Accelerator Architectures for Machine Learning (AAML)

Lecture 1: Basics of AI Accelerator

Tsung Tai Yeh
Department of Computer Science
National Yang-Ming Chiao Tung University

Acknowledgements and Disclaimer

- Slides was developed in the reference with
  Joel Emer, Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, ISCA 2019 tutorial
  Efficient Processing of Deep Neural Network, Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, Joel Emer, Morgan and Claypool Publisher, 2020
  Yakun Sophia Shao, EE290-2: Hardware for Machine Learning, UC Berkeley, 2020
  CS231n Convolutional Neural Networks for Visual Recognition, Stanford University, 2020

Outline

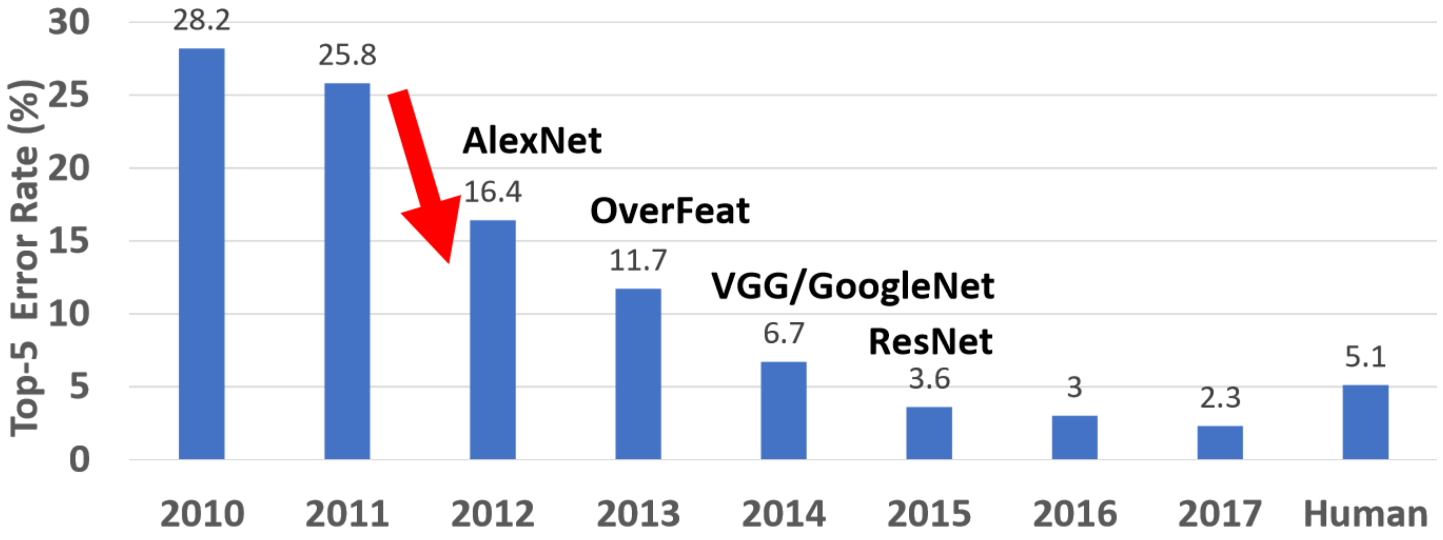
- Dennard Scaling vs Dark Silicon
- Artificial Neural Network (ANN)
- Spiking Neural Network (SNN)
- Neuromorphic architectures
- Digital vs Analog Accelerators

Why do we need accelerators ?

- Previously
  - We focused on designing general-purpose processors
- Why do accelerators have become attractive in recent years?
  - **Dennard Scaling** has ended
    - Dennard Scaling allowed voltage to shrink with transistor size
    - Without Dennard Scaling, we need to find other ways to keep **power** in check
  - **Dark Silicon**
    - Not turn on all transistors on the chip
    - The success of application’s accelerators (encryption, compression …)
    - Applications only use subset of processors/accelerators at a time, such a heterogeneous architecture meets dark silicon phenomenon

Why Deep Neural Network become popular?

