AI on Chip
2024 Spring

Chia-Chi Tsai (蔡家齊)
cctsai@gs.ncku.edu.tw
AI System Lab
Department of Electrical Engineering
National Cheng Kung University

Course Information

- Lecture:
  - Chia-Chi Tsai 蔡家齊
    - cctsai@gs.ncku.edu.tw
    - 電機系館5樓 92510
  - Location: 啟端館一樓階梯教室(96112)
  - Tuesday 09:10~12:00

- TAs:
  - TA Group
    - 羅祥睿、湯詠涵、吳柄葳、林泳陞、金稟鈞、洪翊碩、許峻祐、劉子齊
  - Email course.aislab@gmail.com
    - Please include [AOC2024] to the beginning of the email subject

Course Information

- Lecture Note
  - Slides was developed in the reference with
    - Prof. Chung-Ho Chen's, Prof. Sophia Shao's, Prof. Nathan Zhang's Lecture Notes
    - Cs231n
    - Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, Joel Emer, Efficient Processing of Deep Neural Network, Morgan and Claypool Publisher, 2020
- One semester course, which include knowledge of building an AI processor
  - To learn design principles for AI processor
  - To learn SW/HW co-design for AI processor
  - To learn fundamental knowledge of AI architecture
- Prerequisites
  - Logic Design
  - HDL programming

Course Grade

- Final Exams 20%
- Paper Readings and Review 9%
  - 3 paper reviews and 3% for each
- Assignment 30%
  - Lab exercise related to DNN accelerator design
- AI Processor Design Proposal 16%
  - Architecture, dataflow, NoC and mapping optimization design of AI accelerator
- Final Project 25%
  - DNN accelerator implementation

Course Outline

- Overview of DNNs
- Popular DNNs and Applications
- DNN Kernel Computation
- Designing DNN Accelerator
- Operation Mapping
- Reducing Precision
- Exploiting Sparsity
- Advanced Technologies

Course Timetable

Week	Date	Lecture	Assignment	Paper Review
1	2/20	Introduction		
2	2/27	Overview of DNNs	Lab1: CNN model (SW)	Paper Review1 - AI Models
3	3/5	Popular DNNs and Applications		
4	3/12	Kernel Computation	Lab2 Quantization (SW)	
5	3/19	Kernel Computation		Paper Review2 - Quantization
6	3/26	Designing DNN Processors	Lab3 Multiplication (HW)	
7	4/2	Designing DNN Processors	AI Processor Design Proposal	
8	4/9	Designing DNN Processors	Lab4 PE architecture based on Eyeriss accelerator (HW)	Paper Review3 – AI Processor
9	4/16	Designing DNN Processors		
10	4/23	Designing DNN Processors	Lab5 Use row stationary dataflow to implement convolution (HW)	
11	4/30	Operation Mapping	Final Project	
12	5/7	AI Processor Design Proposal Presentation		
13	5/14	AI Processor Design Proposal Presentation	Lab6 Build CNN Accelerator to run AI model (HW)	
14	5/21	Reducing Precision		
15	5/28	Advanced Technologies		
16	6/4	Final Exam	Lab7 AI compiler (SW&HW)	
17	6/11	Final Project Presentation		
18	6/18	Final Project Presentation		

Paper Readings and Review

- Objective
  - To understand fundamental knowledge of AI
  - To learn up-to-date DNNs and hardware architecture

- Requirement
  - Choose **at least one or more** papers
    - From recommended paper list
    - **Or any other paper as long as it related to the topics**
  - Summarize and write paper review in word/latex format
    - **LaTeX format is highly recommended**
  - Hand in **compiled pdf files** on moodle

Paper Readings and Review

- Reading reviews are free of format
- But the following 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?

Programming Assignments

- Assignment Topics
  - Lab1: CNN model (SW)
  - Lab2 Quantization (SW)
  - Lab3 Multiplication (HW)
  - Lab4 PE architecture based on Eyeriss accelerator (HW)
  - Lab5 Use row stationary dataflow to implement convolution (HW)
  - Lab6 Build CNN Accelerator to run AI model (HW)
  - Lab7 AI compiler (SW&HW)
- **Done Solely**

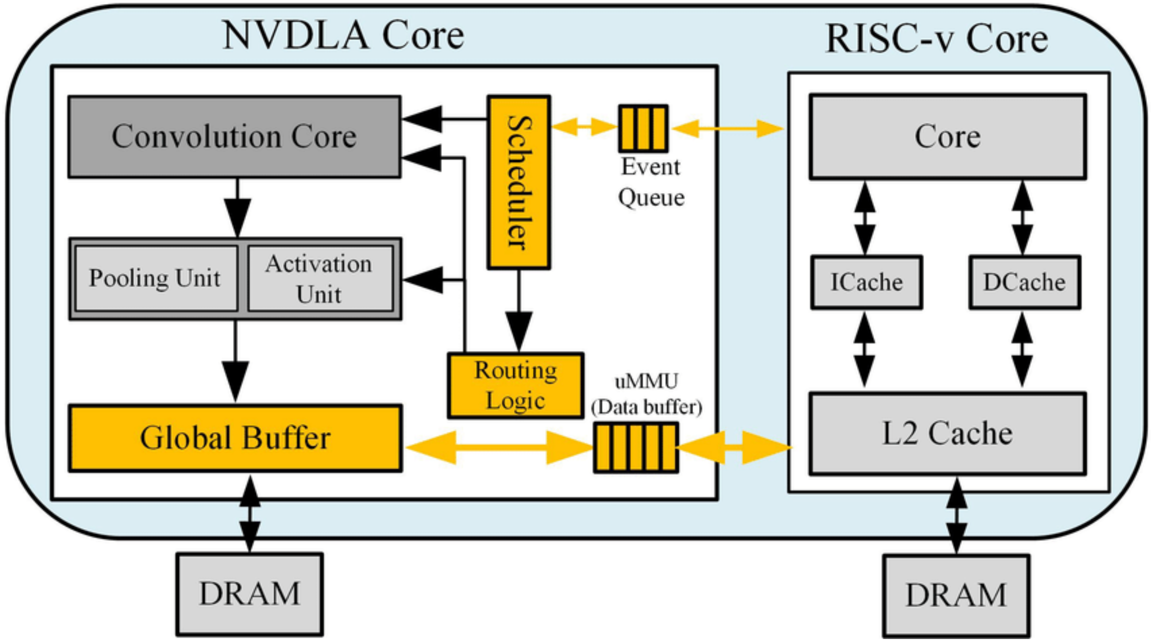
AI Processor Design Proposal

- Design your own AI accelerator
- Innovated design from perspective of
  - Architecture
  - Network on Chip
  - Dataflow
  - Mapping Optimization
  - Co-design
  - Any other novel design
- Detail analytical report of your design is needed
- Implementation free
- **Done with partners**

Final Project

- AI accelerator implementation
  - Realistic design of your own AI accelerator
    - Implement previously proposed design is recommended but not necessary
  - Implement your own accelerator
  - Or Improved from existing designs
  - Co-design with your AI accelerator

