CS5120 VLSI System Design, Spring 2025

Introduction

黃稚存

Chih-Tsun Huang

cthuang@cs.nthu.edu.tw

國立清華大學
NATIONAL TSING HUA UNIVERSITY

資訊工程學系
Computer Science

Lecture 01

聲明

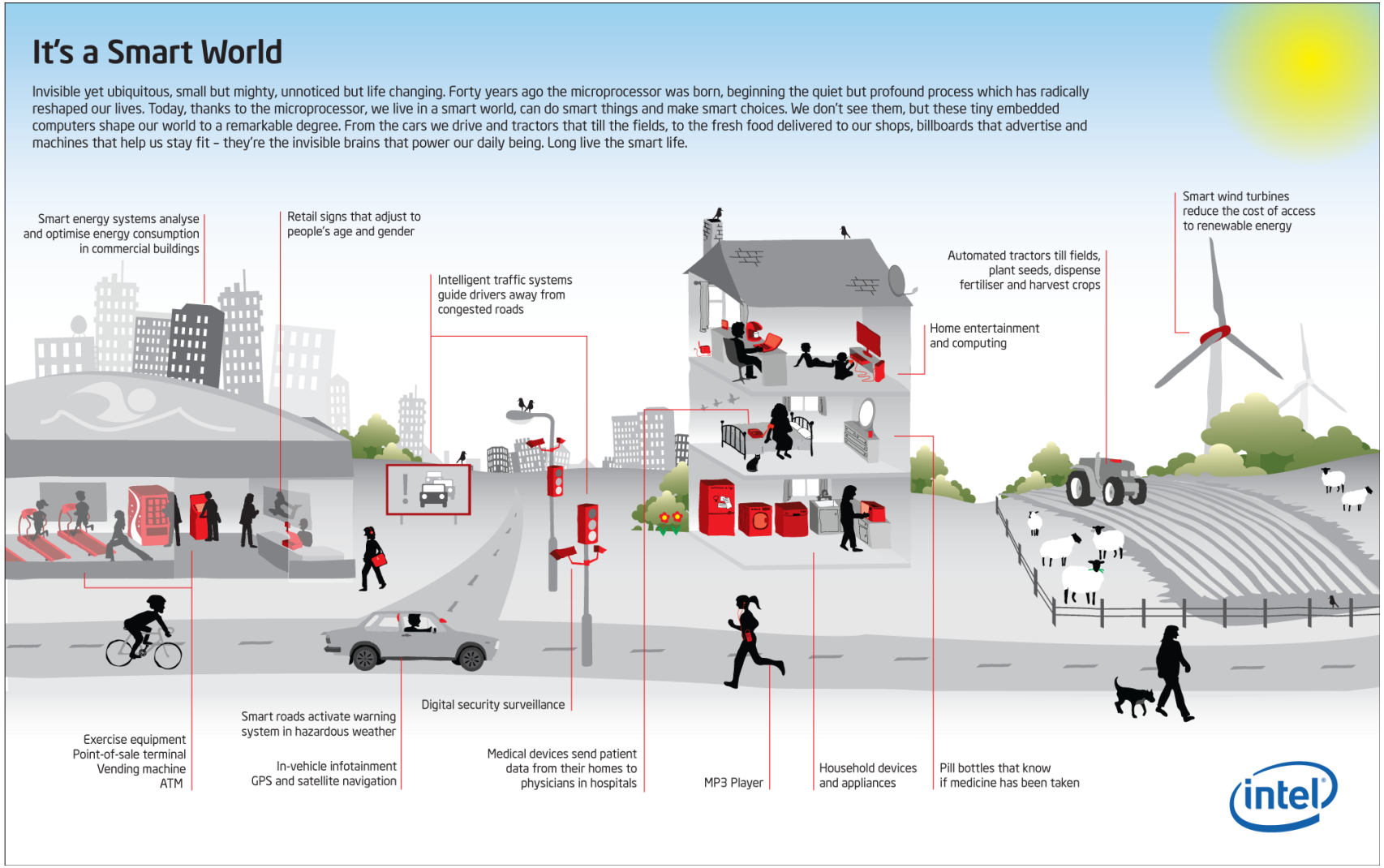
- 本課程之內容 (包括但不限於教材、影片、圖片、檔案資料等)，僅供修課學生個人合理使用，非經授課教師同意，不得以任何形式轉載、重製、散布、公開播送、出版或發行本影片內容 (例如將課程內容放置公開平台上，如 Facebook, Instagram, YouTube, Twitter, Google Drive, Dropbox 等等)。如有侵權行為，需自負法律責任。

