http://inst.eecs.berkeley.edu/~ee290-2/sp21/

Hardware for Machine Learning

Lecture 1: Introduction

Sophia Shao

2pm – 3:30pm

Mondays and Wednesdays

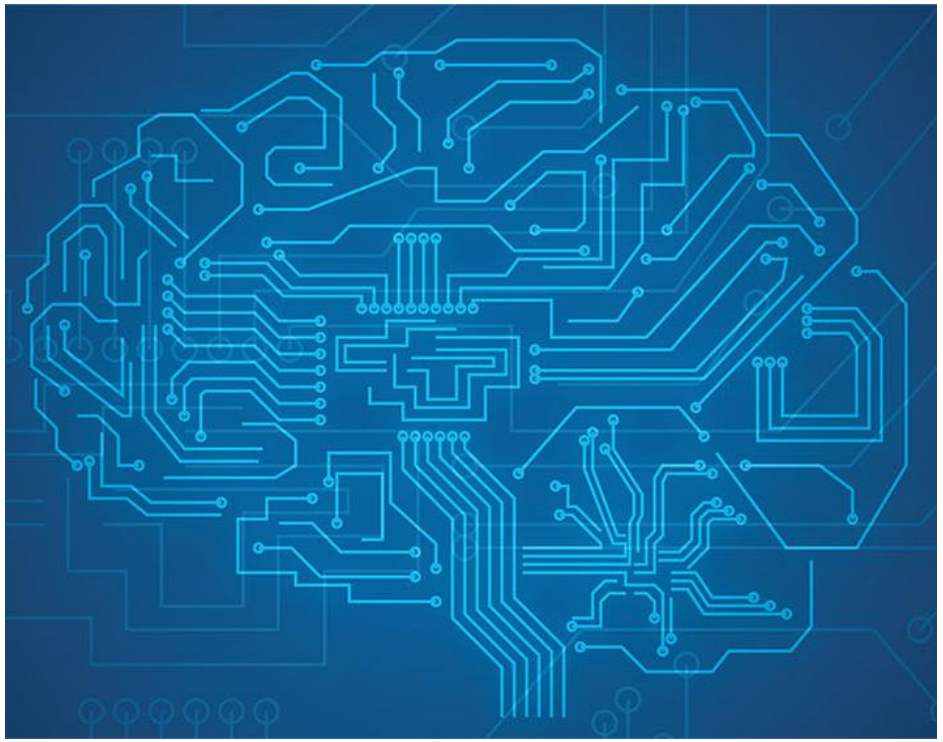
Online

A (relatively) new course!

- A bridge between hardware and machine learning

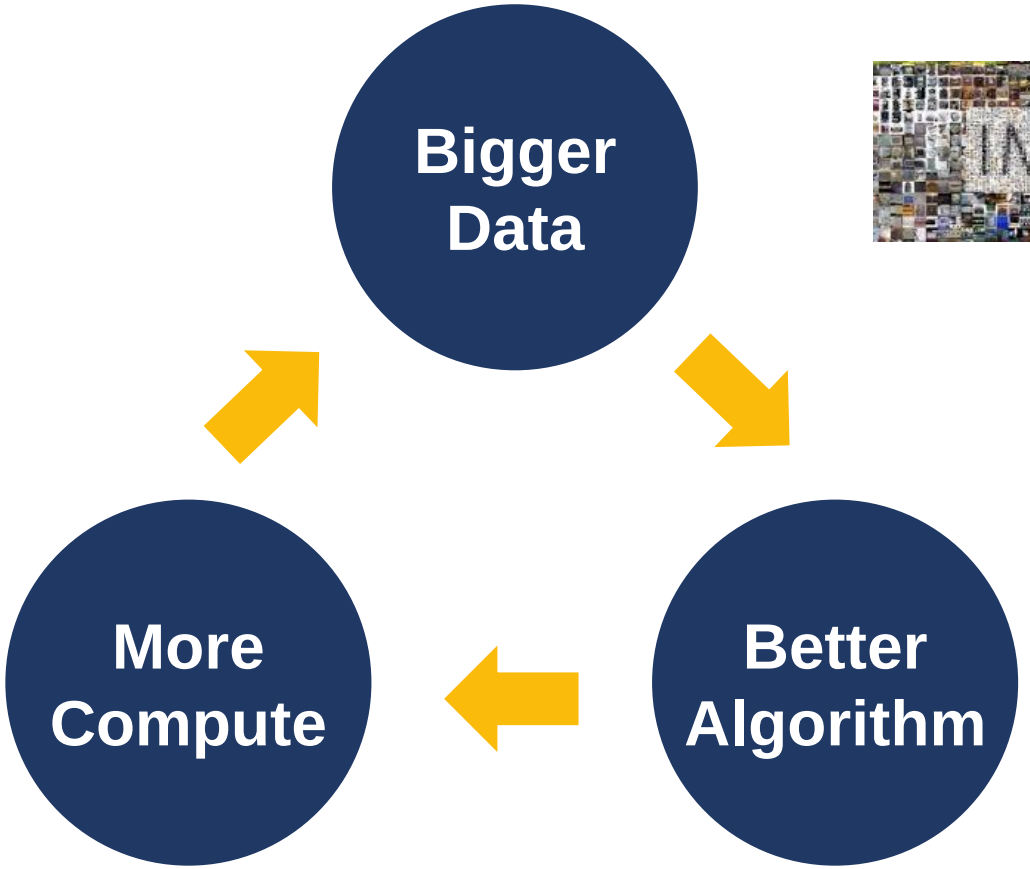
- Goal:
  - **Build** efficient hardware for accelerating machine learning applications.

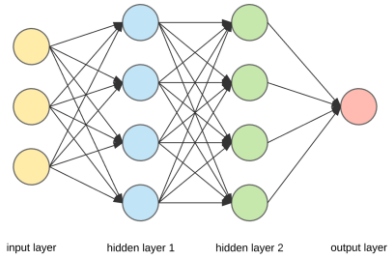
- Approach:
  - Understand key machine learning characteristics
  - Exploit core hardware optimizations
  - Guest lectures with state-of-the-art industry practices