- **Done with partners**

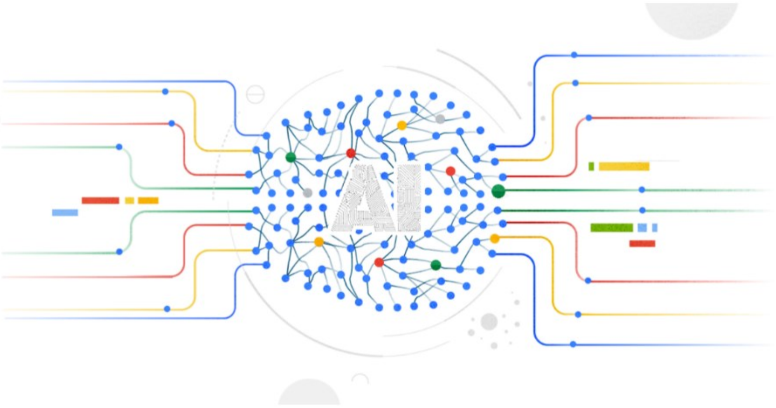


The Virtuous Circle of Deep Learning

Bigger Data

Larger Model

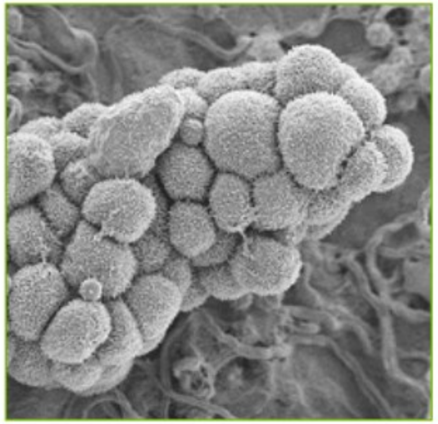
More Compute



Application Domains Scaling

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

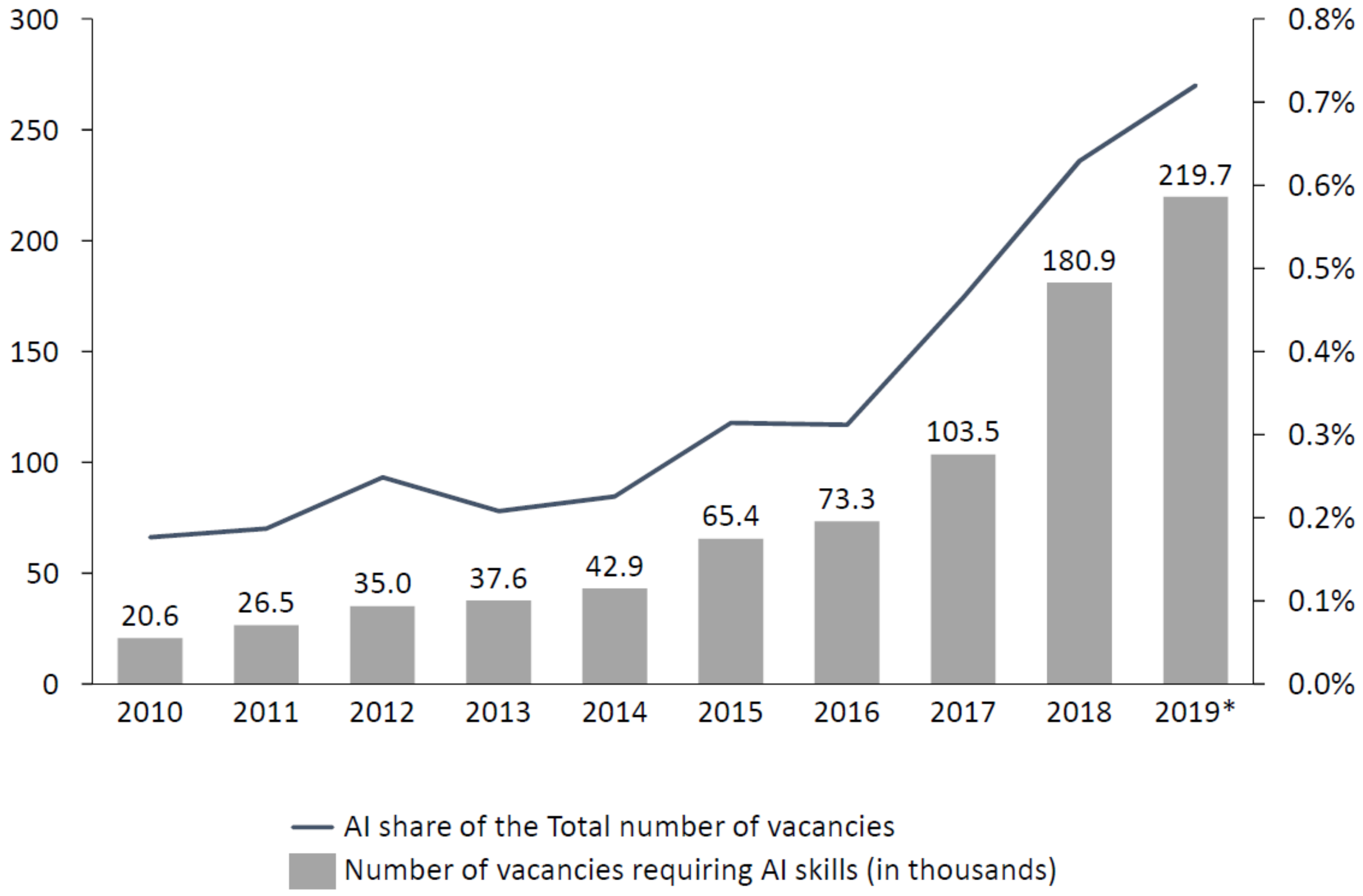
SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

AUTONOMOUS MACHINES

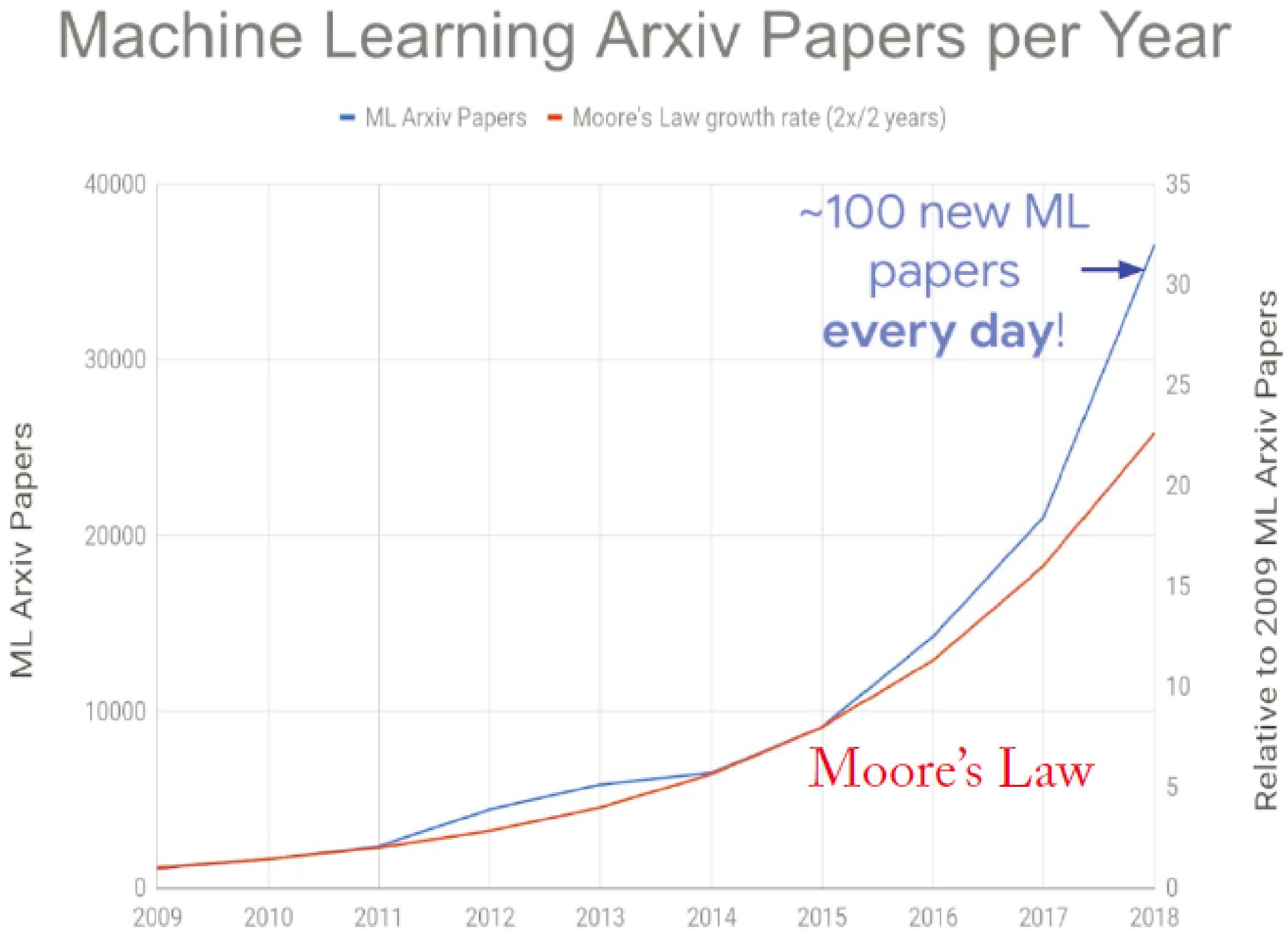
Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