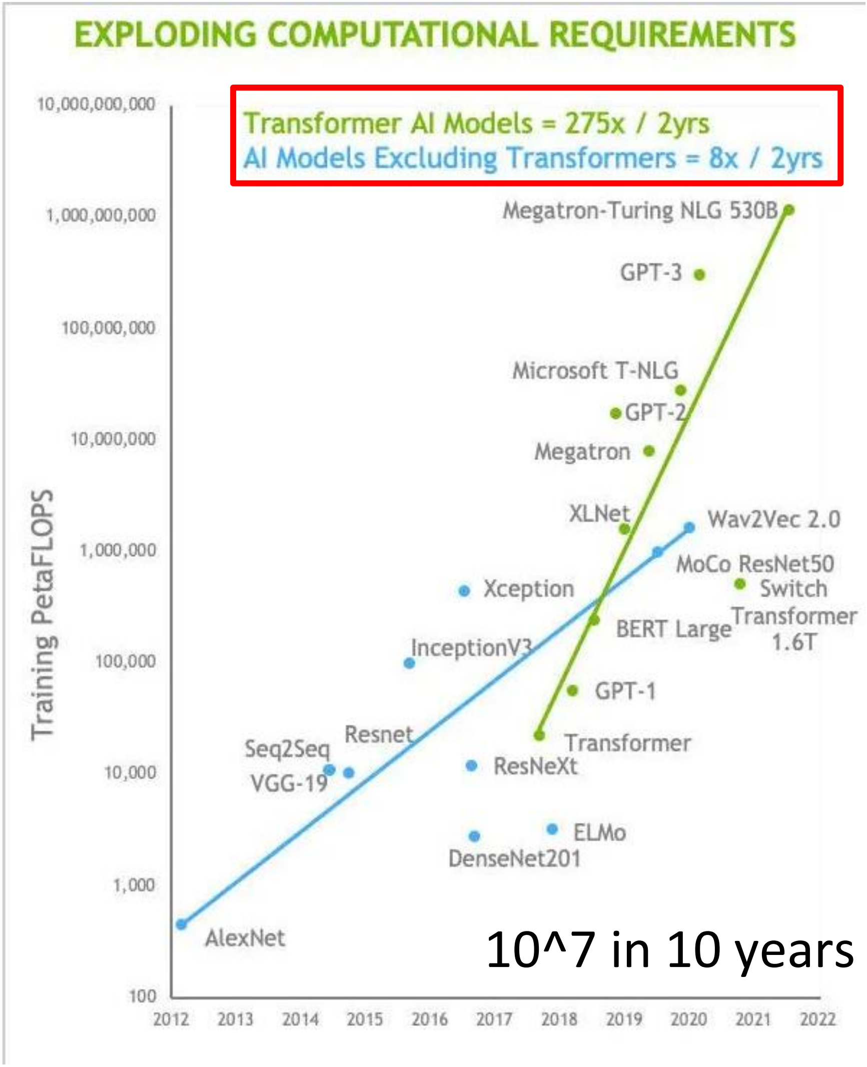
- DNN model outperforms human-being on the ImageNet Challenge



https://arxiv.org/ftp/arxiv/papers/1911/1911.05289.pdf

Computations of DNN

- **Deep Neural Network is getting large**
  - Large model parameters
    - Palm-E (540B)
    - GPT-MoE (1.8T)
  - Why do we need such a large model?



10^7 in 10 years

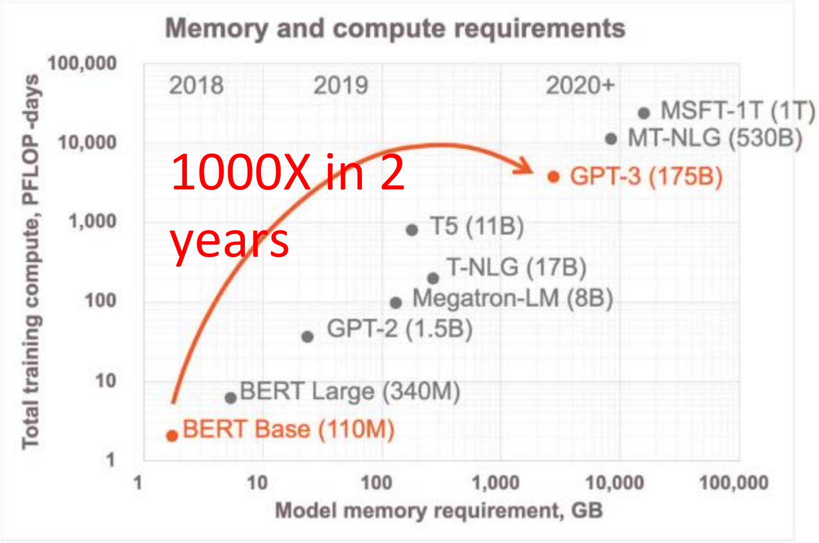
https://blogs.nvidia.com/blog/what-is-a-transformer-model/