Machine Learning

The Virtuous Circle of Deep Learning

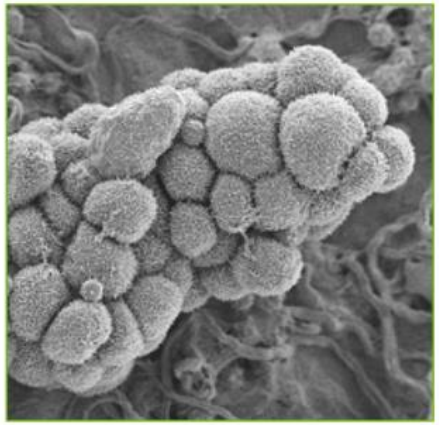




Scaling Application Domains

INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation



MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery

MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

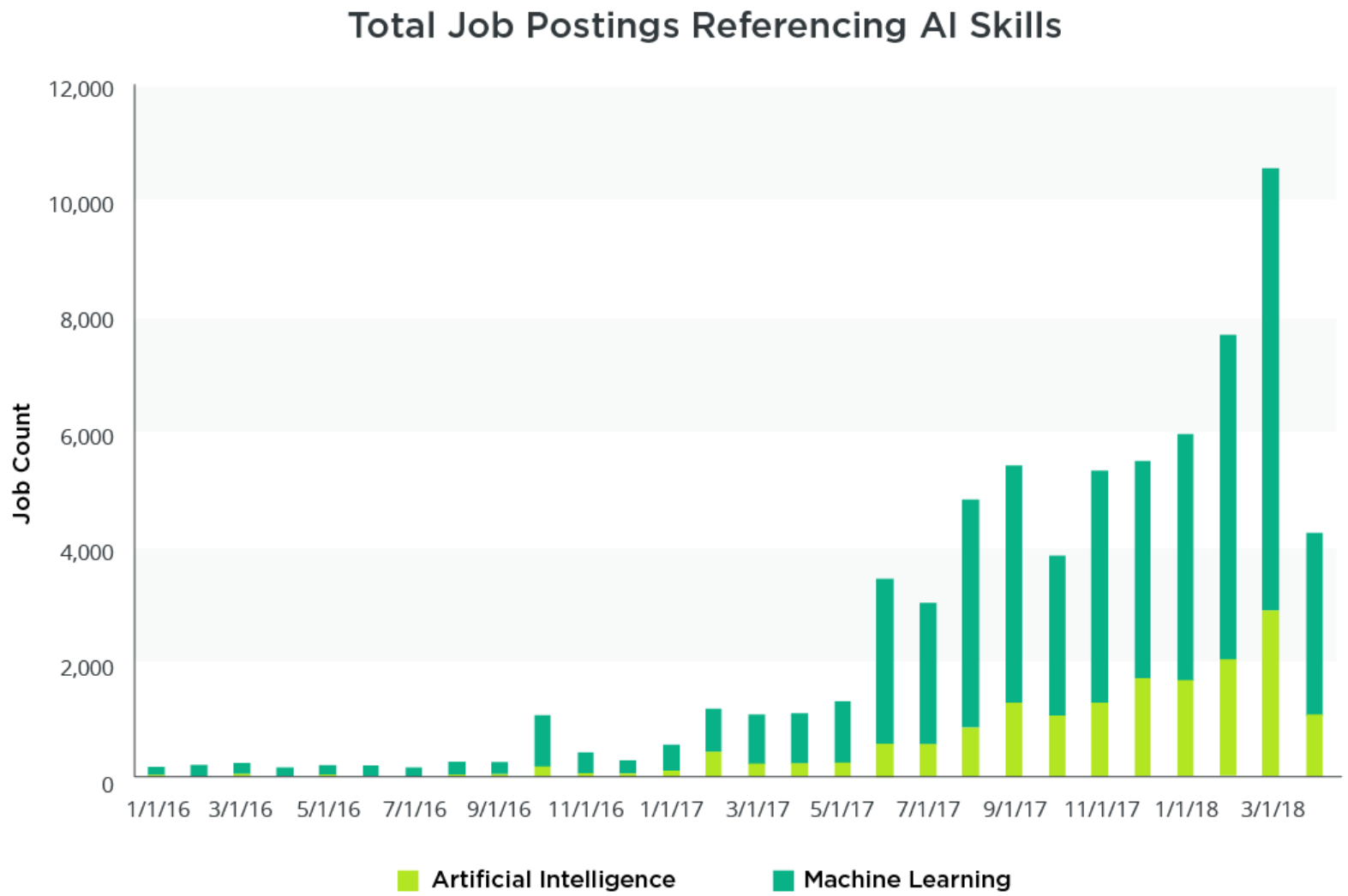
SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

AUTONOMOUS MACHINES

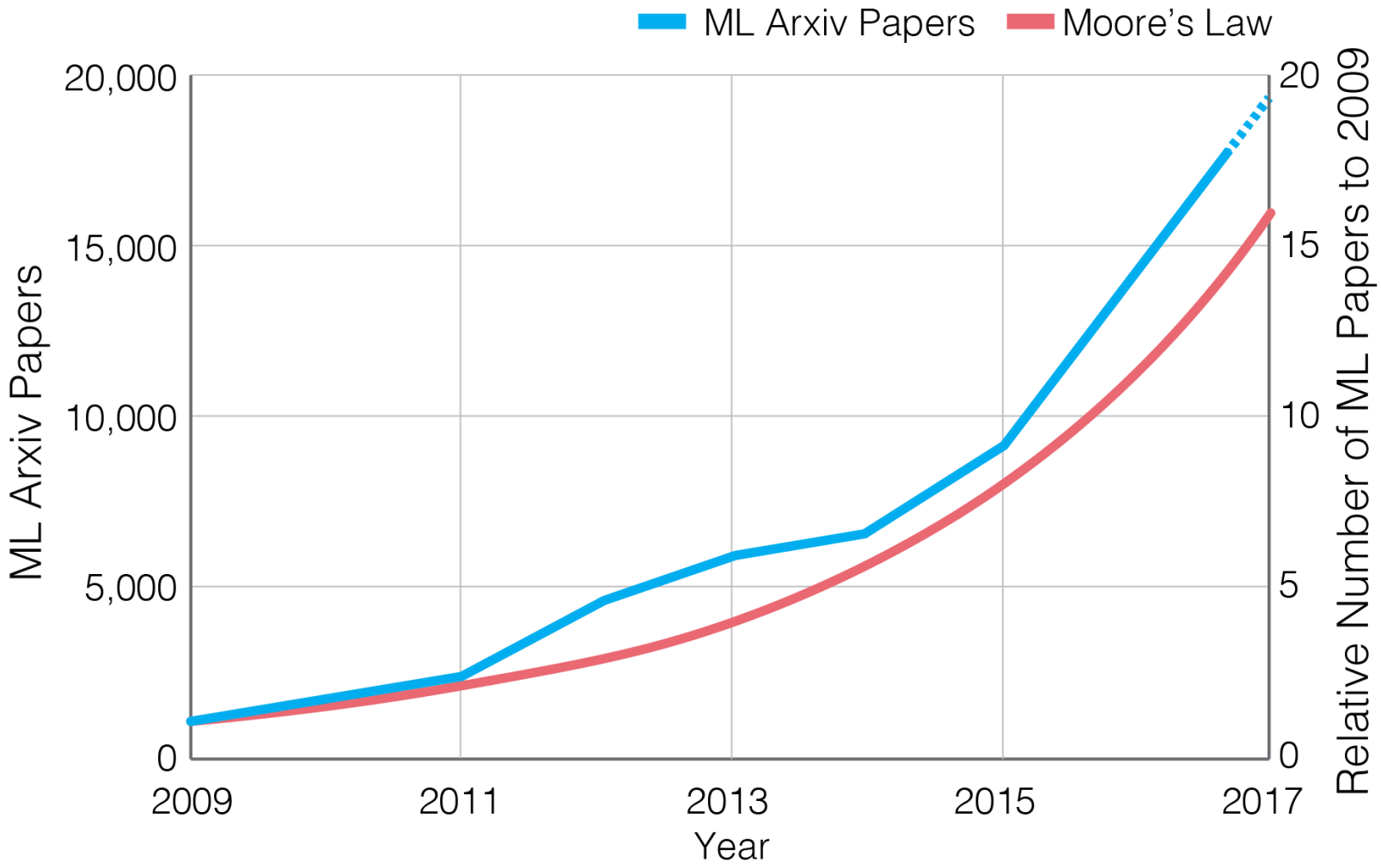
Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