Deep Learning Jobs Scaling



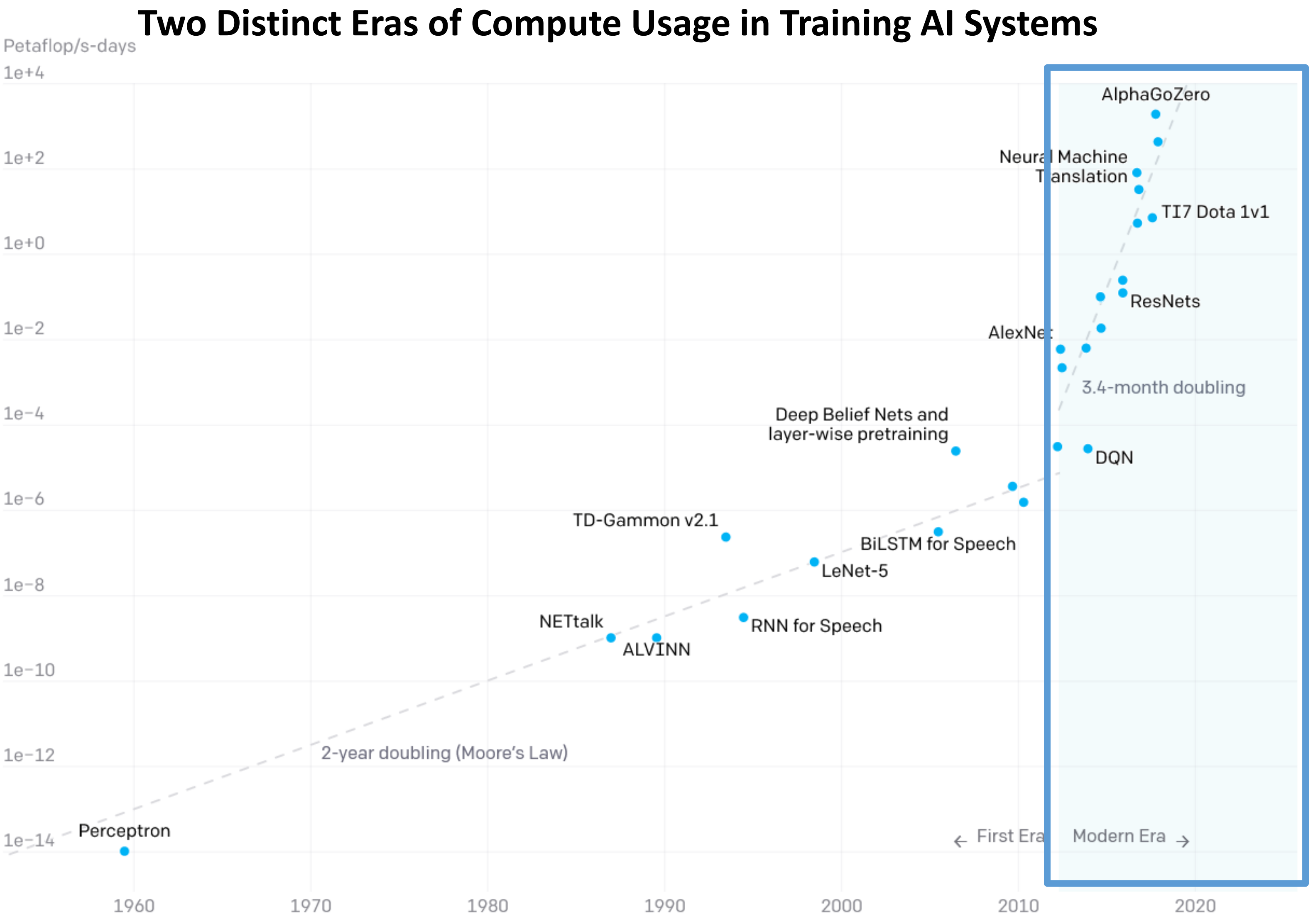
Alekseeva, L., Azar, J., Gine, M., Samila, S., & Taska, B. (2021). The demand for AI skills in the labor market. *Labour Economics*, *71*, 102002.

Deep Learning Papers Scaling



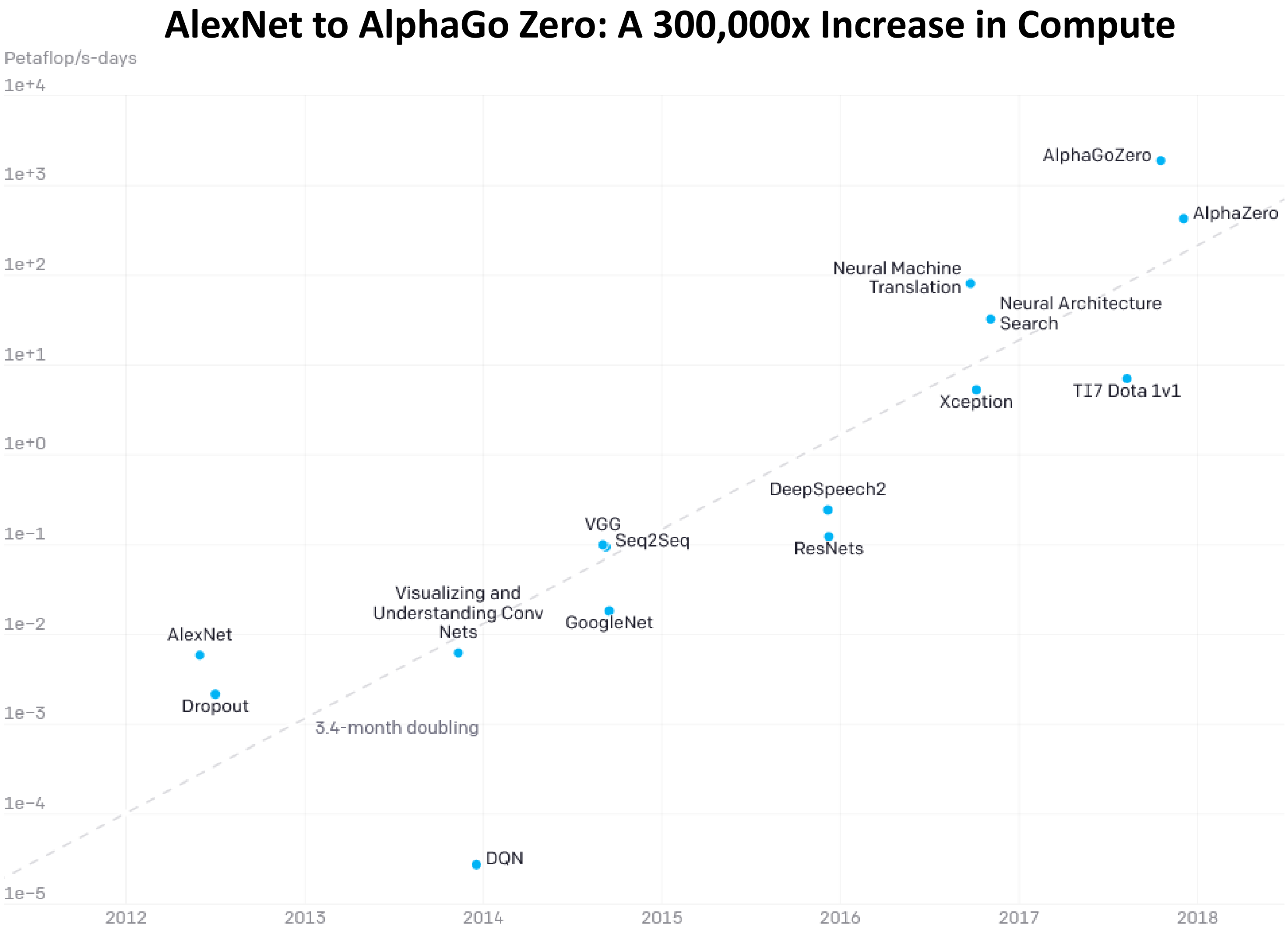
Dean, J., Patterson, D., & Young, C. (2018). A new golden age in computer architecture: Empowering the machine-learning revolution. *IEEE Micro*, *38*(2), 21-29.