Large Language Model

- **A language model is a mathematical mapping**
  - Text -> embedding vector (“representation”)
  - Embedding vectors encode meaning of texts i.e. dog [1,0,0]
  - Train such a model via **next-word prediction** on a large corpus of text data as the **lossy compression**
    - 1000B text token -> 30B model parameters
  - Empirically, LLMs behaves as human as the model size increases

Unsustainable ML Model Growth

- We need a better way to grow models more efficiently
- Get the advantages of larger models but with substantially less compute and memory resources



1000X in 2 years

Sparsity might be one of answer

https://www.cerebras.net/blog/accelerating-llm-training-with-variable-sparse-pre-training-and-dense-fine-tuning/

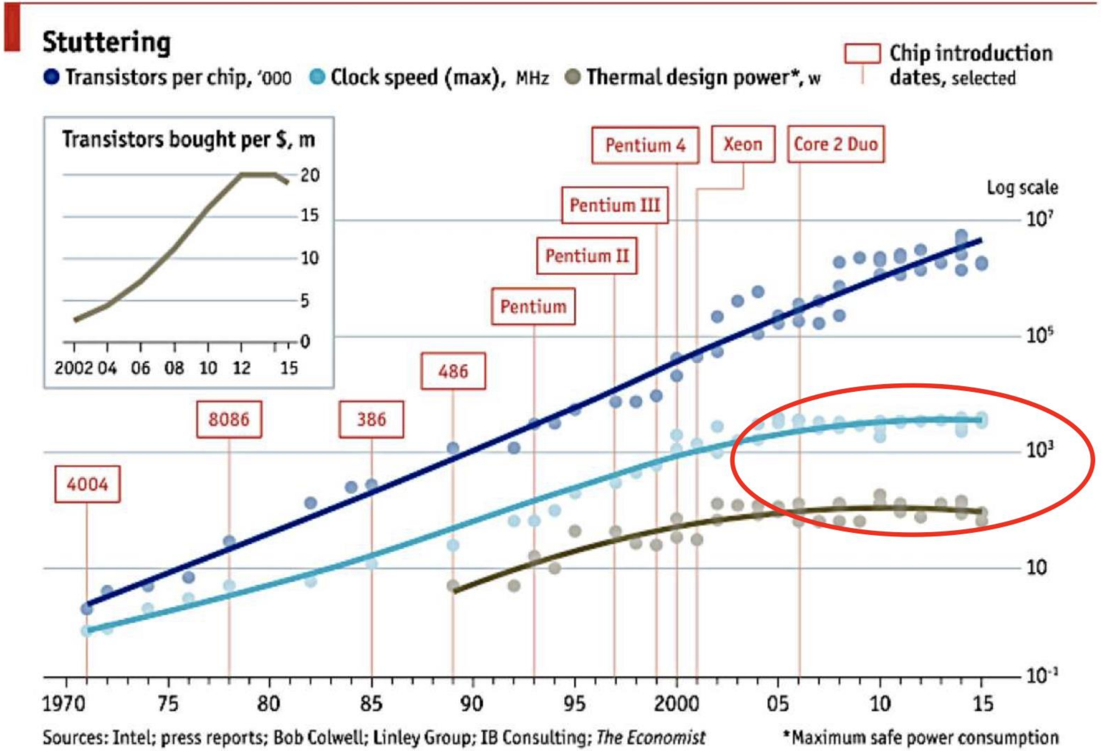
Hardware trends

- **Stagnant single and multi-thread performance on general-purpose cores**
- **What do accelerators matter?**
  - Dark silicon (emphasis on power-efficient throughput)
  - End of scaling
- **Emergence of machine learning**
  - Facilitate the pervasive of hardware acceleration as machine learning emerges as a solution for “everything”.

Commercial Hardware for Machine Learning

- Google TPU (inference and training)
- Nvidia Tensor/transformer cores (Ampere, Hopper)
- Microsoft Brainwave and Catapult
- Intel Loihi NPU
- Cambricon
- Graphcore (training)
- Cerebras (Training)
- Tesla (FSD, Dojo)
- …

Increasing transistors is not getting efficient



Sources: Intel; press reports; Bob Colwell; Linley Group; IB Consulting; *The Economist*

*Maximum safe power consumption

General purpose processor is not getting faster and power-efficient because of **Slowdown of Moore's Law and Dennard Scaling**