Scaling Deep Learning Jobs



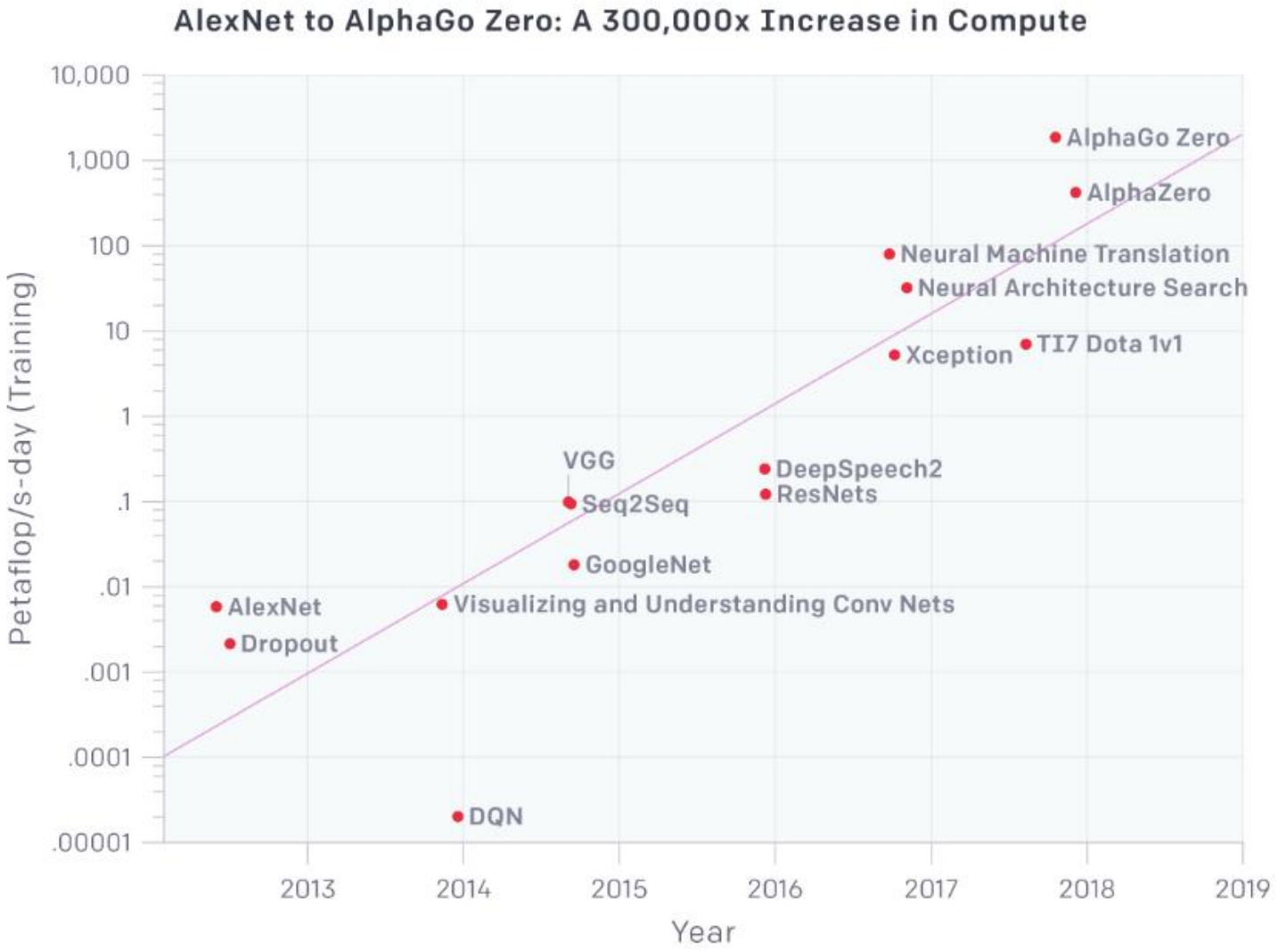
From: ZipRecruiter

Scaling Deep Learning Papers

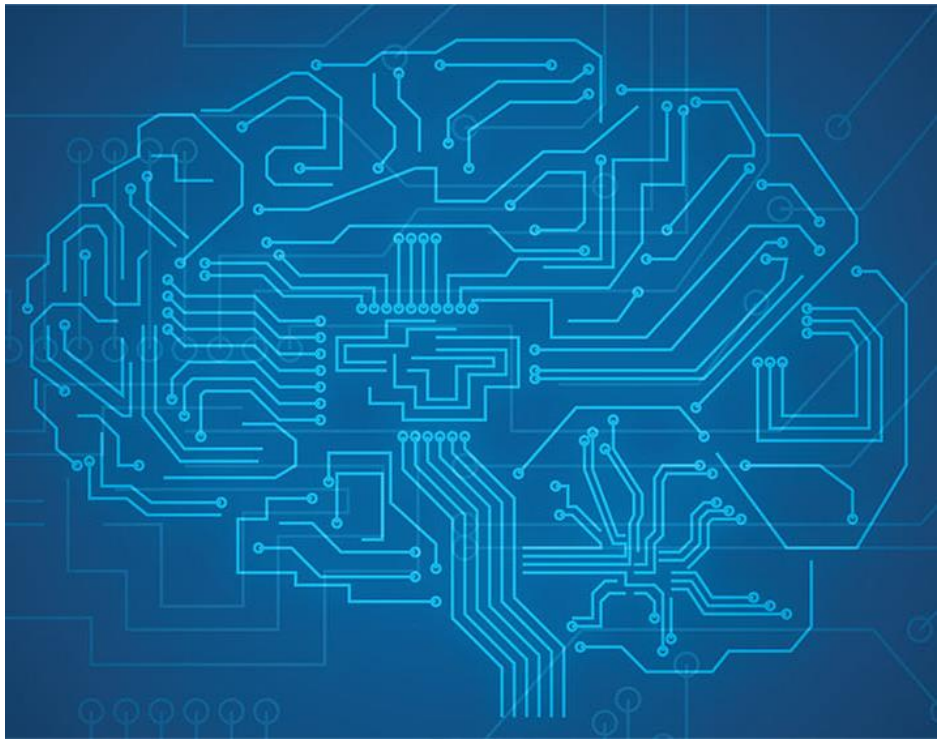


From: J. Dean, 2018

Scaling Deep Learning Models



From: OpenAI



Hardware for Machine Learning

Increasing Demand for Computing

Augmented Reality

Mobile Gaming

4K Video

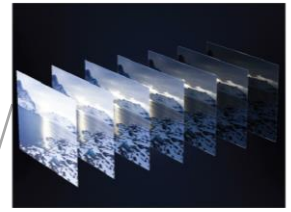
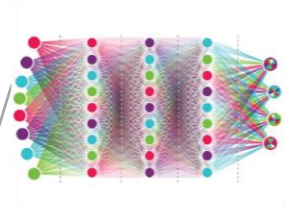


