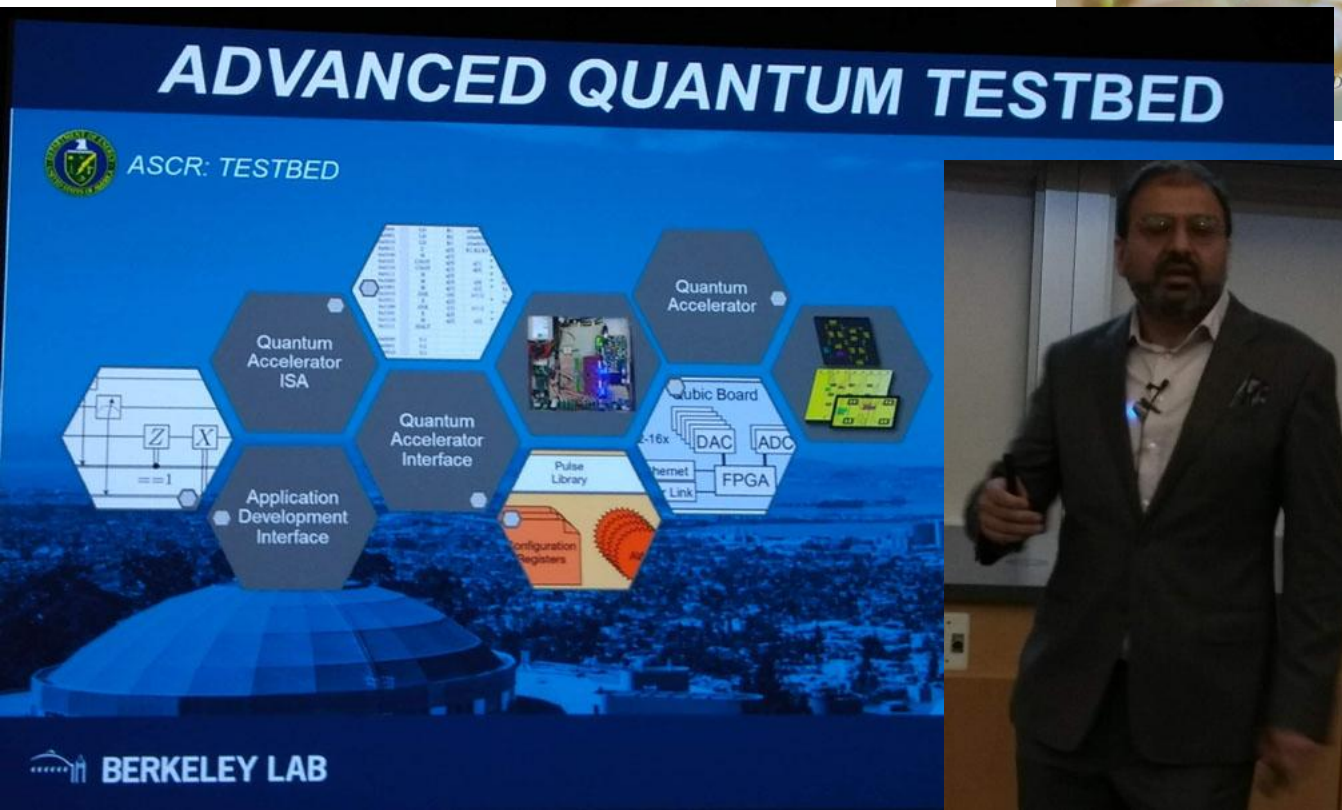
Hardware 2.0

- Quantum Processors
  - Superconducting circuits are the most promising platform
  - UC Berkeley

Hardware 2.0

- Superconducting Quantum Processors

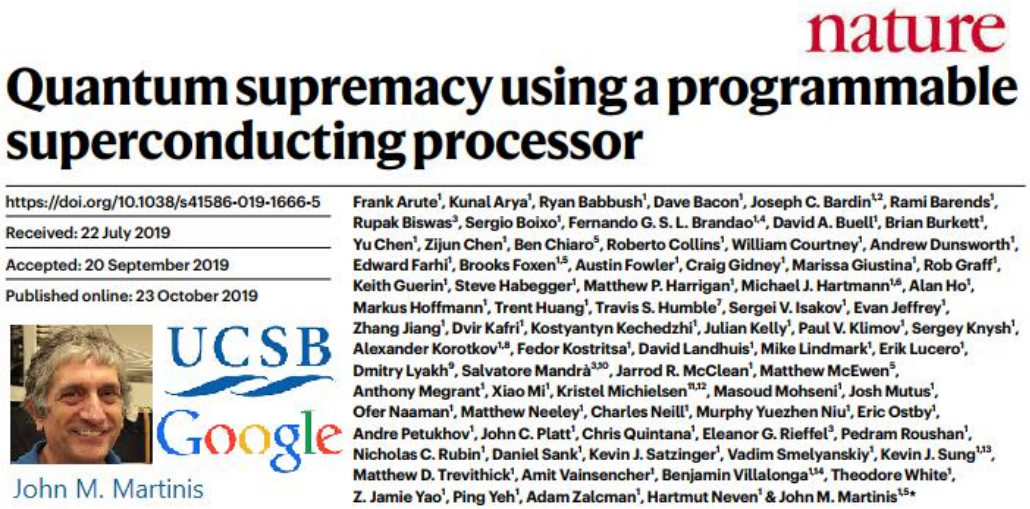




Irfan Siddiqi
(UC Berkeley)
28 May 2019

Hardware 2.0

- October 2019: John Martinis (UC Santa Barbara & Google) announces supercomputing supremacy