The Future of Computing Is Exascale

- Scaling application domains
  - Security
  - Scientific discovery
  - Economic & manufacturing
  - Internet & cloud
  - Healthcare & biology
- New architecture and computing paradigms
  - Advanced process nodes
  - Heterogeneous computing with accelerators
  - Innovative memory hierarchy

Smart World: Ubiquitous Computing

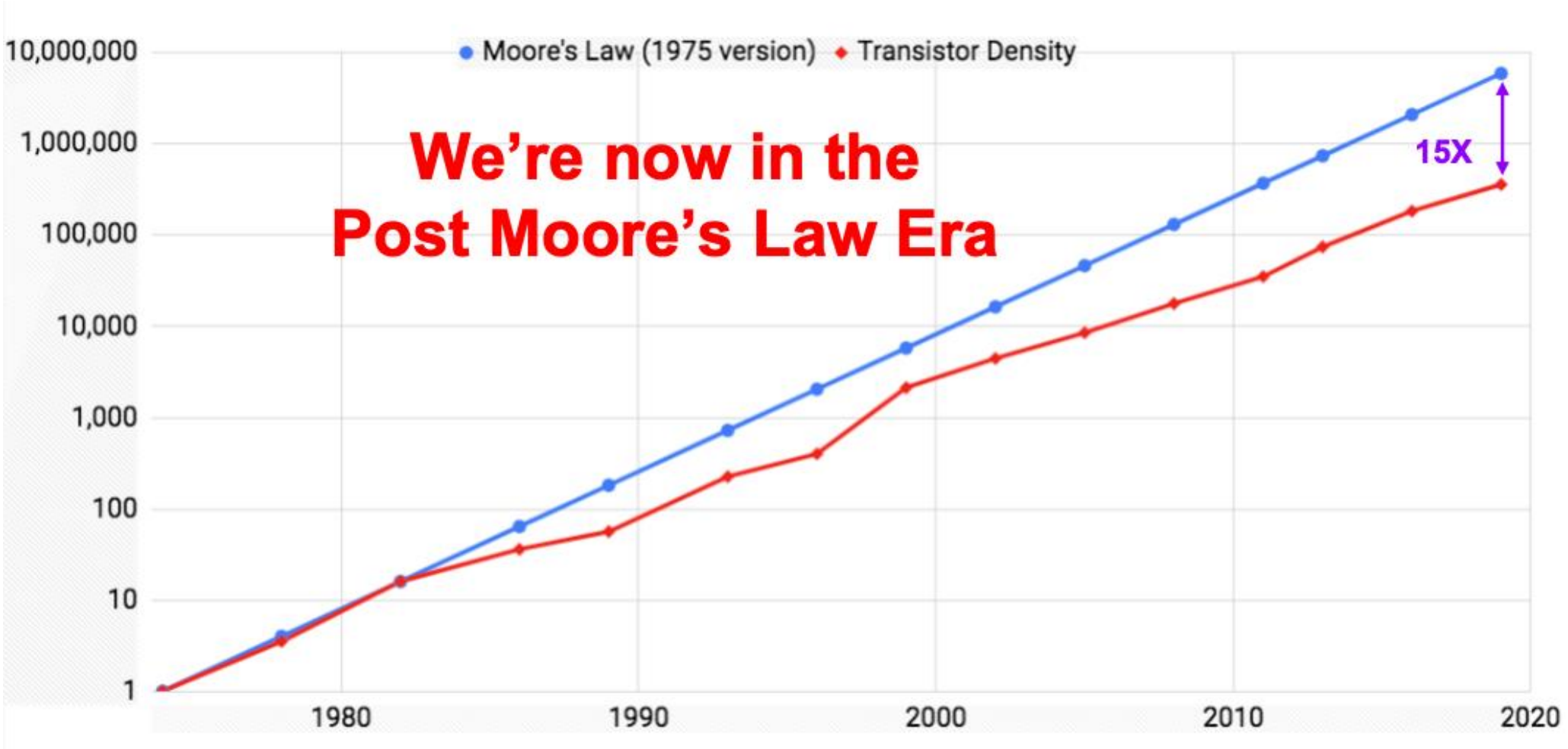
- Diversified, customized computing nodes
  - Cloud
  - Edge
  - End devices