Image Processing

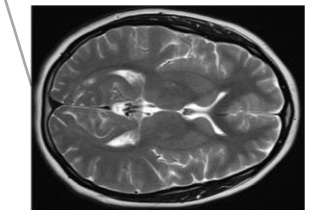
Object Tracking

Obstacle Avoidance



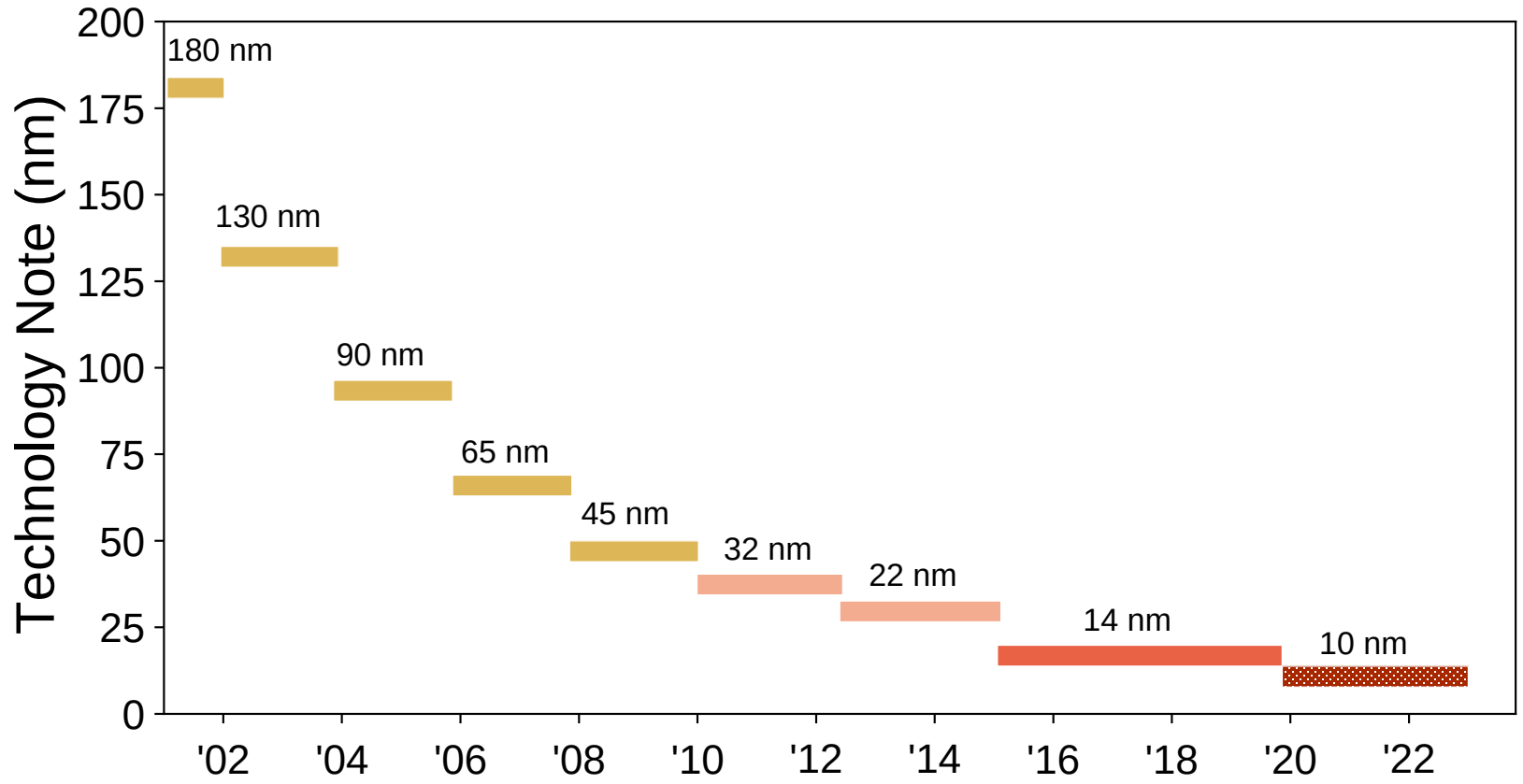
Machine Learning

Genomics



Medical Imaging

Moore's Law Won't Help Us.



* http://www.anandtech.com/show/9447/intel-10nm-and-kaby-lake

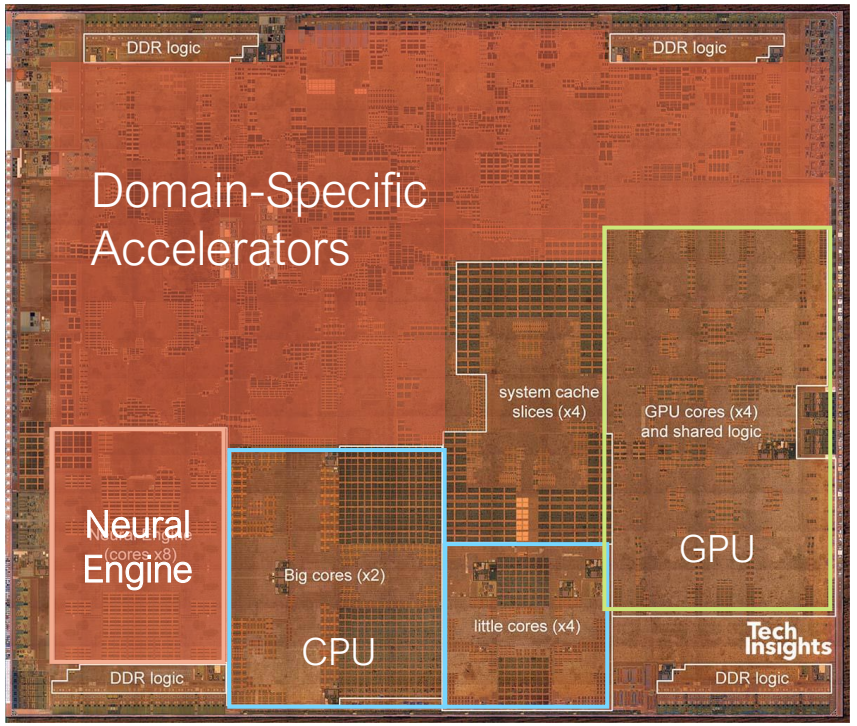
“It was the best of times,
it was the worst of times.”

- *Dickens, A Tale of Two Cities, 1859*

Domain-Specific Accelerators

- Customized hardware designed for a domain of applications.