nature

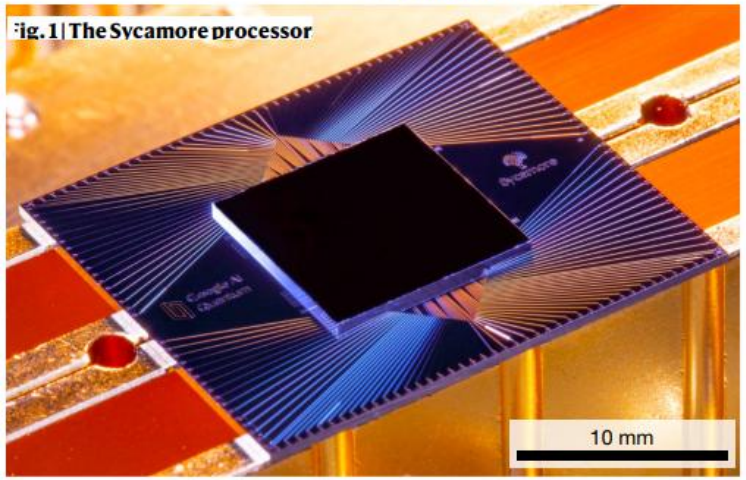
Quantum supremacy using a programmable superconducting processor

https://doi.org/10.1038/s41586-019-1666-5
Received: 22 July 2019
Accepted: 20 September 2019
Published online: 23 October 2019

Frank Arute[1], Kunal Arya[1], Ryan Babbush[1], Dave Bacon[1], Joseph C. Bardin[1,2], Rami Barends[1], Rupak Biswas[3], Sergio Boixo[1], Fernando G. S. L. Brandao[1,4], David A. Buell[1], Brian Burkett[1], Yu Chen[1], Zijun Chen[1], Ben Chiaro[5], Roberto Collins[1], William Courtney[1], Andrew Dunsworth[1], Edward Farhi[1], Brooks Foxen[1,5], Austin Fowler[1], Craig Gidney[1], Marissa Giustina[1], Rob Graff[1], Keith Guerin[1], Steve Habegger[1], Matthew P. Harrigan[1], Michael J. Hartmann[1,6], Alan Ho[1], Markus Hoffmann[1], Trent Huang[1], Travis S. Humble[7], Sergei V. Isakov[1], Evan Jeffrey[1], Zhang Jiang[1], Dvir Kafri[1], Kostyantyn Kechedzhi[1], Julian Kelly[1], Paul V. Klimov[1], Sergey Knysh[1], Alexander Korotkov[1,8], Fedor Kostritsa[1], David Landhuis[1], Mike Lindmark[1], Erik Lucero[1], Dmitry Lyakh[9], Salvatore Mandrà[3,10], Jarrod R. McClean[1], Matthew McEwen[5], Anthony Megrant[1], Xiao Mi[1], Kristel Michielsen[11,12], Masoud Mohseni[1], Josh Mutus[1], Ofer Naaman[1], Matthew Neeley[1], Charles Neill[1], Murphy Yuezhen Niu[1], Eric Ostby[1], Andre Petukhov[1], John C. Platt[1], Chris Quintana[1], Eleanor G. Rieffel[3], Pedram Roushan[1], Nicholas C. Rubin[1], Daniel Sank[1], Kevin J. Satzinger[1], Vadim Smelyanskiy[1], Kevin J. Sung[1,13], Matthew D. Trevithick[1], Amit Vainsencher[1], Benjamin Villalonga[1,14], Theodore White[1], Z. Jamie Yao[1], Ping Yeh[1], Adam Zalcman[1], Hartmut Neven[1] & John M. Martinis[1,5*]