Deep Learning Models Scaling



From: OpenAI

Deep Learning Models Scaling



From: OpenAI

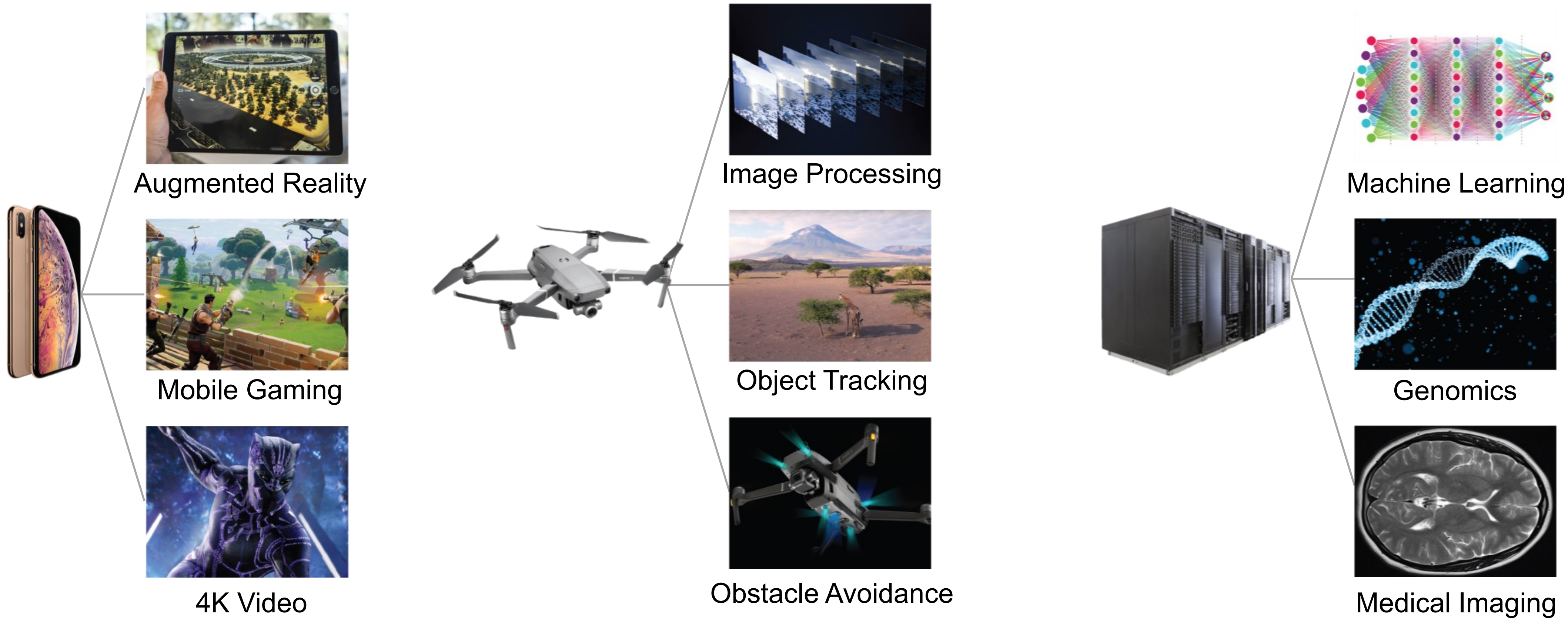
Deep Learning Hardware Scaling

Table 1: 90-epoch training time and single-crop validation accuracy of ResNet-50 for ImageNet reported by different teams.

Team	Hardware	Software	Minibatch size	Time	Accuracy
He *et al.* [5]	Tesla P100 × 8	Caffe	256	29 hr	75.3 %
Goyal *et al.* [4]	Tesla P100 × 256	Caffe2	8,192	1 hr	76.3 %
Codreanu *et al.* [3]	KNL 7250 × 720	Intel Caffe	11,520	62 min	75.0 %
You *et al.* [10]	Xeon 8160 × 1600	Intel Caffe	16,000	31 min	75.3 %
This work	Tesla P100 × 1024	Chainer	32,768	15 min	74.9 %

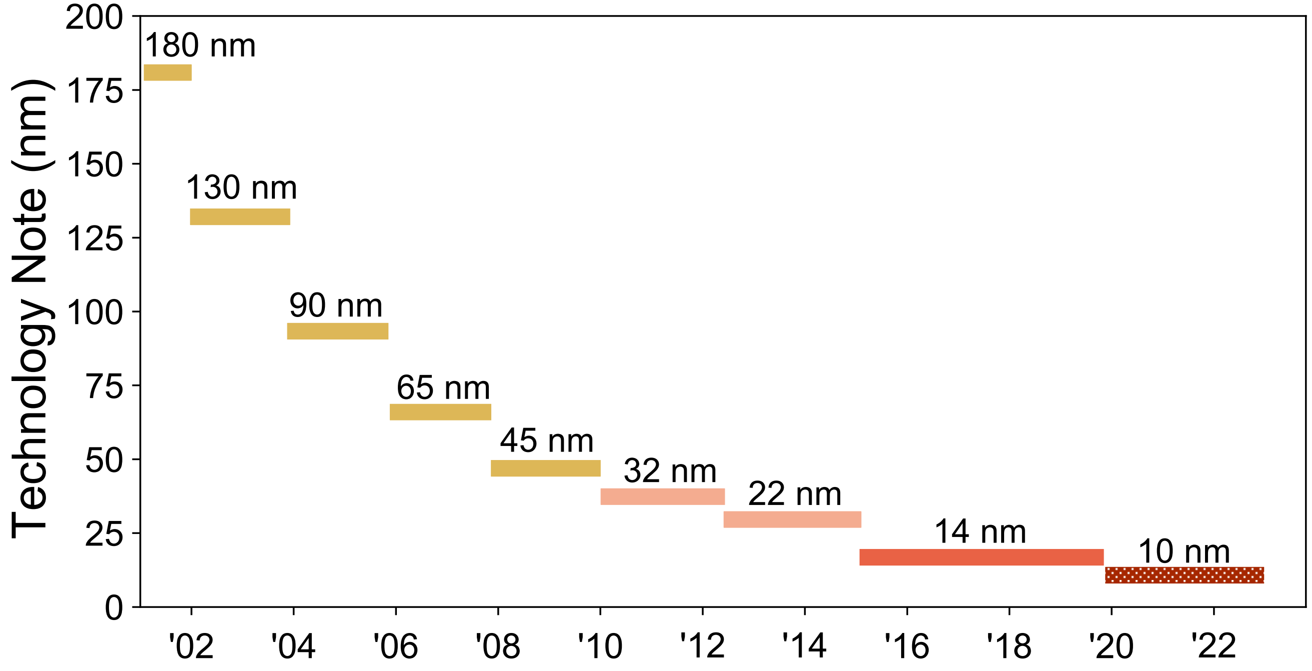
Akiba, T., Suzuki, S., & Fukuda, K. (2017). Extremely large minibatch sgd: Training resnet-50 on imagenet in 15 minutes. *arXiv preprint arXiv:1711.04325*.