Dennard Scaling

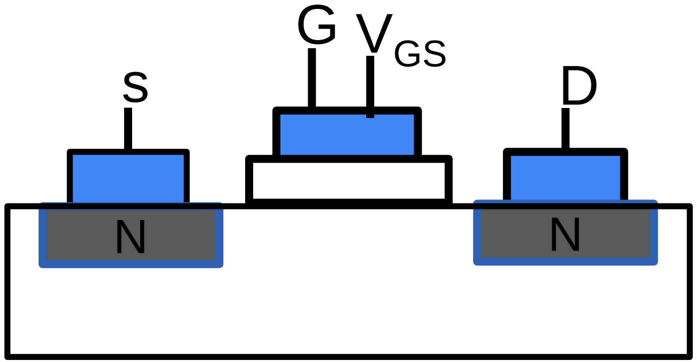
- Dennard scaling allowed voltage to shrink with transistor size
  - E.g. 180 nm -> 1.8 V, 130 nm -> 1.3 V
  - All 4 cores (45 nm) can be worked in full speed
  - Could all 8 cores (28 nm) be worked in full speed, too ? Why ?

Power = alpha x CFV^2
alpha: percent time switched
C: capacitance
F: Frequency
V: Voltage

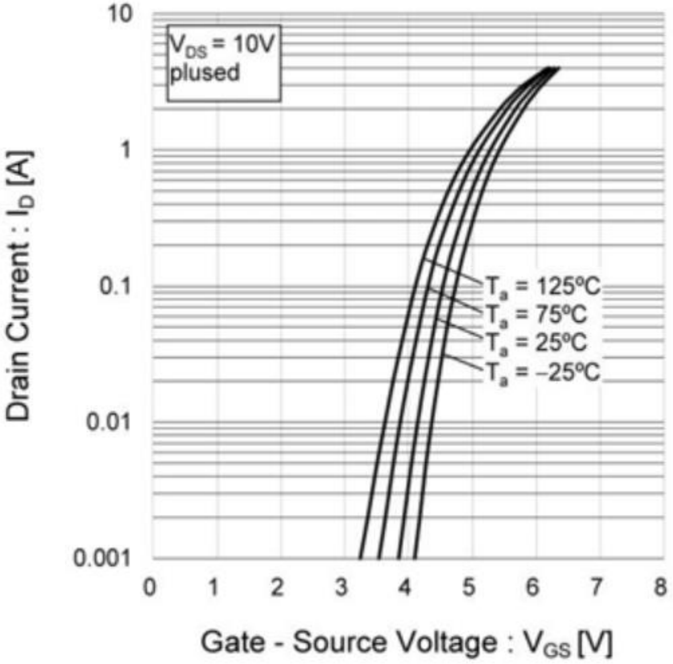
1. Typically, the transistor size reduces K (~1.4) times
2. In the same chip area, **the number of transistor** increases **K^2** times, the **frequency** increases **K** times
3. **The size of capacitance** shrinks **K** times as the reduction of transistor size, and **the voltage** reduces **K^2** times
4. So, we can boost performance of the chip without any compensation of the power

Voltage threshold of MOSFET

- Temperature affects the value of V_{GS} and I_D
  - Ta = 25 d, I_D = 1A and Ta = 75 d, I_D = 1.5 A when fixing V_{GS}
  - Due to $V_{GS\ (TH)}$ constraint, difficult to keep reducing voltage to be proportional to the transistor size below 28 nm



Nch MOSFET



https://techweb.rohm.com.tw/knowledge/si/s-si/03-s-si/5277

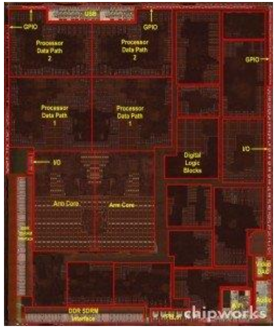
What can we do ?

Power = alpha x CFV^2
alpha: percent time switched
C: capacitance
F: Frequency
V: Voltage

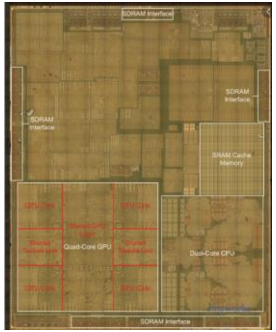
- **Dark silicon**
  - Below 28 nm, the voltage (V) is hard to be changed
  - K^2 = (transistor size as capacitance size) K x frequency (K)
  - The power increases K^2 times
  - Therefore, **not turn on all transistor on the chip**
  - What is the percentage of inactive transistors ?
  - 20 nm: 33%, 16 nm: 45%, 10 nm: 56%, 7 nm: 75%, 5 nm: 80%
- **Dim silicon**
  - Turn all transistor on at low clock speeds