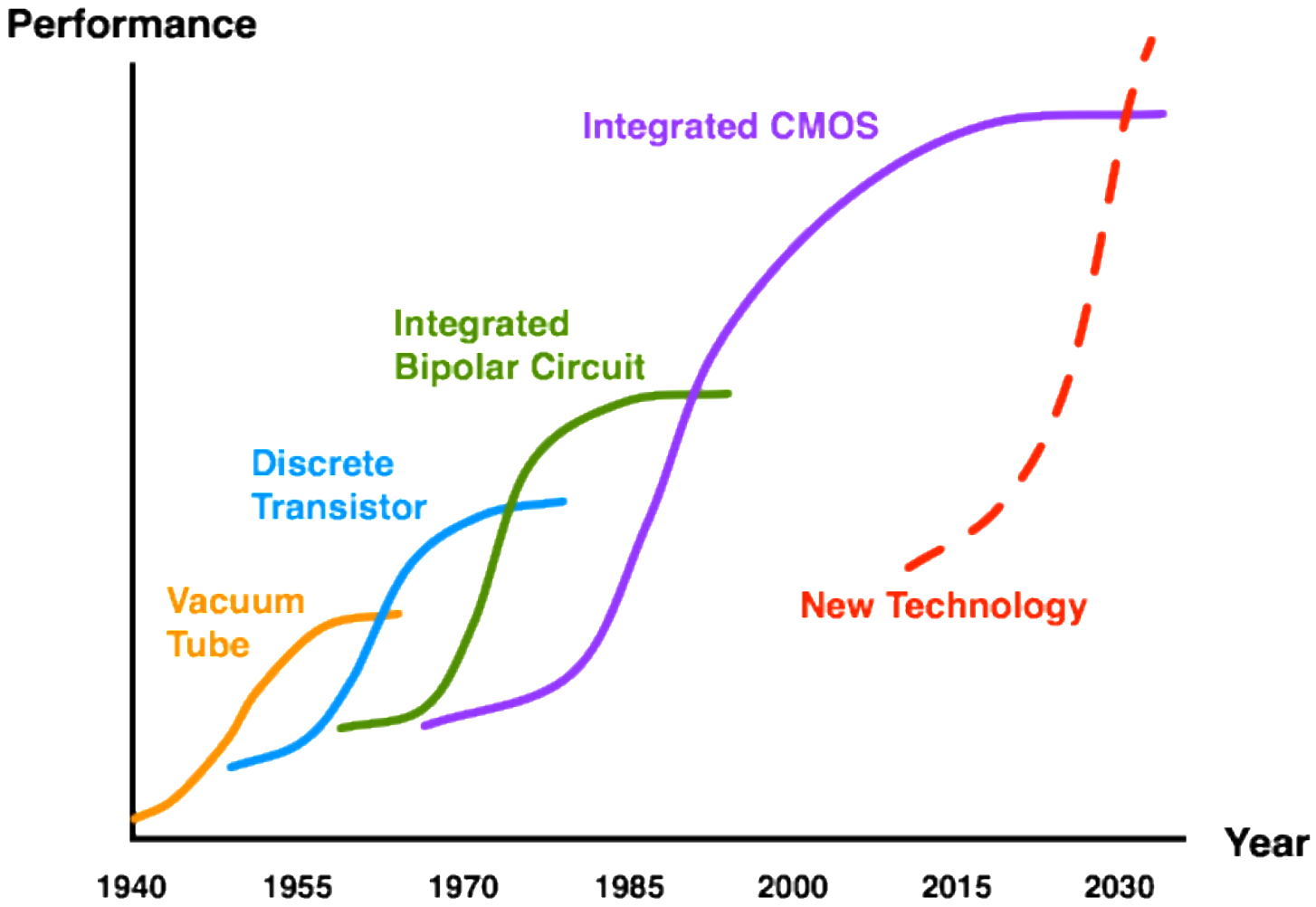
https://visual.ly/community/Infographics/technology/its-smart-world