UCSB Google

John M. Martinis



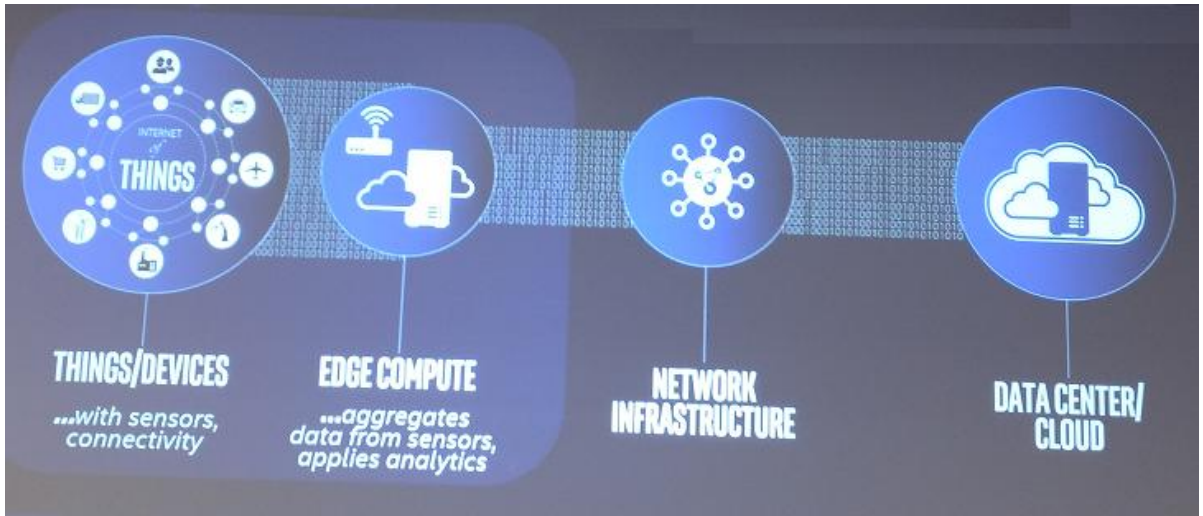
John Martinis (UC Santa Barbara) September 2019

Hardware 2.0

- Superconducting Quantum Processors
  - Suitable for specific applications
  - Suitable for improving classical algorithms, eg “quantum recommendation systems” for Netflix and Amazon

Hardware 2.0

- Edge Computing
  - Edge computing: gathering data from sensors and computing actions in real time
  - Edge server (on board) vs data center (on the cloud)
  - Internet of Things -> Edge server -> Internet -> Data center



Gary Brown of Intel showing an edge server at Stanford (Oct 2019)

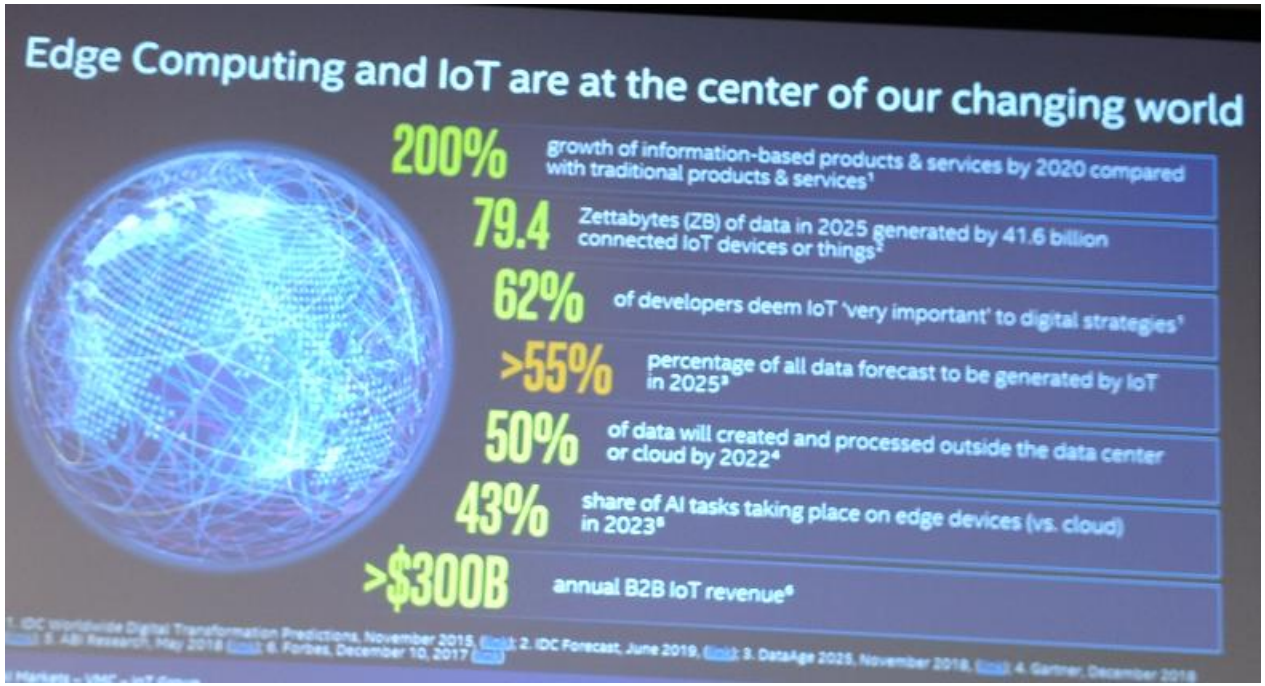
Hardware 2.0

- Edge Computing
  - Why the need for edge computing: pedestrian recognition in a self-driving car cannot depend on cloud computing

Bringing visual intelligence to products,
at the edge and in the cloud.