iPhone XS, 2018

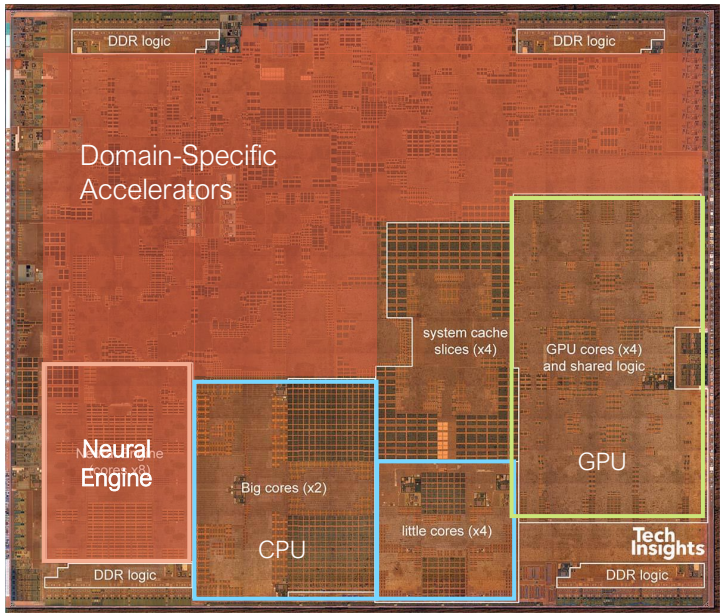


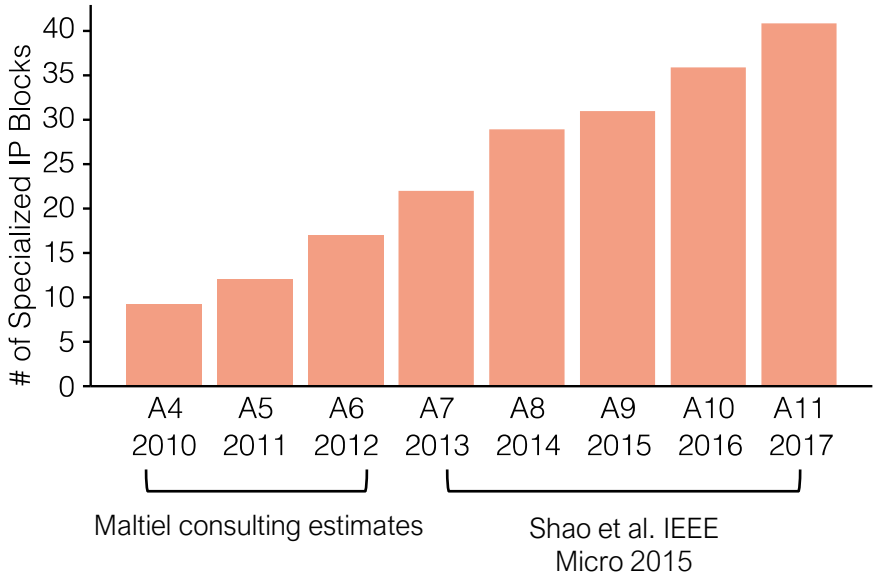
* TechInsights.com Apple iPhone XS teardown

Domain-Specific Accelerators

- Customized hardware designed for a domain of applications.

iPhone XS, 2018





DL has Reinvigorated Hardware.

The New York Times

Big Bets on A.I. Open a New Frontier for Chip Start-Ups, Too

By Cade Metz

Jan. 14, 2018

Today, at least 45 start-ups are working on chips that can power tasks like speech and self-driving cars, and at least five of them have raised more than $100 million from investors. Venture capitalists invested more than $1.5 billion in chip start-ups last year, nearly doubling the investments made two years ago, according to the research firm CB Insights.

Reddi, The Vision Behind MLPerf

DL has Reinvigorated Hardware.

The New York Times

Big Bets on A.I. Open a New Frontier for Chip Start-Ups, Too

By Cade Metz

Jan. 14, 2018

Today, at least 45 start-ups are working on chips that ca
speech and self-driving cars, and at least five of them h
than $100 million from investors. Venture capitalists inv
$1.5 billion in chip start-ups last year, nearly doubling th
made two years ago, according to the research firm CB

INTEL ACQUIRES ARTIFICIAL INTELLIGENCE CHIPMAKER HABANA LABS

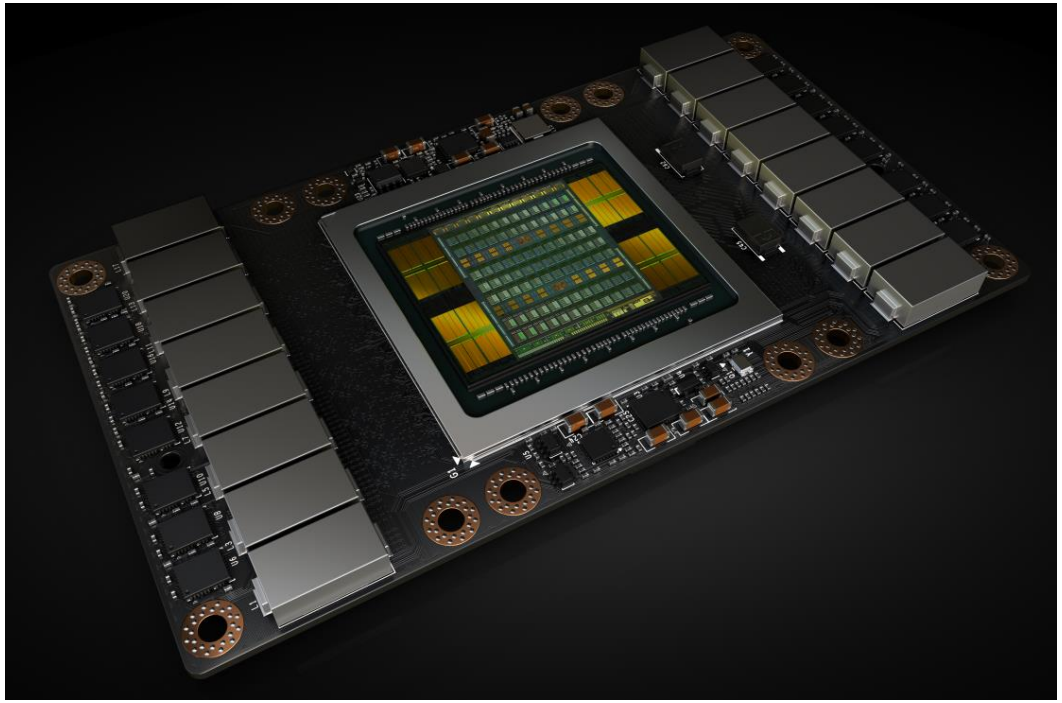
Combination Advances Intel's AI Strategy, Strengthens Portfolio of AI Accelerators for the Data Center

SANTA CLARA Calif., Dec. 16, 2019 – Intel Corporation today announced that it has acquired Habana Labs, an Israel-based developer of programmable deep learning accelerators for the data center for approximately $2 billion. The combination strengthens Intel's artificial intelligence (AI) portfolio and accelerates its efforts in the nascent, fast-growing AI silicon market, which Intel expects to be greater than $25 billion by 2024[1].

"This acquisition advances our AI strategy, which is to provide customers with solutions to fit every performance need – from the intelligent edge to the data center," said Navin Shenoy, executive vice president and general manager of the Data Platforms Group at Intel. "More specifically, Habana turbo-charges our AI offerings for the data center with a high-performance training processor family and a standards-based programming environment to address evolving AI workloads."

NVIDIA GPU

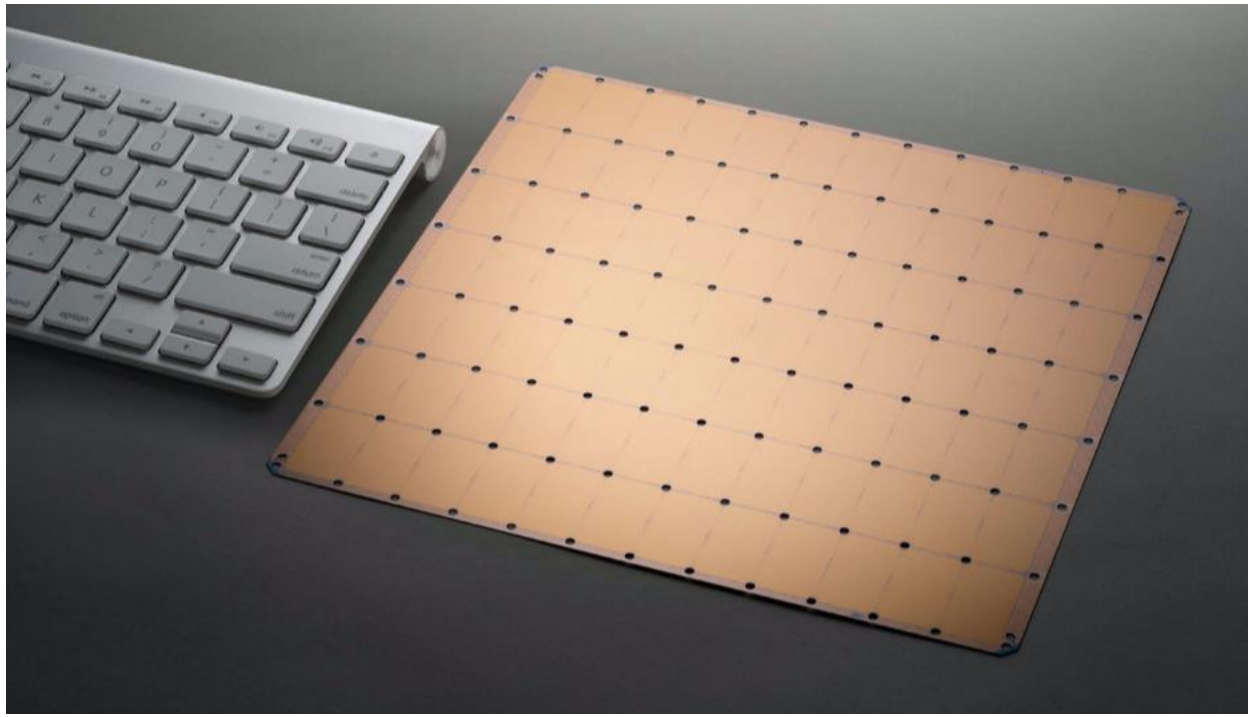
- Volta V100 GPU
  - 21 billion transistors
  - Die size 815 mm2
  - TSMC 12 nm FinFET
  - 15.7 TFLOP/s of single precision (FP32) performance
  - 125 Tensor TFLOP/s of mixed-precision matrix-multiply-and-accumulate
  - TDP 300W