Increasing Demand for Computing



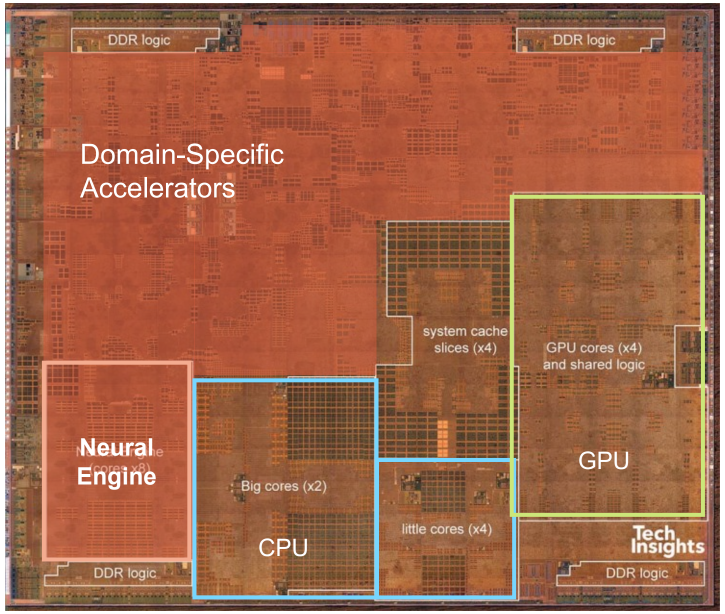
Moore’s Law Won’t Help Us

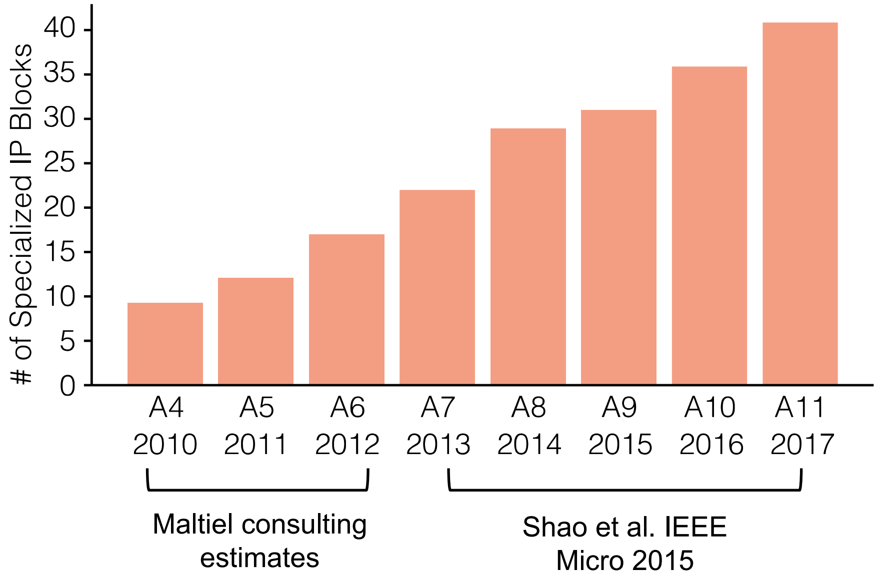
- Increasing transistors is not getting efficient
  - Because of **Slowdown of Moore’s Law and Dennard Scaling**
  - Need **Specialized/Domain-specific accelerators** to improve computing speed and energy



Domain Specific Accelerators

- Customized hardware designed for a domain of applications.





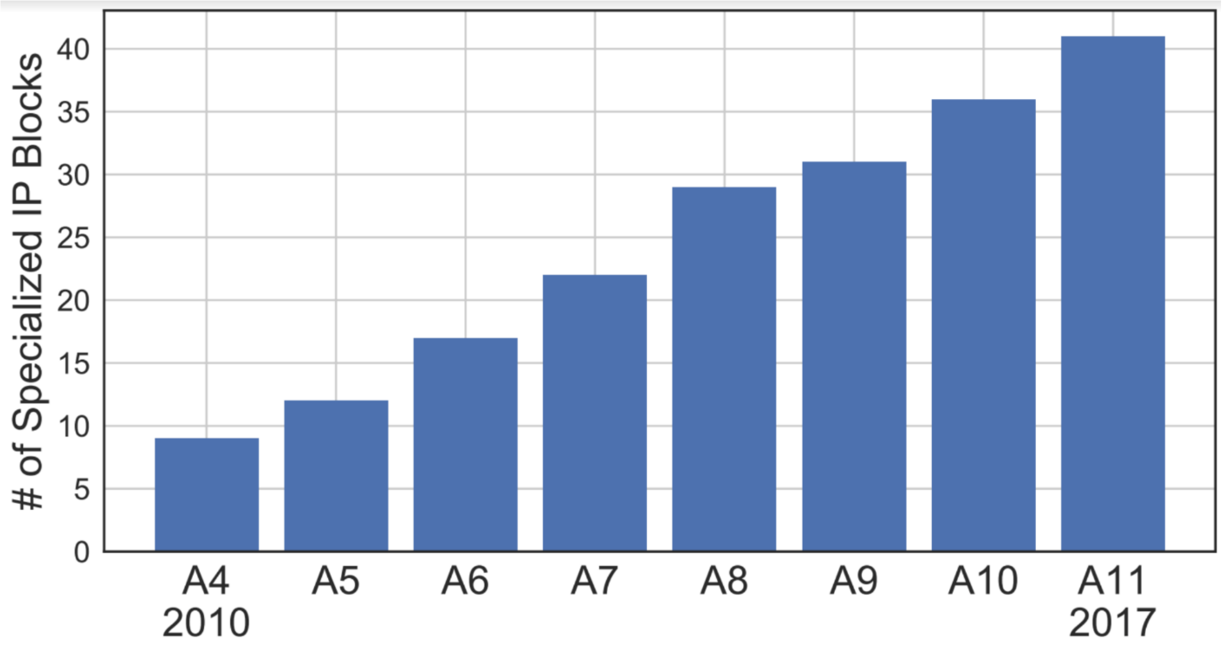
Domain Specific Architecture (DSAs)

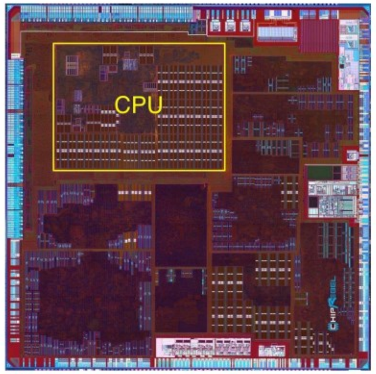
- Achieving higher performance by tailoring characteristics of domain applications to the architecture
  - Need domain-specific knowledge to work out good DSAs
  - Domain Specific Languages (DSLs) + DSAs (not strict ASIC)
  - Specialize to **a domain** of many **applications**
- Examples
  - GPU for computer 3D graphics, virtual reality
  - Neural processing unit (NPU) for machine learning
  - Visual processing unit (VPU) for image processing

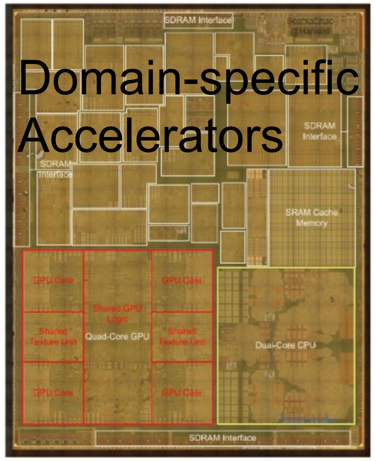
Why DSA

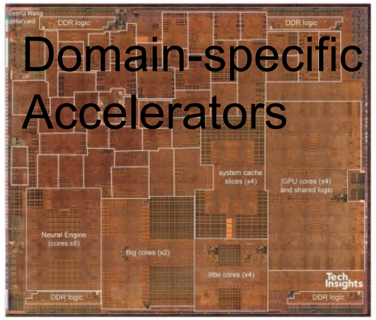
- More effective parallelism for a specific domain
  - SIMD vs. MIMD
  - VLIW vs. Speculative, out-of-order
- More effective use of memory bandwidth
  - User controlled vs. caches
- Eliminate unneeded precision (Quantization)
  - Lower FP/INT data precision (32 bit integers -> 8 bit integers)
- Increase the hardware utilization
  - Reduce the idle time on pipeline and LD/ST

Domain Specific Accelerators in SoCs









2010 Apple A4 65 nm Samsung 53 mm^2 **4 accelerators**	**2014 Apple A8** 20 nm TSMC 89 mm^2 **28 accelerators**	**2019 Apple A12** 7 nm TSMC 83 mm^2 **42 accelerators**

Emma Wang and Sophia Shao, http://vlsiarch.eecs.harvard.edu/accelerators/die-photo-analysis

Domain Specific Languages (DSL)

- DSLs target specific operations on a domain of applications
- Need vector, matrix or sparse matrix operations
- DSLs tailors for these operations
  - OpenGL, TensorFlow, Halide
- Compilers are important if DSLs are architecture-independent
  - Translate, schedule, map ISAs to right DSAs

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.