Heterogeneous SoC

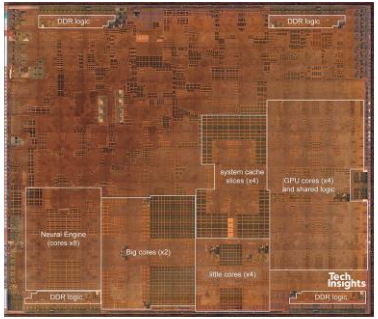
- **Post-Moore era and dark silicon**
  - A suite of accelerators on chip are rising
  - Applications will only use a subset of processors/accelerators at a time
  - Such a heterogeneous architecture is compatible with dark silicon



2010 Apple A4
65 nm TSMC 53 mm^2
4 accelerators



2014 Apple A8
20 nm TSMC 89 mm^2
28 accelerators

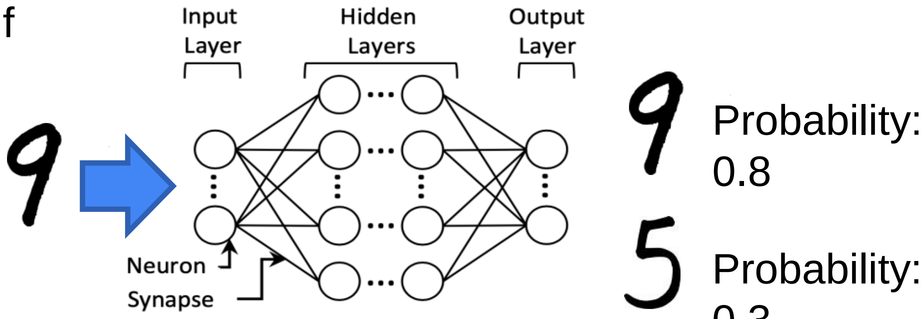


2019 Apple A12
7 nm TSMC 83 mm^2
42 accelerators

https://edge.seas.harvard.edu/files/edge/files/alp.pdf

Artificial Neural Network (ANN)

- **Most machine learning algorithms**
  - Perceptron or artificial neuron
  - Receiving synchronous inputs, and performs math, then produce outputs
  - Measuring the **“strength” (z)** of weighted inputs
  - **Z = x1 * w1 + x2 * w2** where (x is the input of the neuron, w is the weight (determined by training))
  - **Activation function** a = f(z) to decide if a neuron should fire or not
  - **Training performs back-propagation** with gradient descent

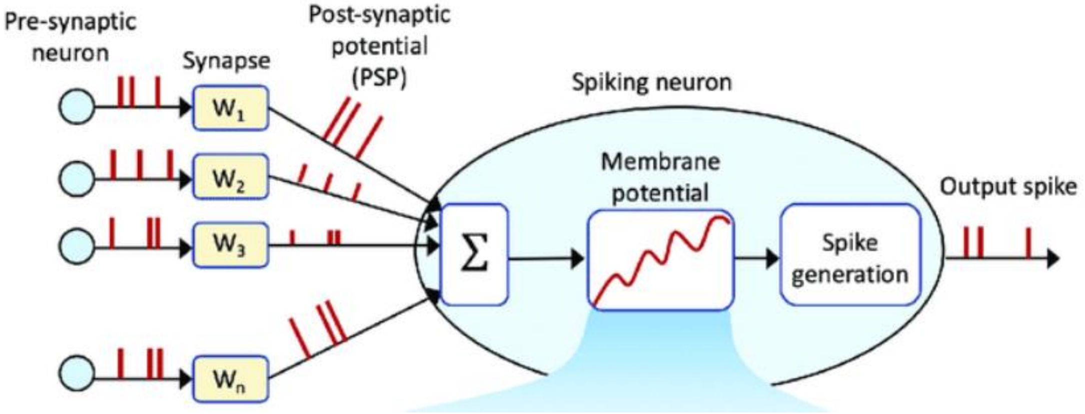


Multilayer perceptrons (MLPs)

https://arxiv.org/pdf/2005.01467.pdf

Spiking Neural Network (SNN)

- Spiking neurons resembles chemical reactions in our brains
  - **A neuron has a certain potential** that represents inputs received
  - The potential **rises and falls** depending on the relative importance of those inputs and leaks away when no receiving inputs
  - When the potential reaches a **threshold**, **the neuron fires**
  - All inputs/outputs are in the form of **binary spikes**



https://www.researchgate.net/figure/A-Schematic-of-a-spiking-neural-network-consisting-of-an-array-of-plastic-synapses_fig1_342414706