Cerebras: Wafer-Scale Deep Learning

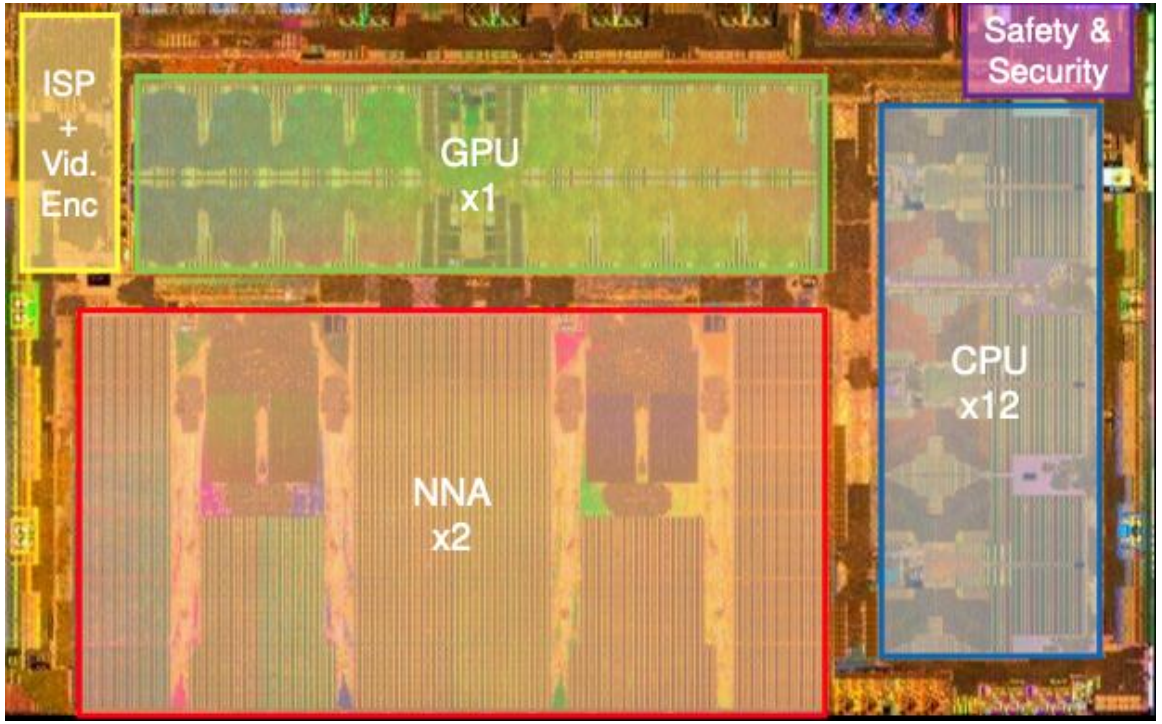
- Largest Chip Ever Built!

- 46,225 mm2 silicon
- 1.2 trillion transistors
- 400,000 optimized AI cores
- 18 GB of on-chip memory
- TSMC 16nm process