DL has Reinvigorated Hardware

INTEL ACQUIRES ARTIFICIAL INTELLIGENCE CHIPMAKER HABANA LABS

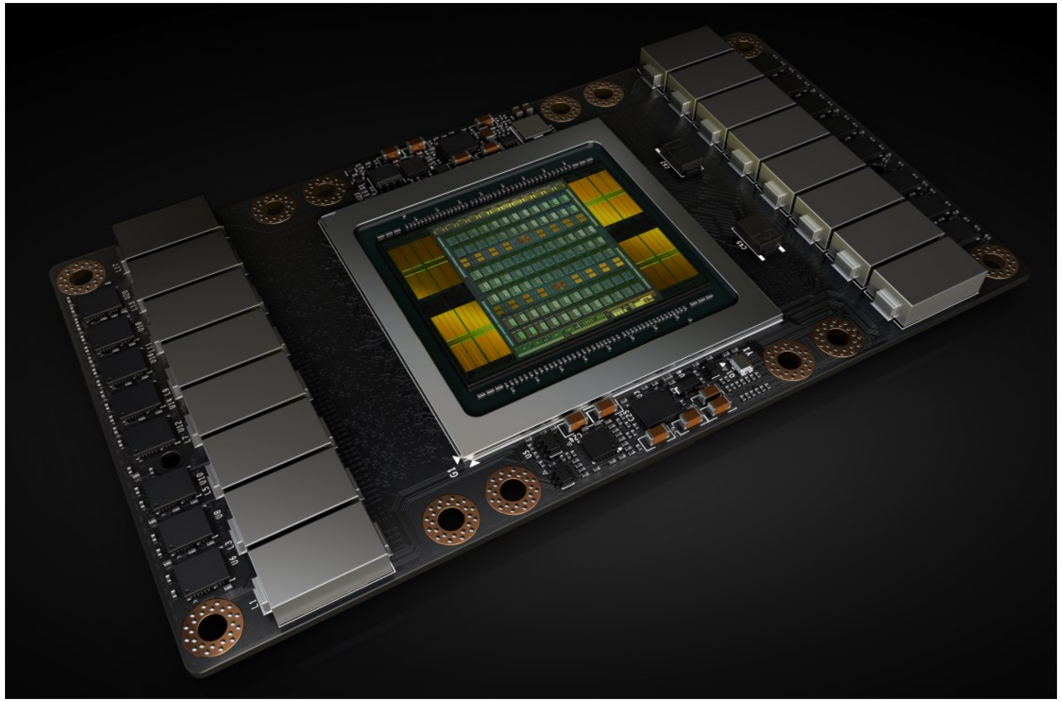
Combination Advances Intel's AI Strategy, Strengthens Portfolio of AI Accelerators for the Data Center

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."

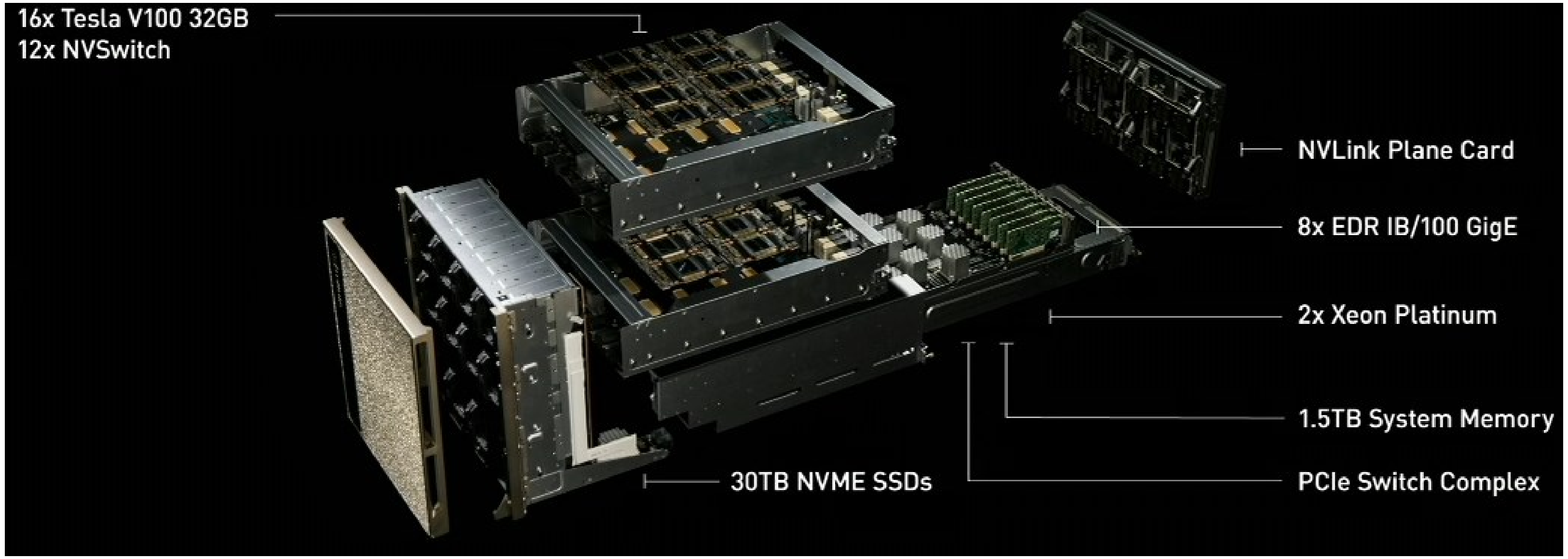
NVIDIA GPU

- 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



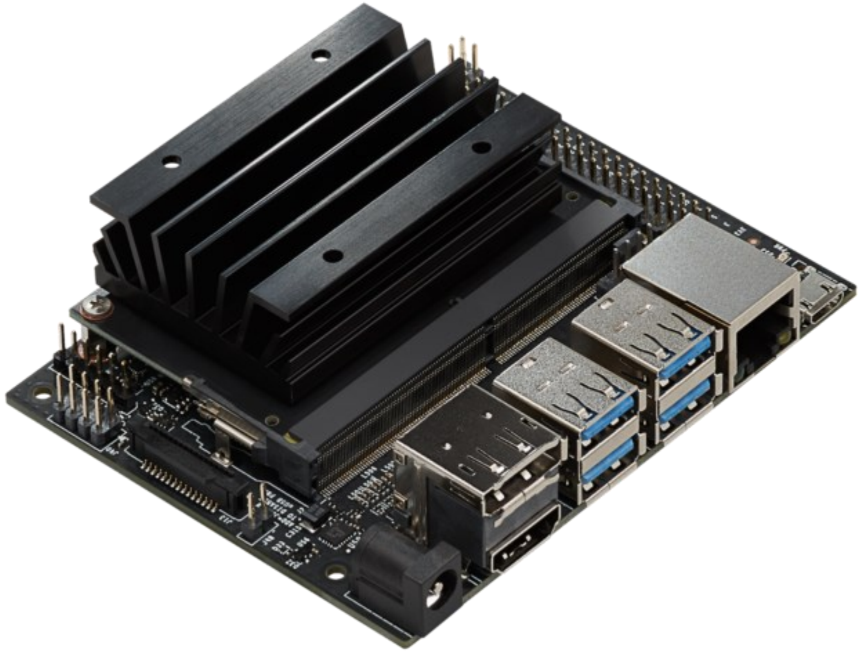
NVIDIA DGX-2

- World's First 2 PetaFlops System
  - 16X Tesla V100
  - Max Power: 10kW
  - $399,000