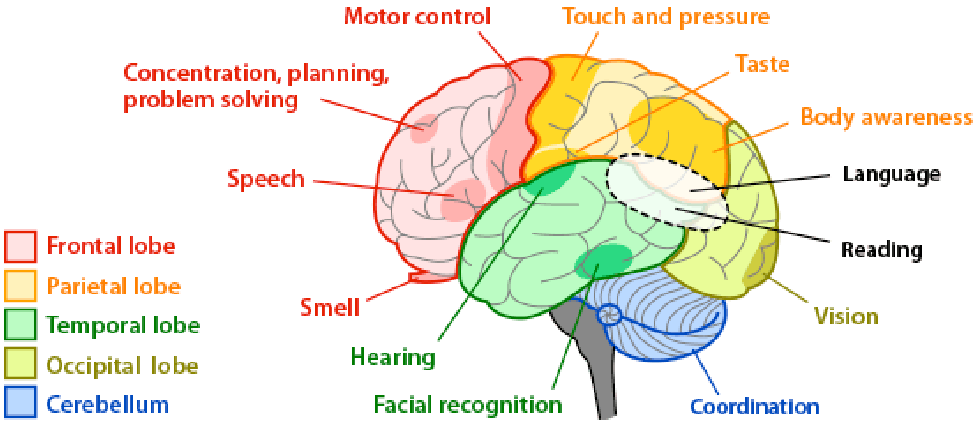
ANN vs. SNN

- **ANN**
  - Perceptron, 8-bit or 16-bit multiplications, complex activation functions
  - High accuracy, supervised learning (inference and training)
- **SNN**
  - Don’t achieve very high accuracies, not well understood
  - A neuron has state that is a more powerful construct for applications that have a notion of time, e.g. video and language analysis
  - Carry a large amount of information in a few bits
  - Unsupervised learning

Uncover Your Brain

2400 kcal/24 hr = 100 kcal/hr = 27.8 cal/sec = 116.38 J/s = 116 W
20% x 116 W = 23.3 W

- The computer as a brain that comprises **specialized accelerators**
- **Low power** – the brain consumes only about 20W
- **Fault tolerant** – the brain loses neurons all the time



Yang, Eric. Think Dinner. *Mac Evolution, 1998*

https://askabiologist.asu.edu/sites/default/files/resources/articles/nervous_journey/brain-regions-areas.gif

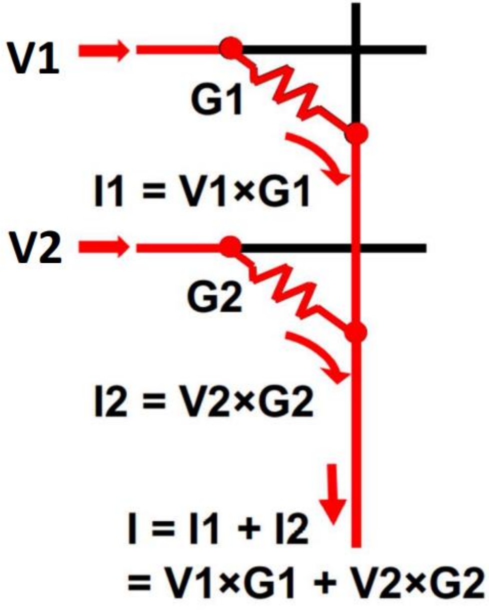
Neuromorphic architectures

- Architectures inspired by neuron behavior
- **Two major flavors**
  - Artificial Neural Network (ANN)
    - Operations on perceptrons
  - Spiking Neural Network (SNN)
    - Mimic operations in the brain
- **Two major implementation styles**
  - Digital
  - Analog

Neuromorphic Hardware

- **Emulating the human brain**
  - Low power – the brain consumes only 20 W
  - Fault tolerant – the brain loses neurons all the time
  - No programming required – the brain learns by itself
- **Examples:**
  - SpiNNaker, Spikey, TrueNorth