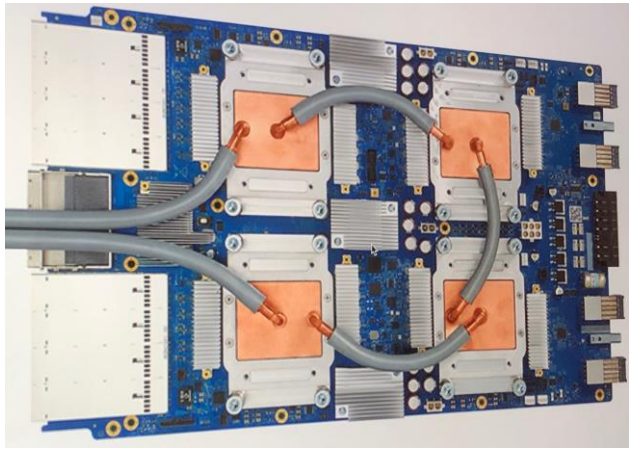
Tesla: Full-Self-Driving Computer

- 2 independent instances
- 2Ghz+ Design
- 32 MB SRAM / instance
- 96*96 MACs

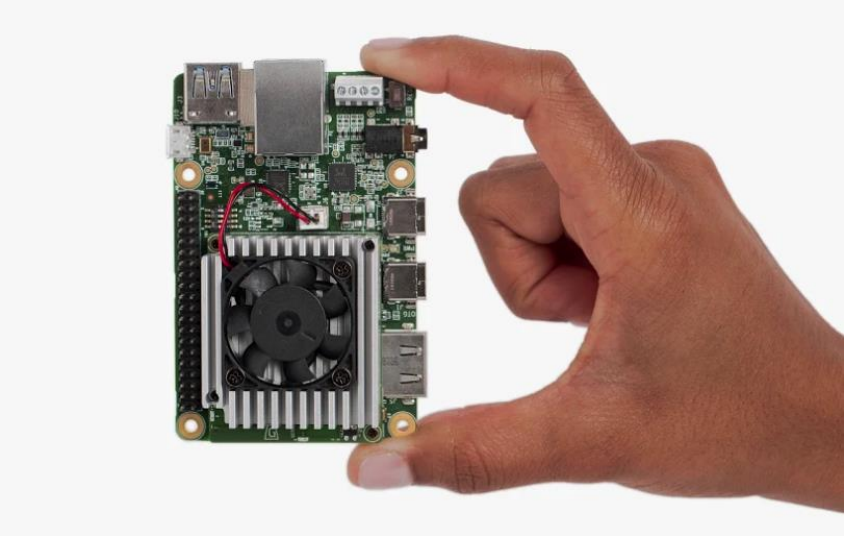


Google TPU

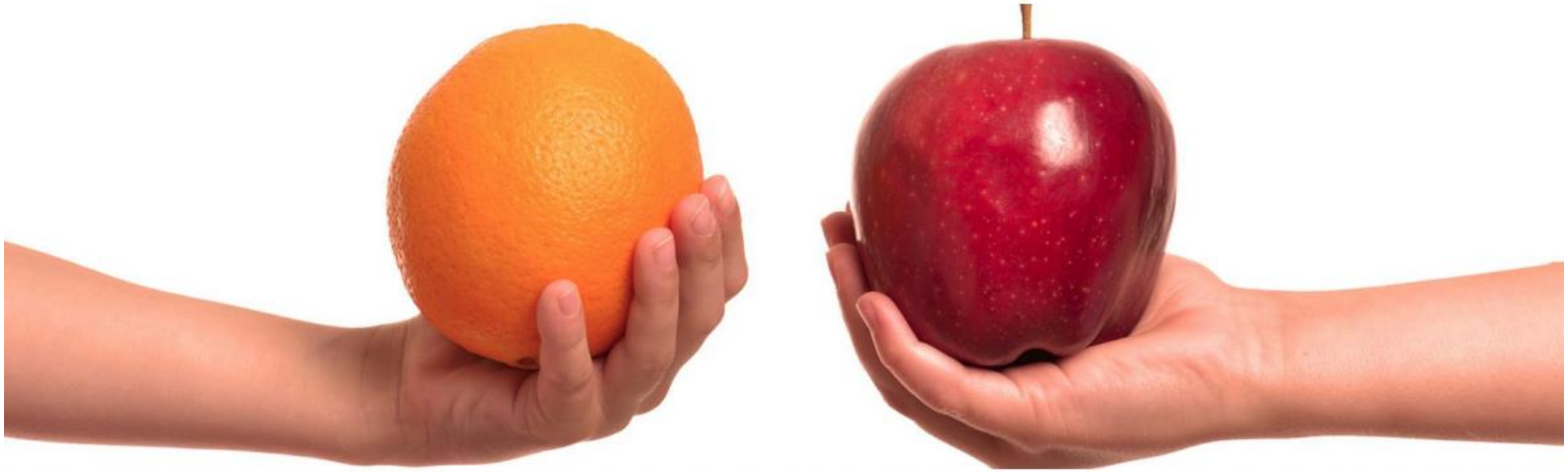
- Systolic-array-based architecture
  - V1: Inference only
  - V2: Training with bfloat
  - V3: 2x powerful than v2
  - V4:?



- Edge TPU
  - Coral Dev Board
  - 4 TOPS
  - 2 TOPS/Watt
  - Supports TensorFlow Lite



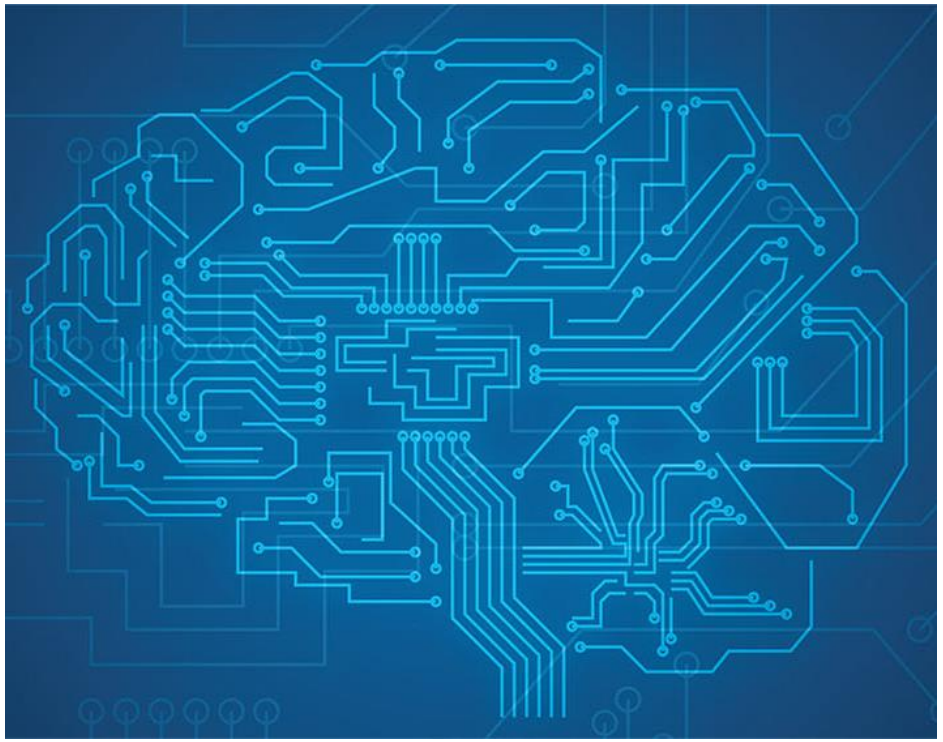
How do we compare the hardware?



Reddi, The Vision Behind MLPerf

MLPerf

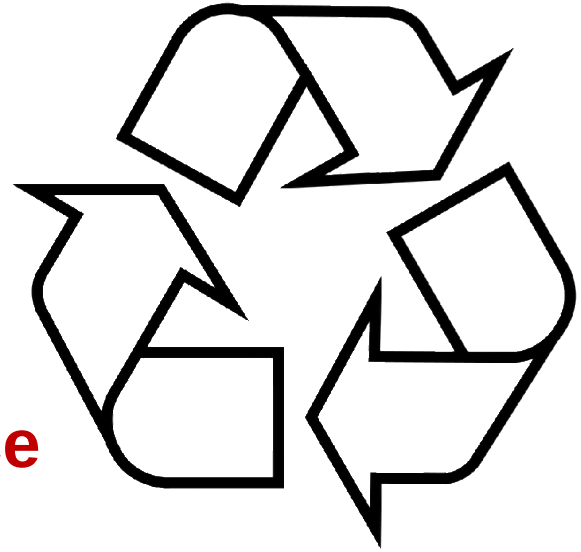
https://mlperf.org/



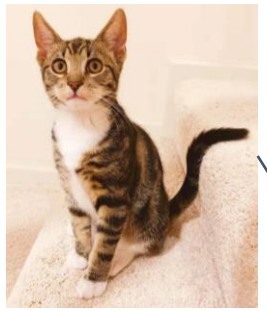
Build Your Own System!

Build Your Own System!

- Build better algorithms
- Build better runtimes
- Build better hardware



At the end of this course



"a visually stunning rumination on love"