Hardware 2.0

- Edge Computing
  - Market

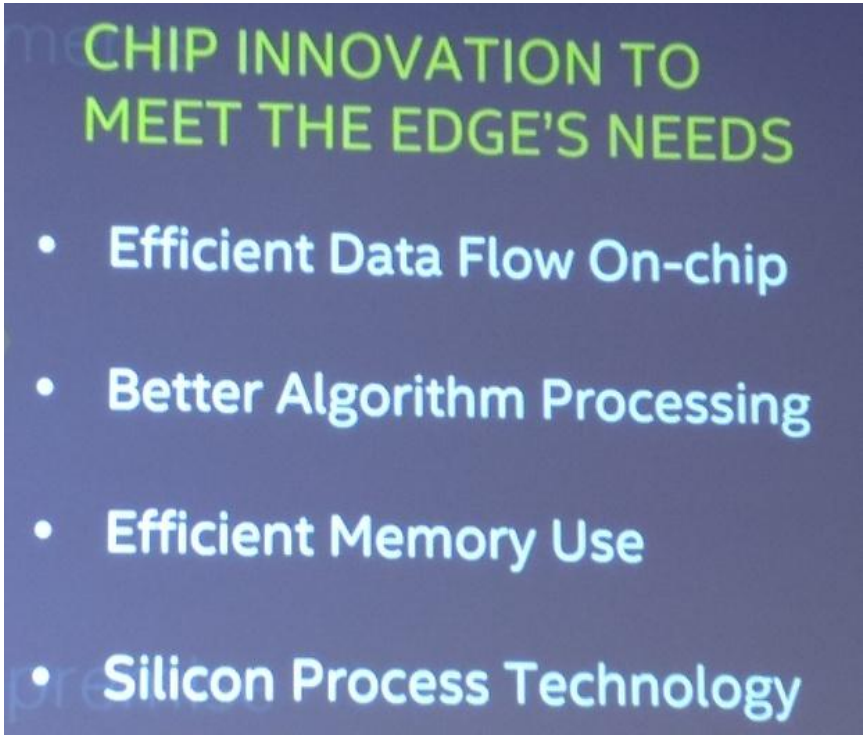


Hardware 2.0

- Edge Computing
  - Applications

Hardware 2.0

- Edge Computing
  - Requirements on high-performance chips



Hardware 2.0

- Edge Computing
  - New chips at Hot Chips 2019

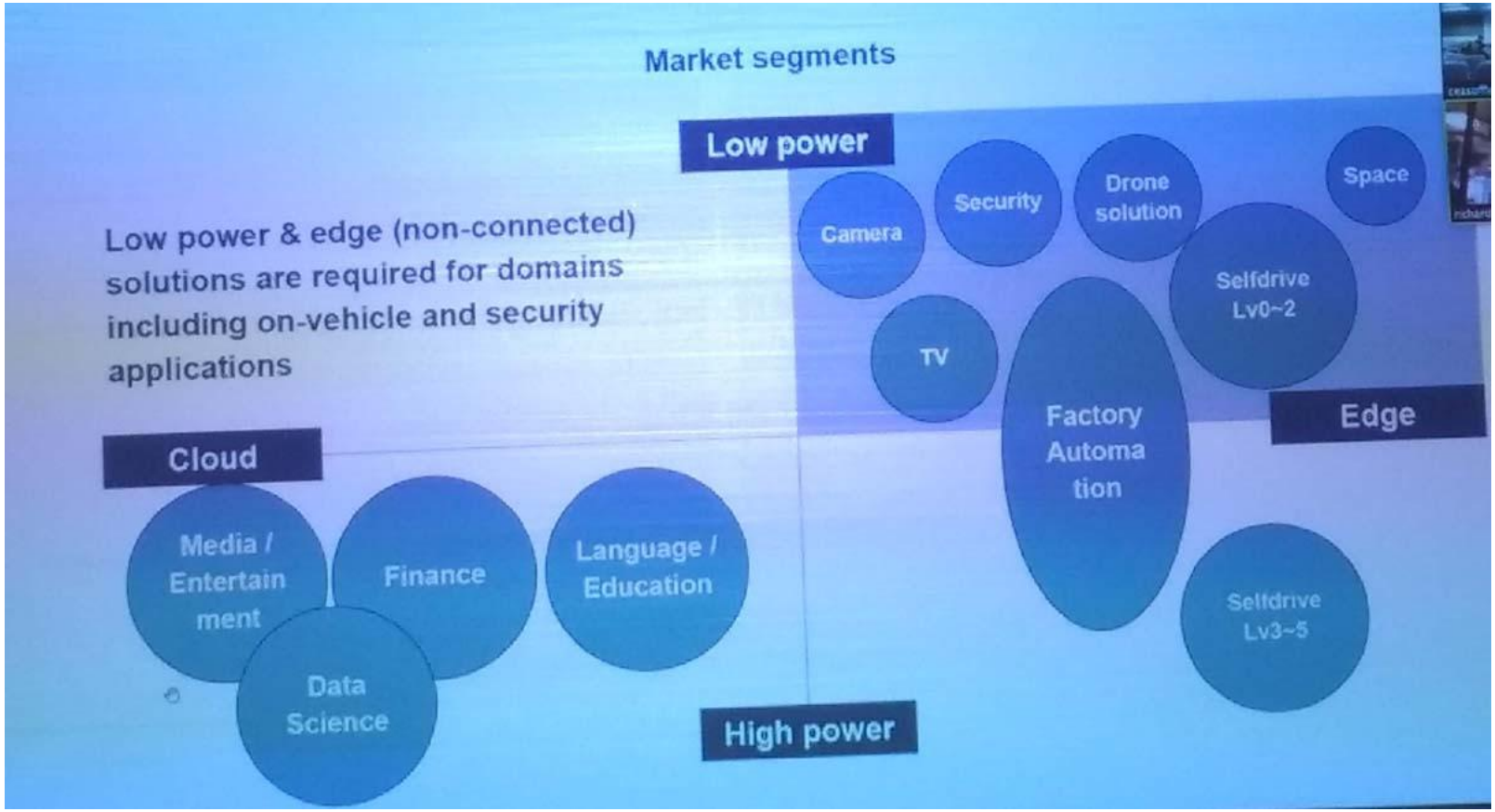
Hot Chips: A Symposium on High Performance Chips