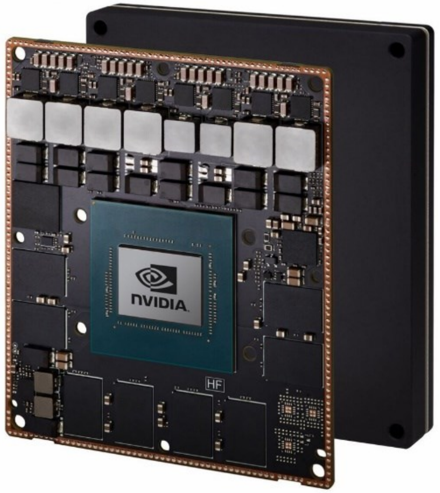
NVIDIA Jetson Nano

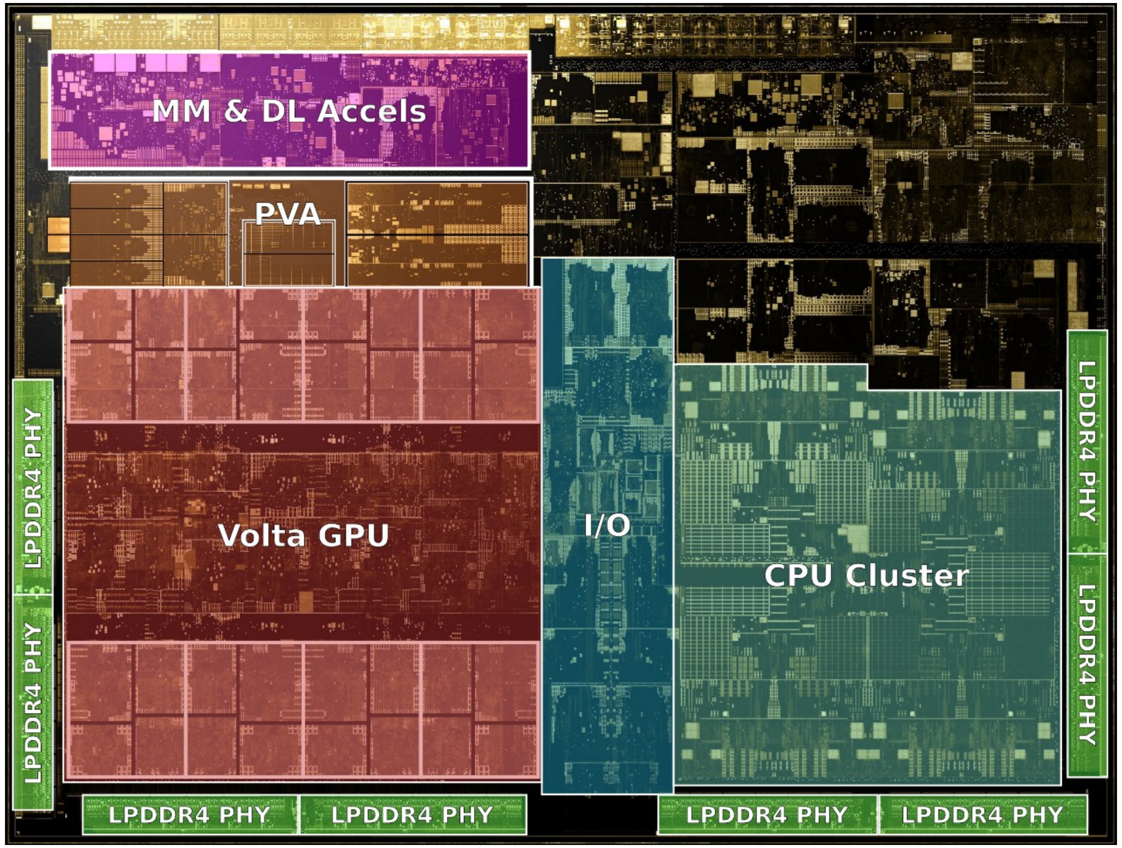
- $99 computer for edge devices
- 472 GFLOPS
  - Quad-core 64-bit ARM CPU
  - 128-core GPU
  - 5W/10W



Nvidia Tegra AGX Xavier GPU

- 32 TOPS
- 512-core Volta GPU with 64 Tensor Cores
- 8-core Carmel ARM v8.2 64-bit CPU
- 32GB 256-Bit LPDDR4x

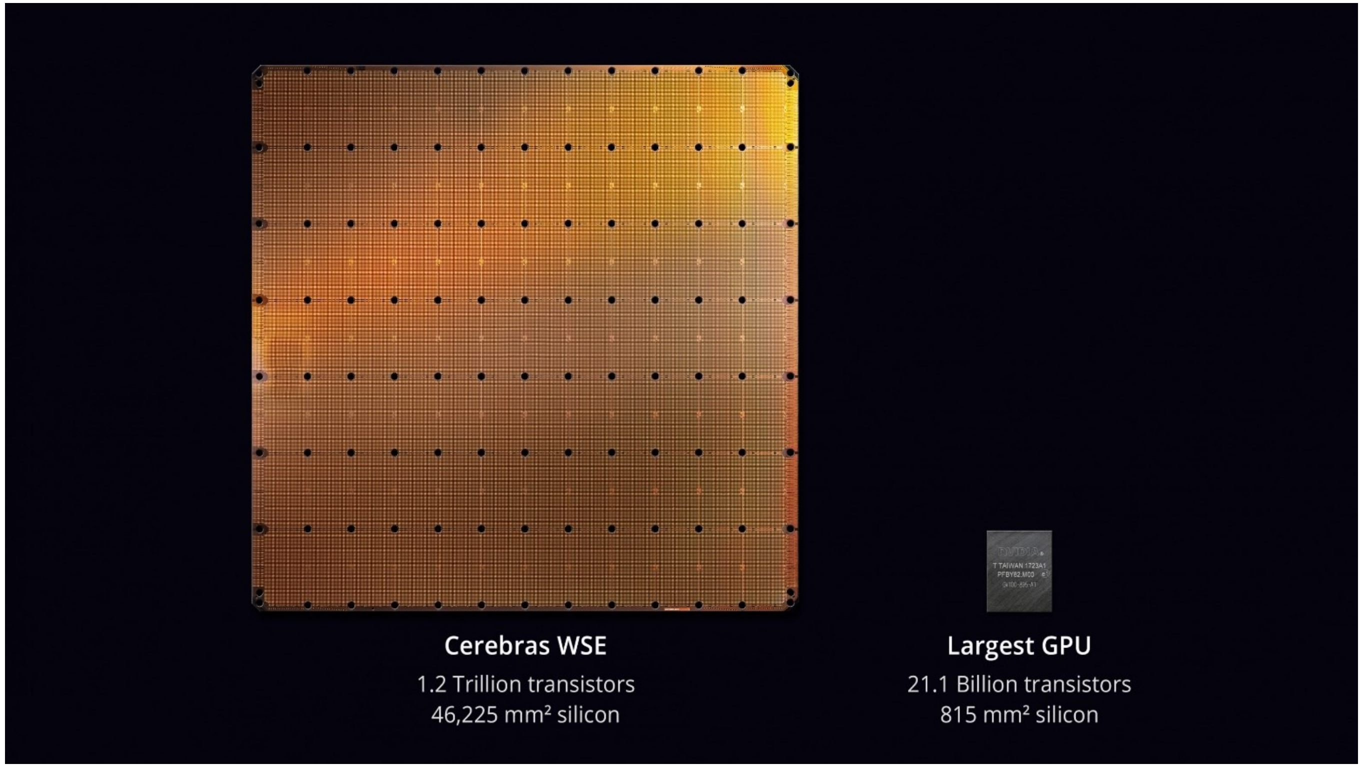




Cerebras: Wafer-Scale Deep Learning

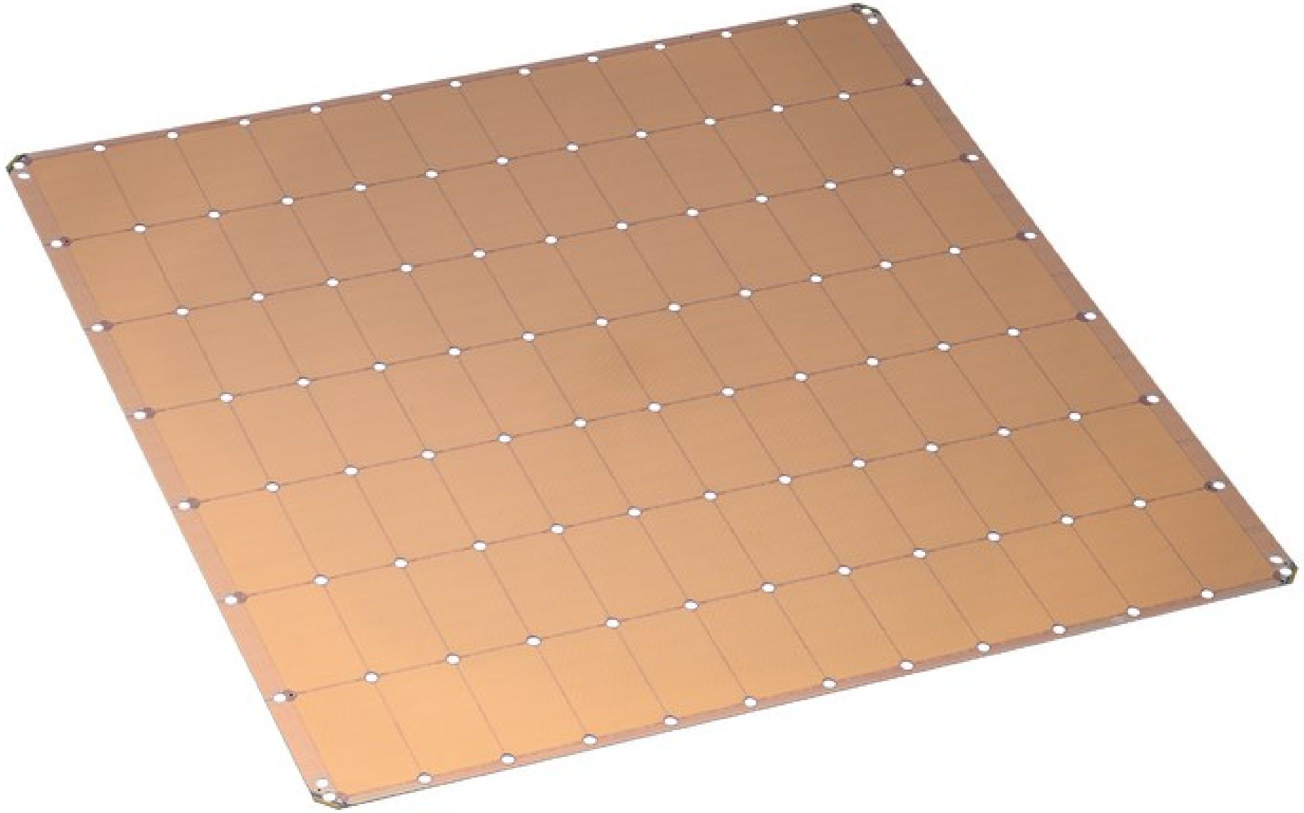
- Largest Chip Ever Built!