Digital vs. Analog

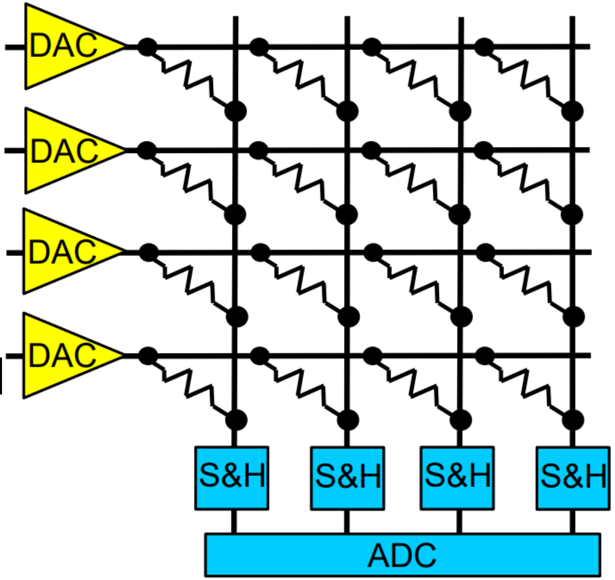


ISAAC, ISCA 2016

- **A single analog device**
  - Perform multiple multi-bit operations
  - Analog has **challenges, e.g., noise/precision**
  - The **current** in a wire or the **charge** in a capacitor represent **a rational number**
  - Perform **addition** by merging the currents in two wires
  - **Multiplication** can be represented by the current that emerges when a voltage is applied to a conductor
  - **Instability** as temperature changes, currents change
- **Digital device**
  - Use CMOS transistors and gates, exclusively deal with 0s and 1s

Crossbar for vector-matrix multiplication

- A grid of **resistances** and horizontal and vertical **wires**
    - The **input voltages** are provided on the horizontal wires (**wordlines**)
    - Each **column** represents a different **neuron**
    - Each column computes a different **dot-product based on conductances** in that col
    - Analog current is sent through an analog-to-digital converter (**ADC**).
    - **S&H** is the sample-and-hold circuit that feeds signals sequentially to the ADC



ISAAC, ISCA 2016

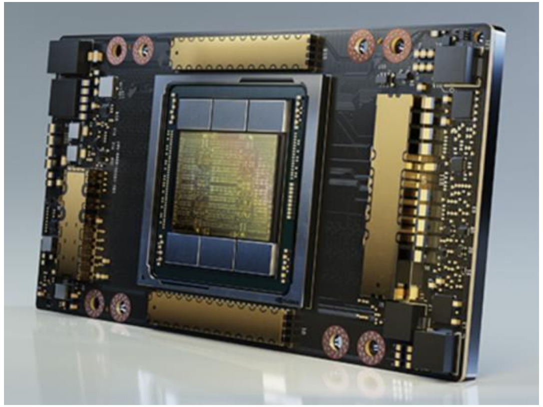
Challenges of analog devices

- **High ADC/DAC area/energy**
  - Long stay in analog needs expensive **analog buffering**, introduces **significant noise** that accumulates across network layers
  - Some ADC overheads increase exponentially with resolution
    - The number of bits coming out of a bitline is a function of **the bits of info in the voltage (v)**
    - **The bits of info in the weight (w)**
    - **The number of rows (R)** being added
  - To increase the parallelism and storage density – high v, w, and R
    - Demanding an expensive high-resolution ADC
  - SNN is amenable to analog, why ?

Digital (I) GPU

	Nvidia V100 GPU (2019)	**Nvidia A100 GPU (2020)**
Transistor count	21 billion	54 billion
FP32 performance	15.7 TFLOP/s	19.5 TFLOP/s
Tensor FP32	125 TFLOP/s	156 TFLOP/s
TDP	300 W	250 W
Die size	815 mm^2	862 mm^2
	TSMC 12 nm	TSMC 7 nm