Stanford University, Palo Alto, California, Sunday-Tuesday, August 18-20, 2019.

Company	What was Presented
Intel	DL Inference chip "NNP-I" (codename Spring Hill)
Habana	DL Inference chip "Goya"
Huawei	DL Inference IP core "Da Vinci"
Nvidia	DL Inference Accelerator multi-chip module
Tesla	FSD* computer with 2 chips *FSD = Full Self-Driving

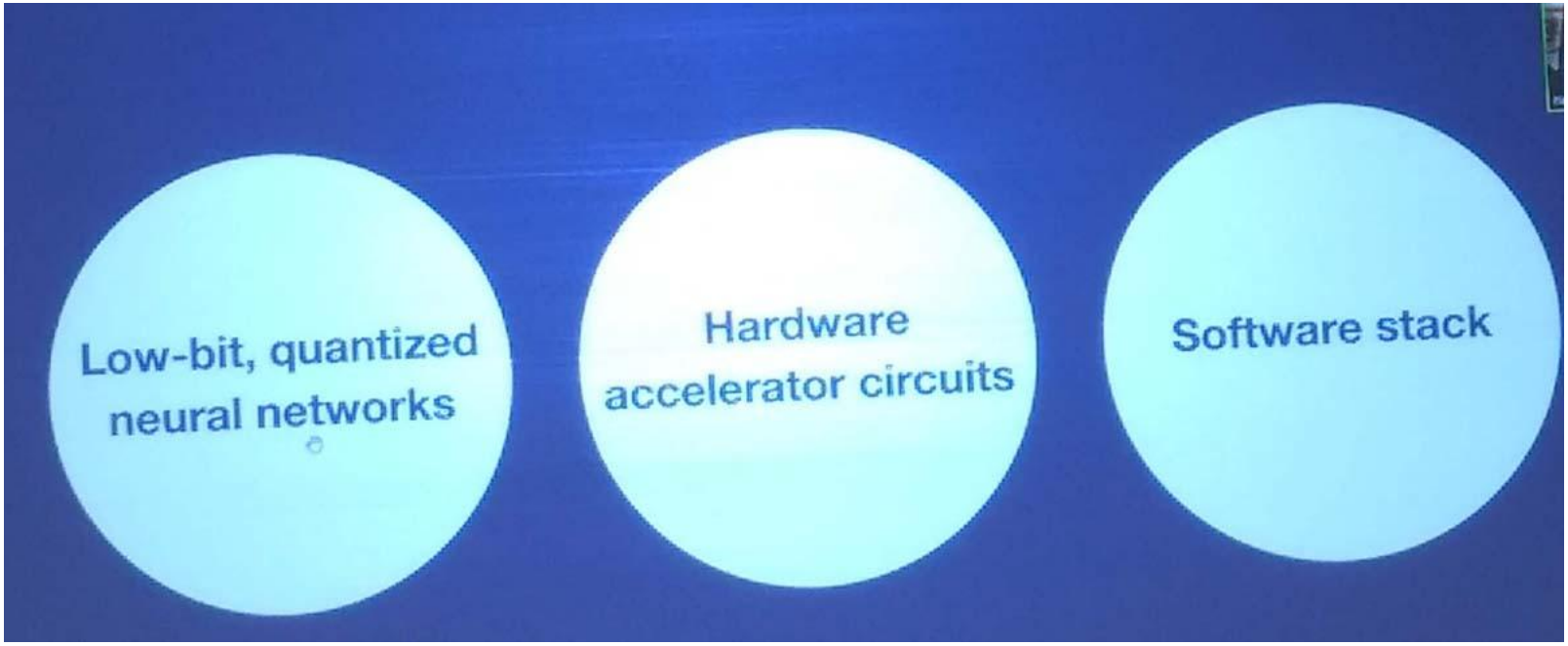
Hardware 2.0

- Edge Computing
  - Low-power and non-connected solutions



Hardware 2.0

- Edge Computing
  - Deep Learning solutions



Hardware 2.0

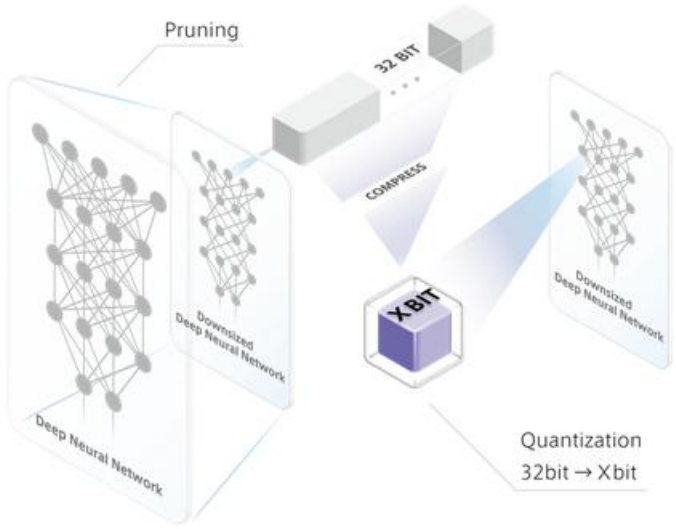
- Edge Computing
  - Deep Learning solutions: example of target device in 2019 (Terasic DE10-Nano, built around the Intel® Cyclone® V System-on-Chip (SoC) FPGA)

Hardware 2.0

- Edge Computing
  - Edge servers

https://leapmind.io/

Our technologies for energy-efficient deep learning on edge will be the key to implementing intelligent machines in every aspect of our world.



Deep learning compression and acceleration

Dr. Atsunori Kanemura, Chief Research Officer and Chief Scientist, LeapMind

NOVEMBER 14, 2019 - 4:30PM TO 6:00PM

SKILLING AUDITORIUM, 494 LOMITA MALL, STANFORD, CA

LeapMind (Japan)

Hardware 2.0

- Edge Computing
  - Edge servers

Intel NUC mini-PC

Hardware 2.0

- Edge Computing
  - Edge servers

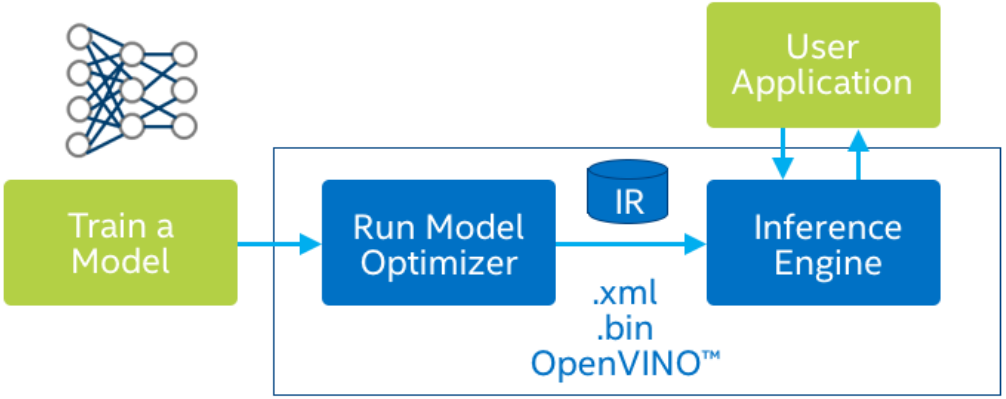
Visionseed (Japan)