No More Moore's Law?

Gordon Moore in 1965 that the number of transistors per silicon chip doubles every year.

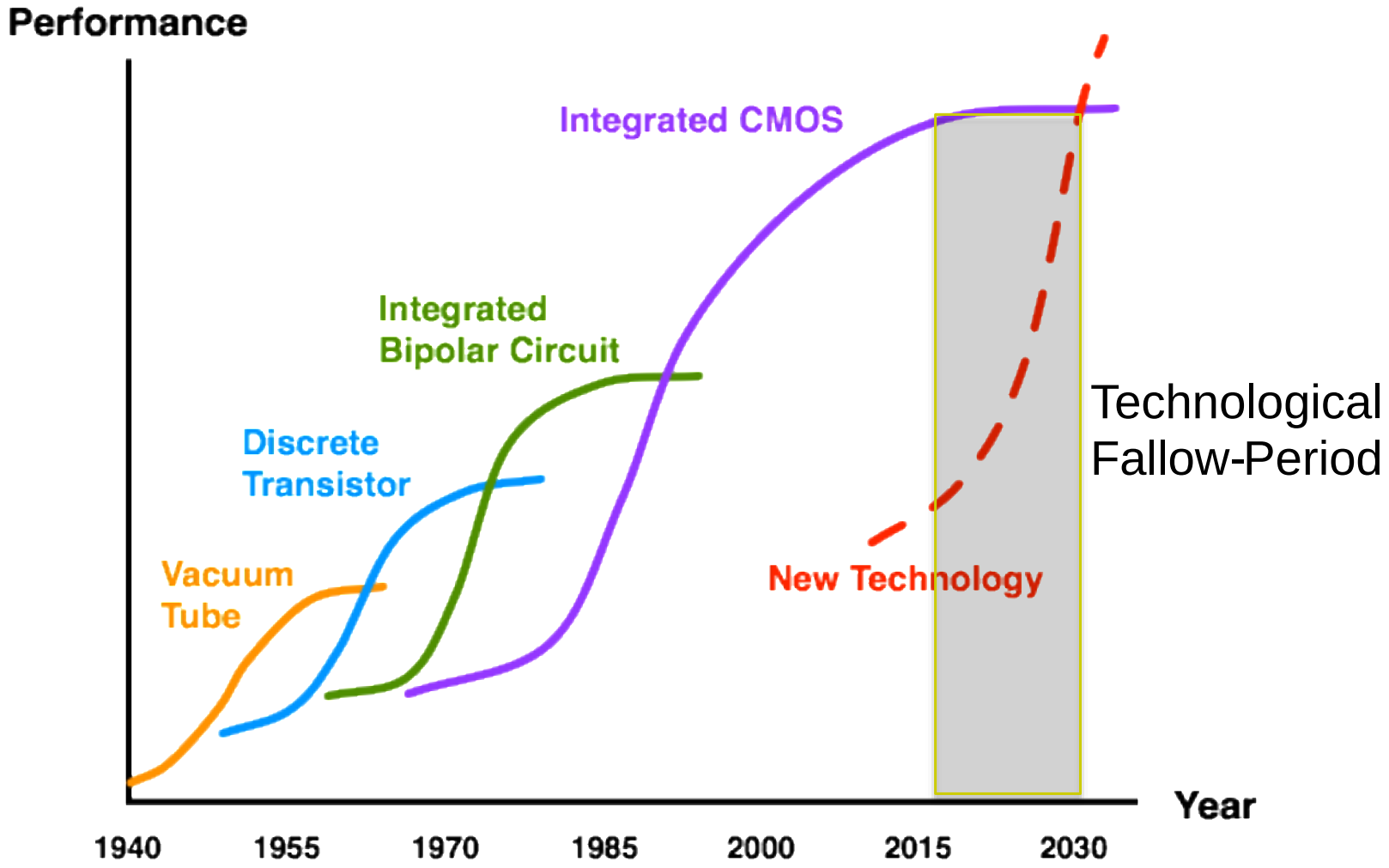