Your Algorithms Optimizations

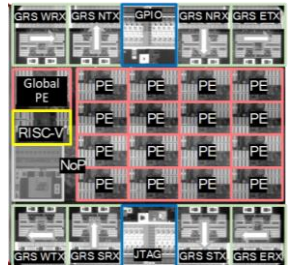
Your Software Optimizations

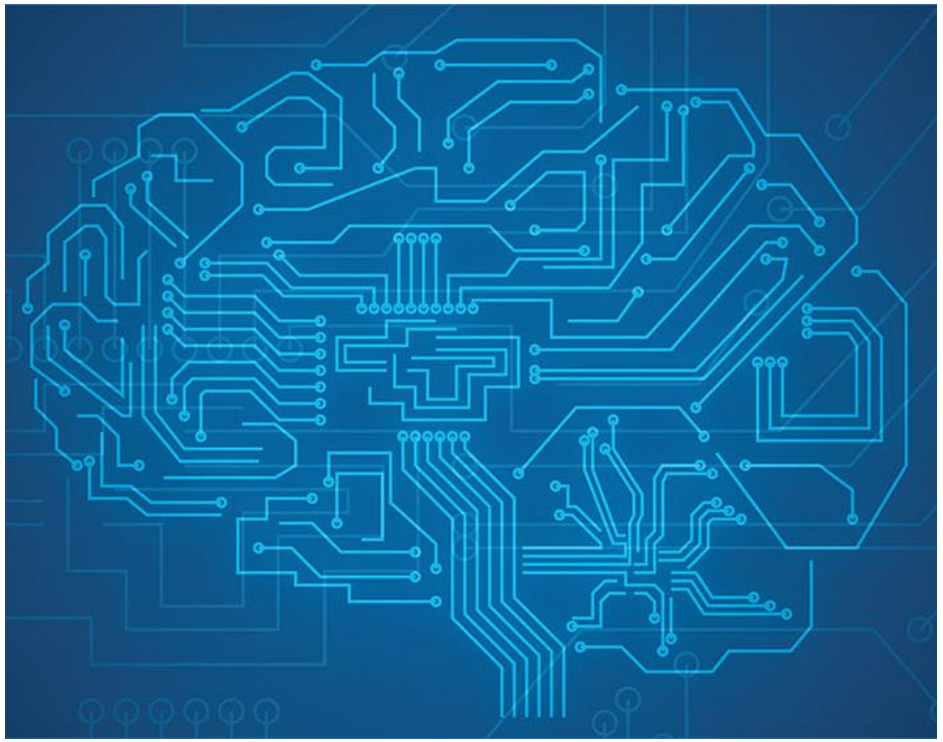
Your Hardware Optimizations

Cat!

Positive!

Competitive Performance, Power, Area!





Course Information

Course Staff

Professor Sophia Shao

ysshao@berkeley.edu
570 Cory Hall
Office Hours:
Tue 1 - 2pm

Abraham (Abe) Gonzalez

abe.gonzalez@berkeley.edu
Office Hours:
Thu 10-11am

Course Information

- Class website:
  - http://inst.eecs.berkeley.edu/~ee290-2/sp21/

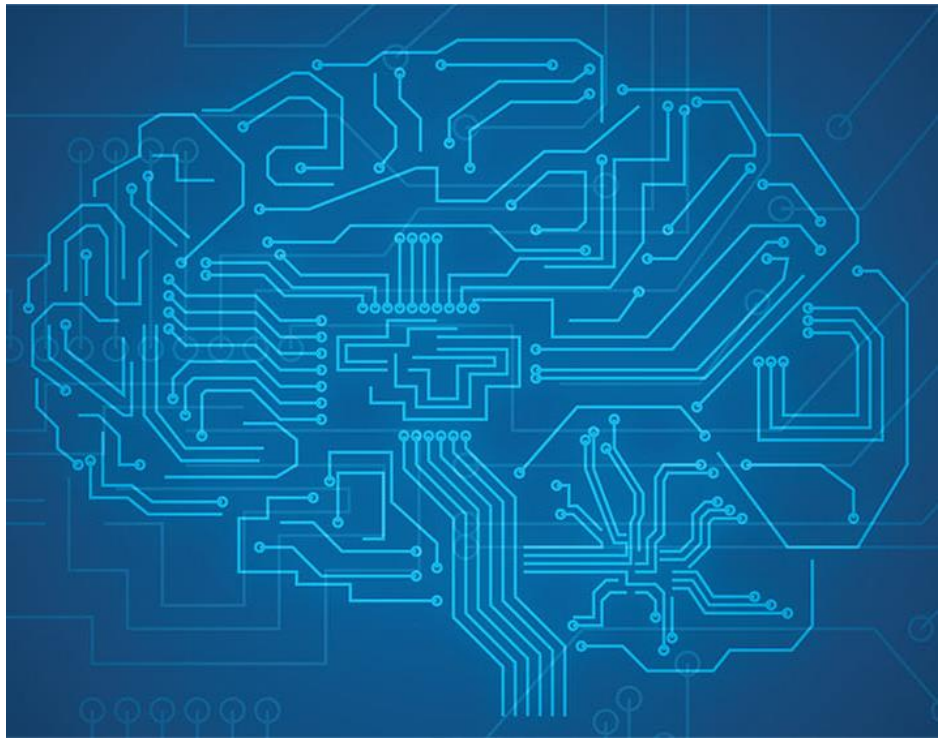
  - Lecture notes
  - Lab and project information
  - Piazza
  - Gradescope for lab/project reports
  - Google form for paper reviews

Course Organization

- Lectures
  - 2 – 3:30pm Mondays and Wednesdays
  - ~8 Guest lectures from industry
  - One paper/week to read and review before lecture.
- Office hours
- 3 Labs (2 weeks / lab)
- Project
- No midterms/final

Course Topics

- Core topics:
  - Deep Neural Networks
  - Quantization
  - Development Platforms
  - Kernel Computation
  - Dataflows
  - State-of-the-art Accelerators
  - Mapping
  - Sparsity
  - Hardware-Software Co-design
  - Other networks: RNN, NLP, RM, RL
  - Advanced Technology
  - Training
  - ML for Hardware Design



Readings

- One required paper to read every week.
- Submit your paper review via Google form by the end of Wednesday (link posted on the course website).

- The review questions guide you through the paper reading process.
  - What are the **motivations** for this work?
  - What is the **proposed solution**?
  - What is the work's **evaluation** of the proposed solution?
  - What is your **analysis** of the identified problem, idea, and evaluation?
  - What are **future directions** for this research?
  - What **questions** are you left with?

Labs and Project

- Three labs with two weeks / lab
  - Lab 1: Quantization
  - Lab 2: Processing Element design (Verilog)
  - Lab 3: Application Mapping
  - Done solo.
- Project:
  - Open-ended research project.
  - We'll provide a range of HW evaluation platforms, including PYNQ, AWS F1, Jetson Nano, and Coral TPU.
  - Done with a partner (recommended).
  - Project demo/poster session RRR week
  - Project report due RRR week

Piazza

- For interactions between faculty, GSIs and fellow students – we are using Piazza
- http://piazza.com/berkeley/spring2021/ee2902

Honor Code

- If you turn in someone else's work as if it were your own, you are guilty of academic misconduct. This includes problem sets, answers on exams, lab exercise checks, project design, and any required course turn-in material.
- Also, if you knowingly aid in academic misconduct, you are guilty.
- We have software that compares your submitted work to others.
- However, it is okay to discuss with others lab exercises and the project (obviously, okay to work with project partner). Okay to discuss homework with others. But everyone must turn in their own work.
- If we catch your academic misconduct, you will get negative points on the assignment: It is better to not do the work than to cheat! If it is a midterm exam, final exam, or final project, you get an F in the class. All cases of academic misconducts reported to the office of student conduct.

Grading Breakdown

- Readings/Paper Review: 10%
  - Read 1 paper and submit 1 review (via Google Form) per week.
- Labs: 40%
  - 3 labs
- Project: 50%
  - Do great research!
- Course participation (e.g., lecture, Piazza, Gemmini/Chipyard code contribution): 5%
  - Extra credit. Get involved in research!
- 7 late days.

Prerequisites

- EECS151/251A or CS152

- What we mean:
  - Basic understanding of computer architecture and digital logic design.
  - Verilog experience is required.
  - Comfortable with programming in C/C++ and Python.

Getting started

- Paper reading starts next week.
- Lab 1 starts next week.
- Register Piazza as soon as possible.
- Register your instructional server account here:
  - inst.eecs.berkeley.edu/webacct
- Pre-semester survey:
  - https://forms.gle/tc2JKYh7r8LqiQf86