Hardware 2.0

- Edge Computing
  - Development Kits

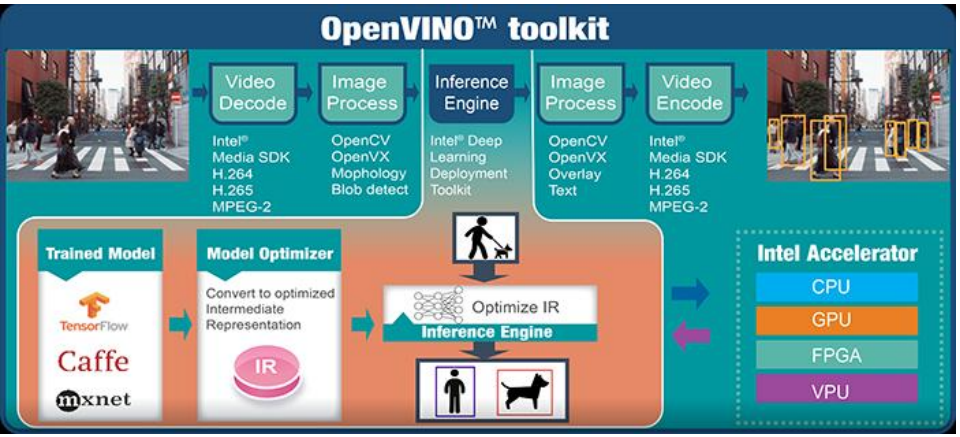
VISION PRODUCT OF THE YEAR 2019

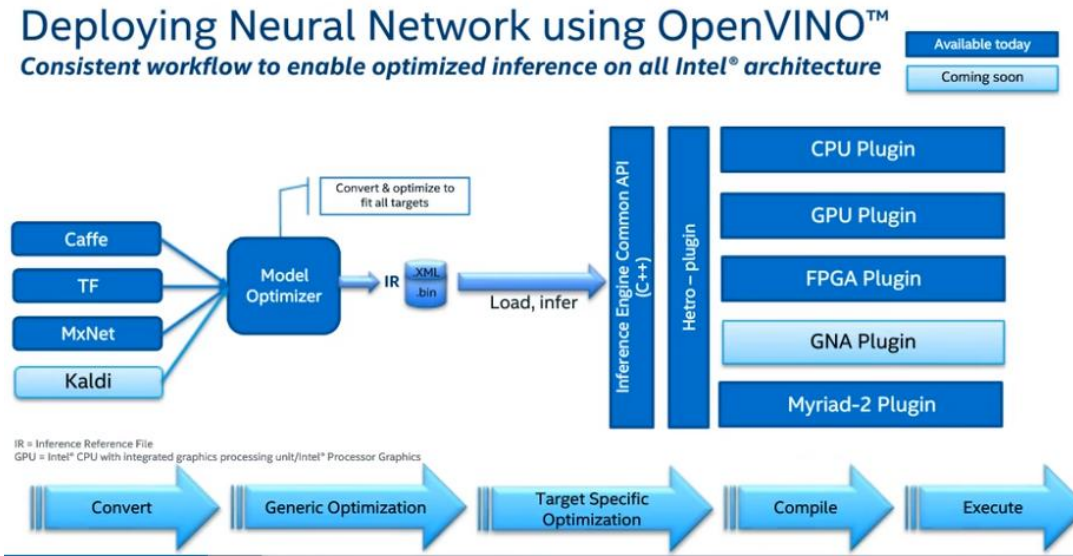
Intel OpenVino
(Open Visual Inferencing and Neural Network Optimization)



Hardware 2.0

- Edge Computing
  - Development Kits



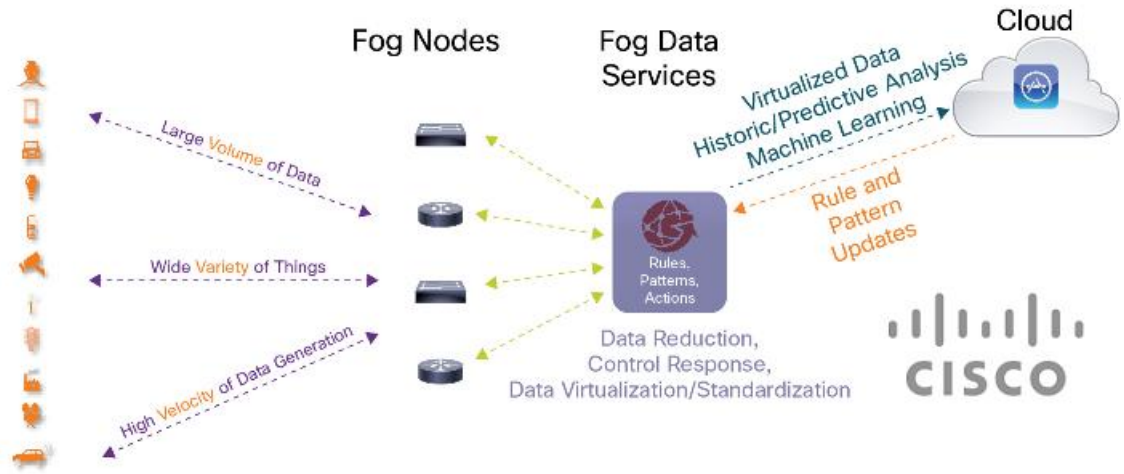


Hardware 2.0

- Edge Computing