CMOS Technology Scaling



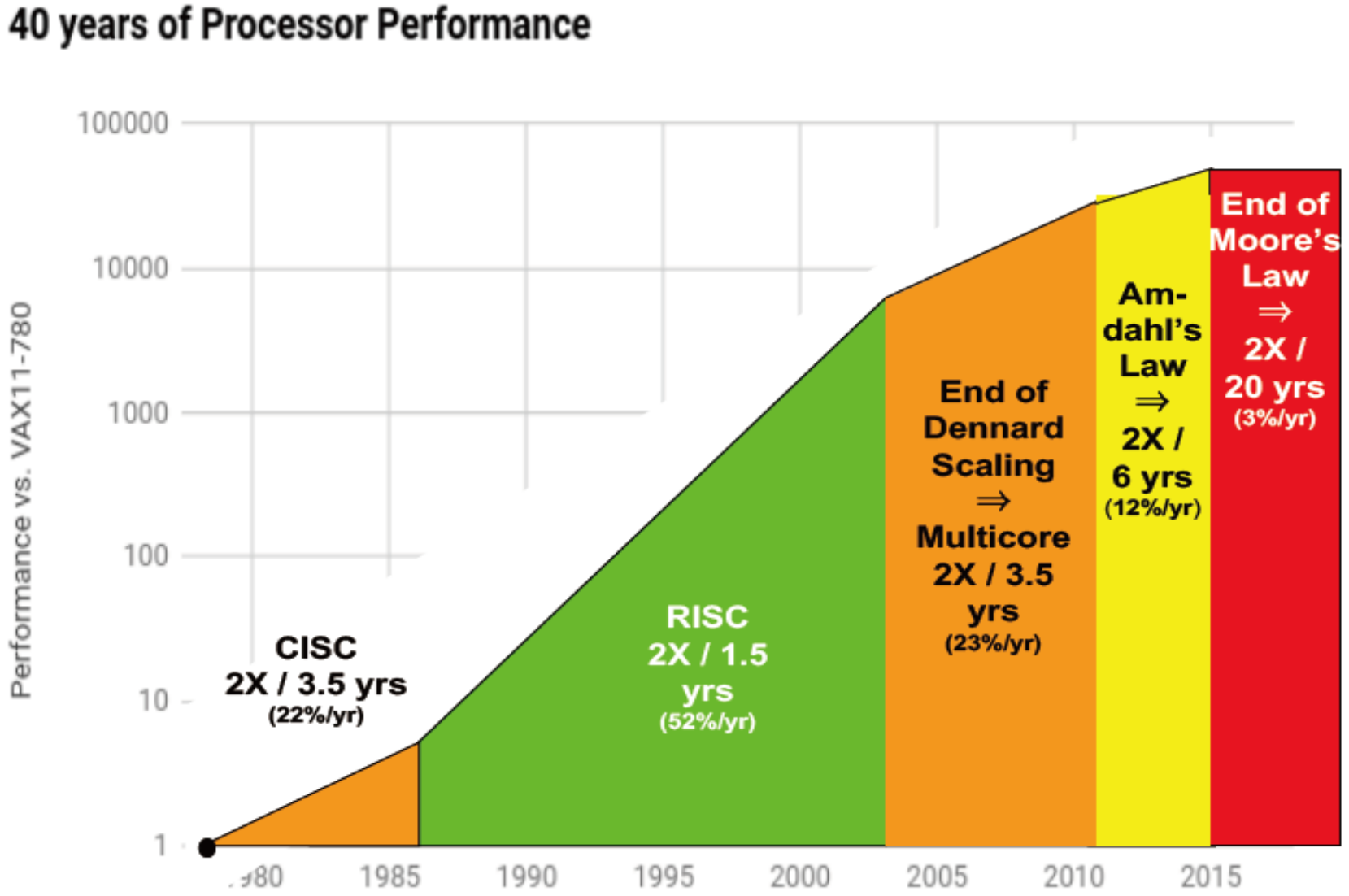
[Source: Prof. David Brooks, Harvard Univ.]