- 46,225 mm^2 silicon
- 1.2 trillion transistors
- 400,000 optimized AI cores
- 18 GB of on-chip memory
- TSMC 16nm process



Cerebras Wafer-Scale Engine 2 (WSE-2)

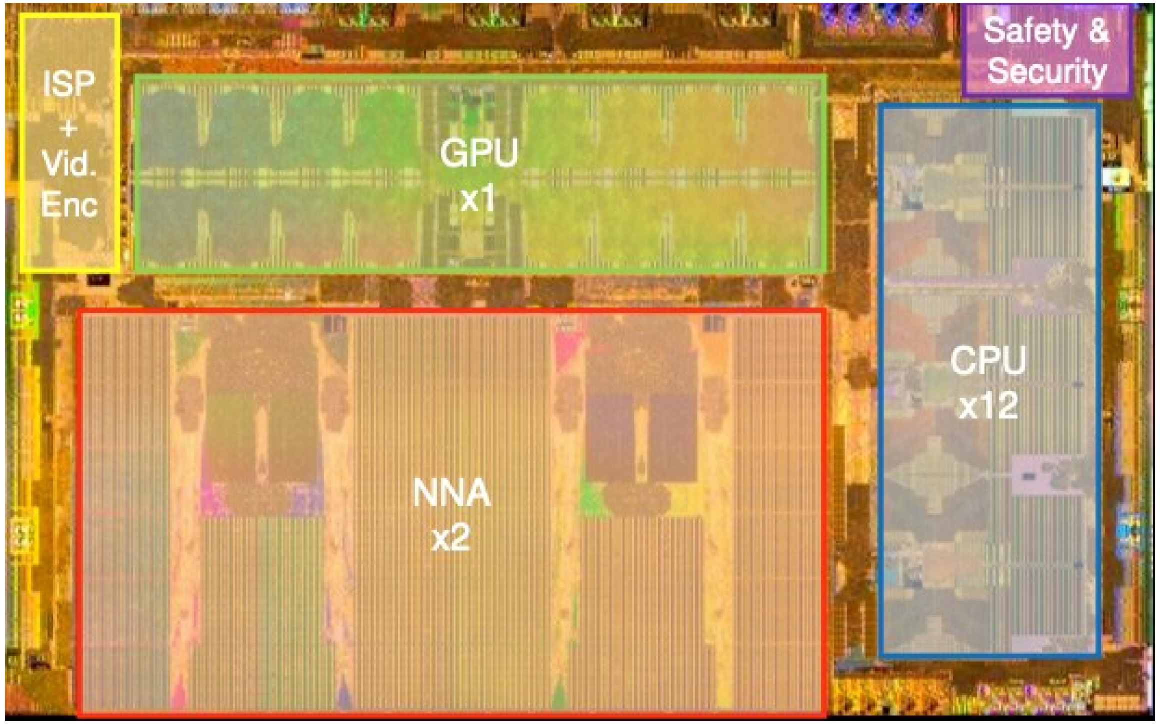
- 46,225 mm^2 silicon
- 2.6 trillion transistors
- 850,000 optimized AI cores
- 40 GB of on-chip memory
- TSMC 7nm process



Power dissipation?
Packaging?
Wafer Yield?
Power supply?

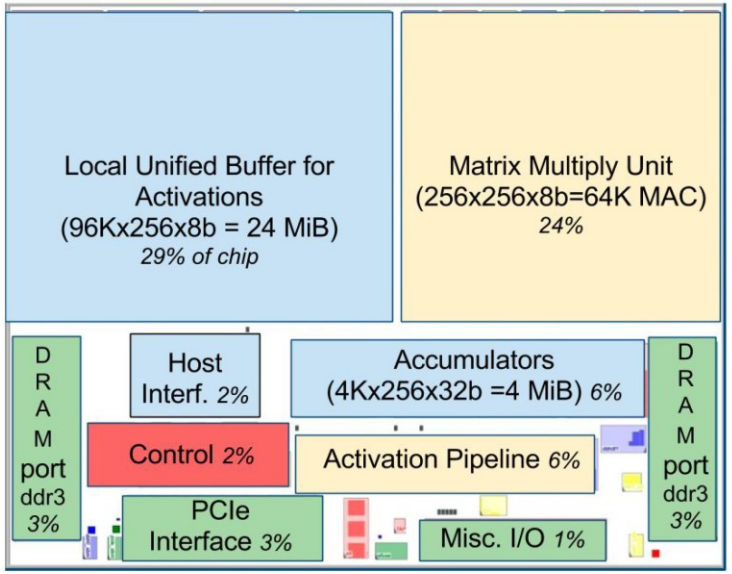
Tesla: Full-Self-Driving Computer

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

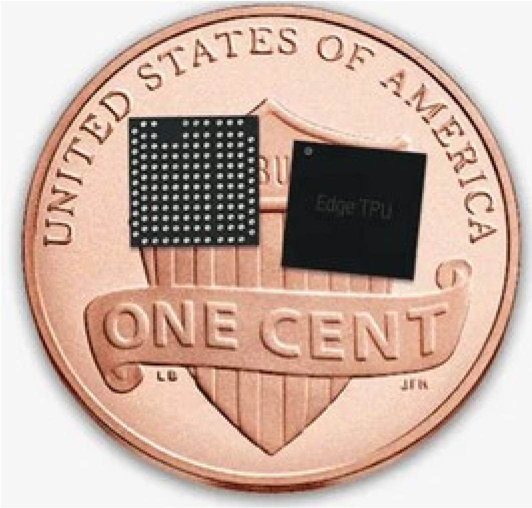


Google TPU

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



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



MLPerf Benchmark

- DL benchmarks for DL models with different DSLs
- Address three challenges
  - The diversity of models
  - The variety of deployment scenarios
  - The array of inference systems

MLPerf

Fair and useful benchmarks for measuring training and inference performance of ML hardware, software, and services.

MLPerf Inference	**Machine learning tasks**
Cloud/ Datacenter	Image classification, object detection, image segmentation, speech, language processing, recommendation, reinforcement learning
Edge	Image classification, object detection, image segmentation, speech, language processing
Mobile	Image classification, object detection, image segmentation, language processing
Tiny	Image classification, Object detection, Anomaly detection, Speech
MLPerf training	
Cloud/ Datacenter	Image classification, object detection, image segmentation, natural language processing, recommendation, reinforcement learning
HPC	Climate segmentation, cosmology parameter prediction

Build Your Own AI Accelerator System

- Build better **algorithms**
- Build better **runtimes**
- Build better **hardware**
- **Mapping Optimization** from DNN operations to hardware