  - A new kind of software, e.g. Morpho (Japan)
    - Sep 2019: 3 billion licenses worldwide!

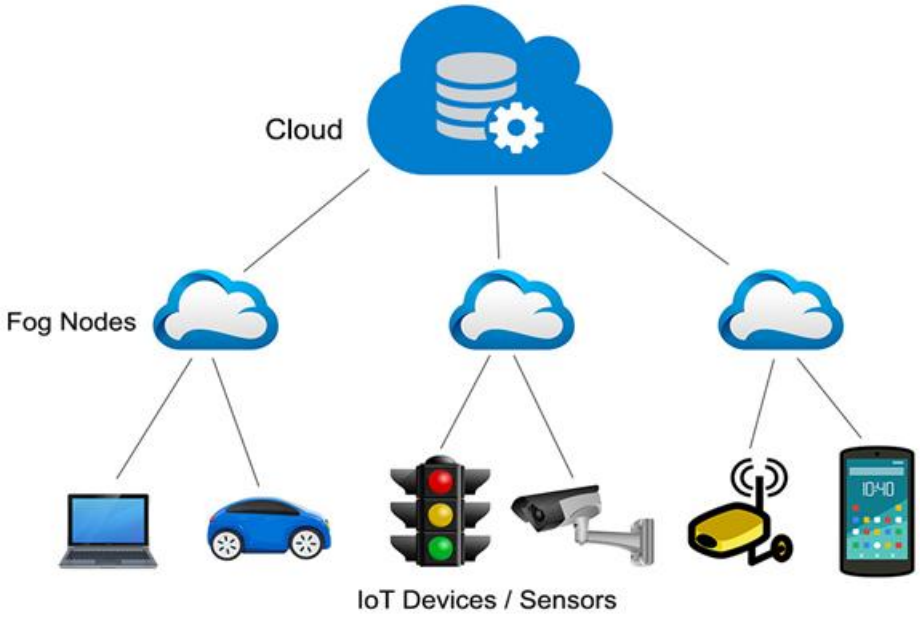
Hardware 2.0

- Edge Computing
  - Fog Computing (Cisco, 2014) pushes data toward the Edge and away from the Cloud
  - Cloud computing: centralized system
  - Fog computing: a distributed decentralized system
    - Low latency
    - High security