Technological Fallow Period?



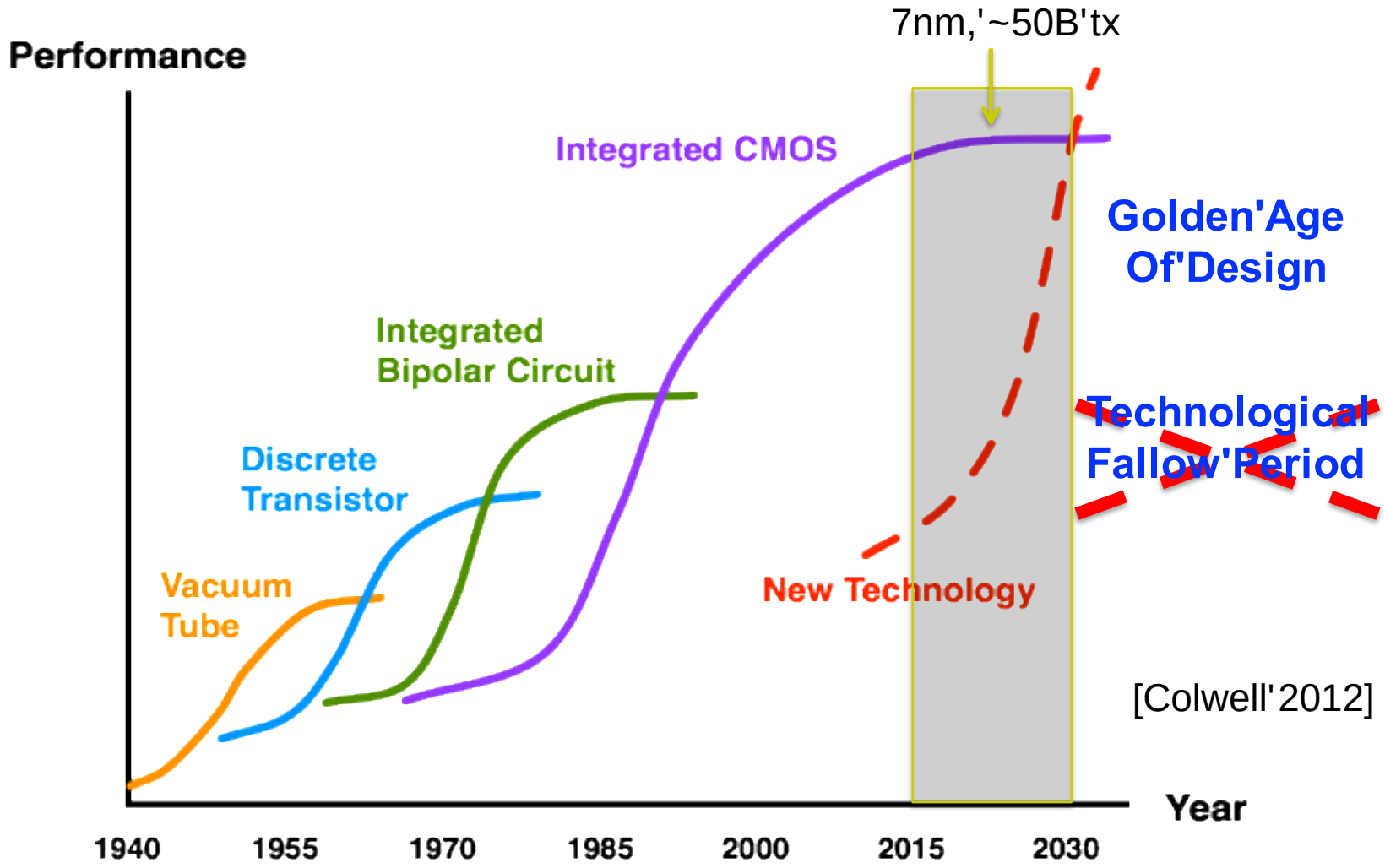
[Source: Prof. David Brooks, Harvard Univ.]