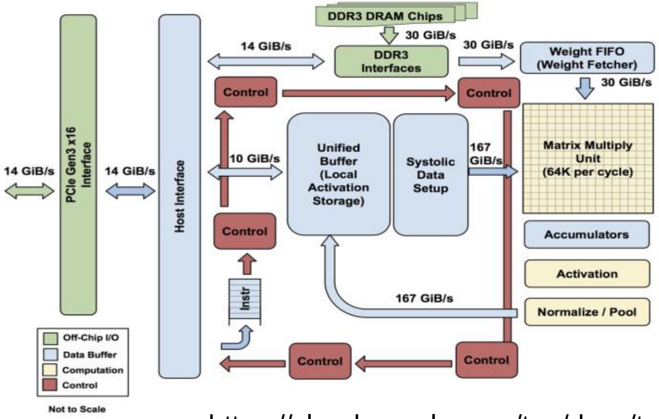
2.57 X
1.24 X
1.25 X



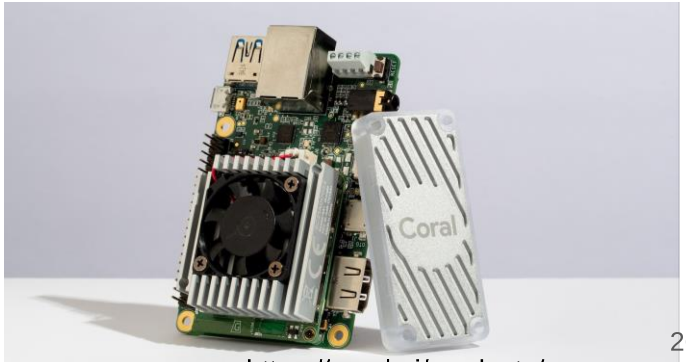
https://www.nvidia.com/en-us/data-center/a100/

Digital (II) Google Tensor Processing Unit (TPU)

- Systolic-array accelerator
  - V1: Inference only
  - V2: Training with bfloat
  - V3: 2X powerful than v2
- Edge TPU
  - Coral Dev Board
  - 4 TOPS
  - 2 TOPS/Watt
  - Support TensorFlow Lite



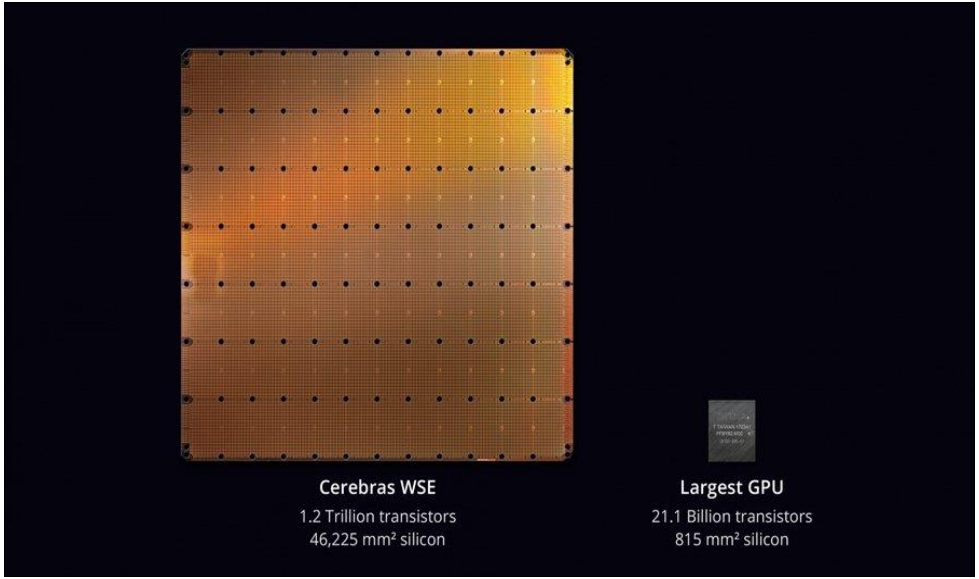
https://cloud.google.com/tpu/docs/tpus



https://coral.ai/products/

Digital (III) Cerebras: Wafer-Scale DL Engine

- Largest DL Chip Ever Built!!
- 46225 mm^2 (WoW !!)
- 1.2 trillion transistor
- 400,000 optimized AI cores
- 18 GB on-chip memory
- TSMC 16 nm process



https://twitter.com/CerebrasSystems/status/1163443985714753537

In summary

- **Learning from History**
  - Neural network (NN) booms, but fades away when it ceases to be fashionable -> support vector machines (SVM) took over
  - General-purpose processors and GPU quickly outpace ASICs
- **Today**
  - NNs > SVM
  - GPPs and GPUs will stagnate in performance, but ML is hot
  - ML accelerators (hardware + ML software perspective) include many implementation operations
  - Neuroscience + emerging technology

Takeaway Questions

- What does dark silicon tell us ?
  - (A) We should turn all transistor on at low clock speeds
  - (B) We cannot turn on all transistors on a chip
  - (C) Allowed voltage to shrink with transistor size
- Why does SNN have the potential for low-energy computations and communication ?
  - (A) Skipping connections
  - (B) Complex SNN computation
  - (C) Not involve in multiplications or complex activation functions

Takeaway Questions

- What are the challenges of analog accelerators ?
  - (A) High ADC/DAC area and energy
  - (B) Limited parallelism
  - (C)Non-programmable