Hardware 2.0

- Edge Computing
  - Fog Computing



Hardware 2.0

- Edge Computing
  - Consolidation of workload on edge server

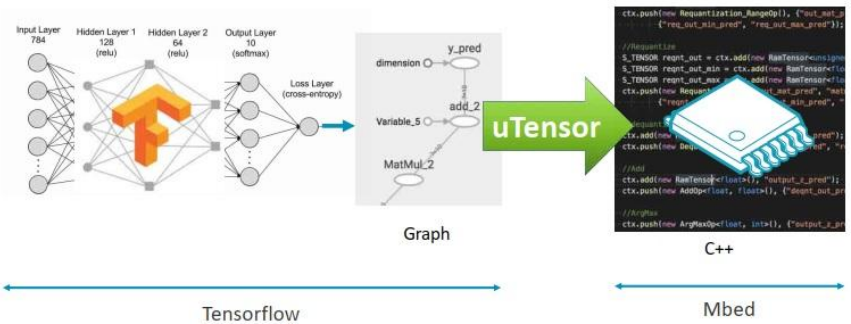
Hardware 2.0

- Tiny machine learning (tinyml)
  - Machine-learning technology on ultra-low power hardware (custom integrated circuits) capable of processing sensor data
  - Enabling a new class of devices: always-on battery-operated devices
  - First tinyML Summit: March 2019
  - Low-power applications for vision and audio
  - Software platforms for embedded environments: Google TensorFlow Lite
  - Problems: no standards, no open-source hardware, no software abstraction layers

TensorFlowLite + arm microTensor

https://www.tensorflow.org/lite/microcontrollers/overview

uTensor http://utensor.ai

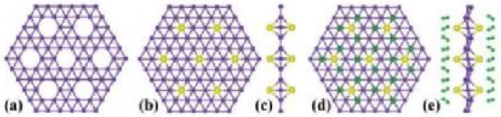


Hardware 2.0

- Processors for robots
  - Eg Nvidia's Jetson

New materials

- Graphene: a single layer of carbon atoms
- Borophene: a single layer of boron atoms
- Two-dimensional materials
- Very strong, flexible and superconducting



Borophene

- First synthesized in 2015 by Nathan Guisinger's team in the USA and Kehui Wu's team in China
- Free standing borophene synthesized in 2019 by Nathan Guisinger's team

Science *Science*. 2015 December 18

Synthesis of borophenes: Anisotropic, two-dimensional boron polymorphs

Andrew J. Mannix[1,2], Xiang-Feng Zhou[3,4], Brian Kiraly[1,2], Joshua D. Wood[2], Diego Alducin[5], Benjamin D. Myers[2,6], Xiaolong Liu[7], Brandon L. Fisher[1], Ulises Santiago[5], Jeffrey R. Guest[1], Miguel Jose Yacaman[5], Arturo Ponce[5], Artem R. Oganov[8,9,3,*], Mark C. Hersam[2,7,10,*], and Nathan P. Guisinger[1,*]

[1] Argonne National Laboratory
[2 7] Northwestern University,
[5]Department of Physics, Univ of Texas San Antonio
[3] Stony Brook University, Stony Brook
[4] Nankai University, Tianjin
[6]NU*ANCE* Center, Northwestern Univ

Nathan Guisinger (Argonne)

Argonne NATIONAL LABORATORY

nature chemistry

PUBLISHED ONLINE: 28 MARCH 2016

Experimental realization of two-dimensional boron sheets

Baojie Feng[1], Jin Zhang[1], Qing Zhong[1], Wenbin Li[1], Shuai Li[1], Hui Li[1]*, Lan Chen[1]* and Kehui Wu[1,2]* **Peng Cheng[1], Sheng Meng[1,2],**

Institute of Physics, Chinese Academy of Sciences, Beijing

Kehui Wu

中国科学院物理研究所
表面物理国家重点实验室 SF09 组

May 14, 2019

Freestanding borophene synthesized

Nathan Guisinger (Argonne)

Argonne NATIONAL LABORATORY

Rectennas

- Rectennas: devices that convert AC electromagnetic waves into DC electricity
- Eliminate batteries: convert Wi-Fi signals to electricity
- Traditional rectennas use either silicon or gallium arsenide for the AC-to-DC rectifier
- Traditional rectennas operate at low frequencies: they cannot feed on the gigahertz frequencies of cell phone and Wi-Fi signals

Rectennas

- Tomás Palacios (MIT, 2019): a rectenna made of a 2-D (three atoms thick) semiconducting material, molybdenum disulfide (MoS2), that is flexible and captures high frequencies