End of Moore's Law!
End of Growth of Performance?



[Source: Prof. David Patterson]

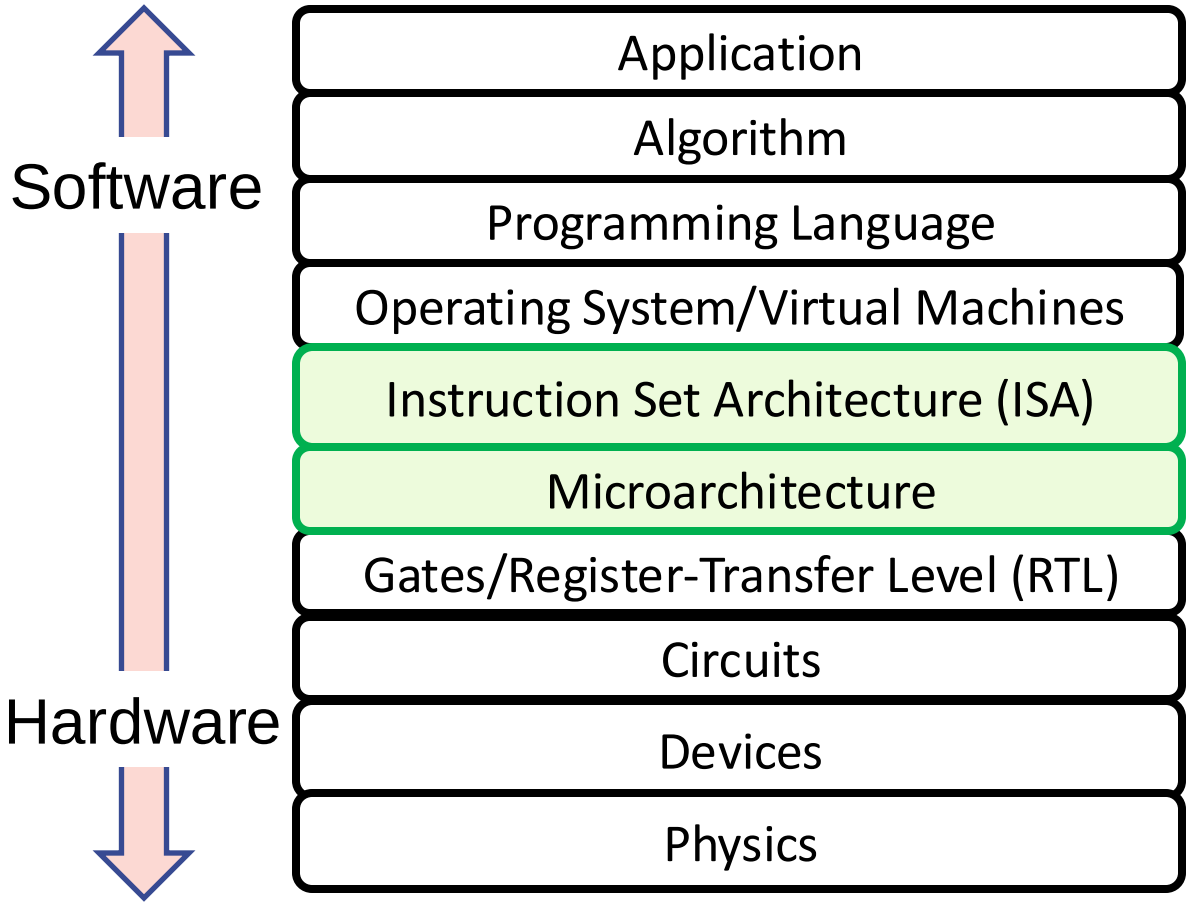
It Comes The Golden Age of Design!



[Source: Prof. David Brooks, Harvard Univ.]

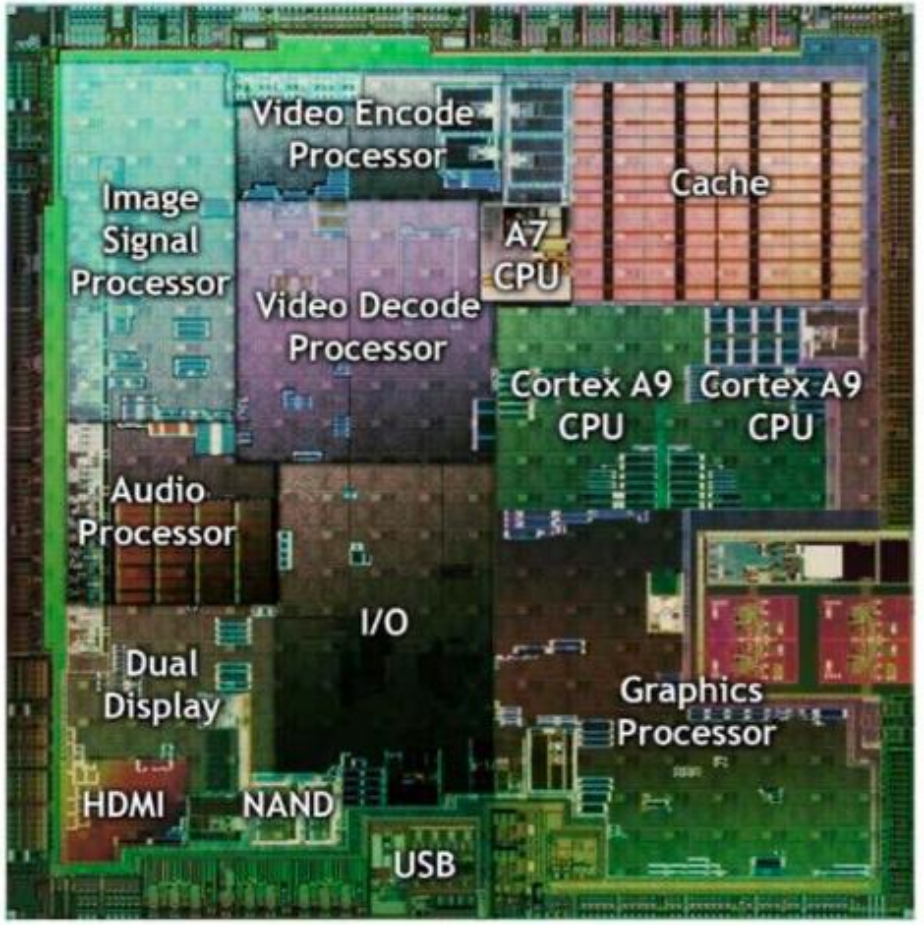
Abstraction Layers in Traditional Computer Systems

- *Design of the abstraction layers* that allow us to implement information processing applications efficiently using available manufacturing technologies.
- Computer architecture acts as the intermediate between programmers and devices (e.g., VLSI)



Today, Many ISAs on One SoC (System-on-Chip)

- Applications processor (usually ARM)
- Graphics processors
- Image processors
- Radio DSPs (Digital Signal Processors)
- Audio DSPs
- Security processors
- Power-management processor
- > dozen ISAs on some SoCs – each with unique software stack
- Why?
  - Apps processor ISA too big, inflexible for accelerators
  - IP bought from different places, each proprietary ISA
  - Engineers build home-grown ISA cores

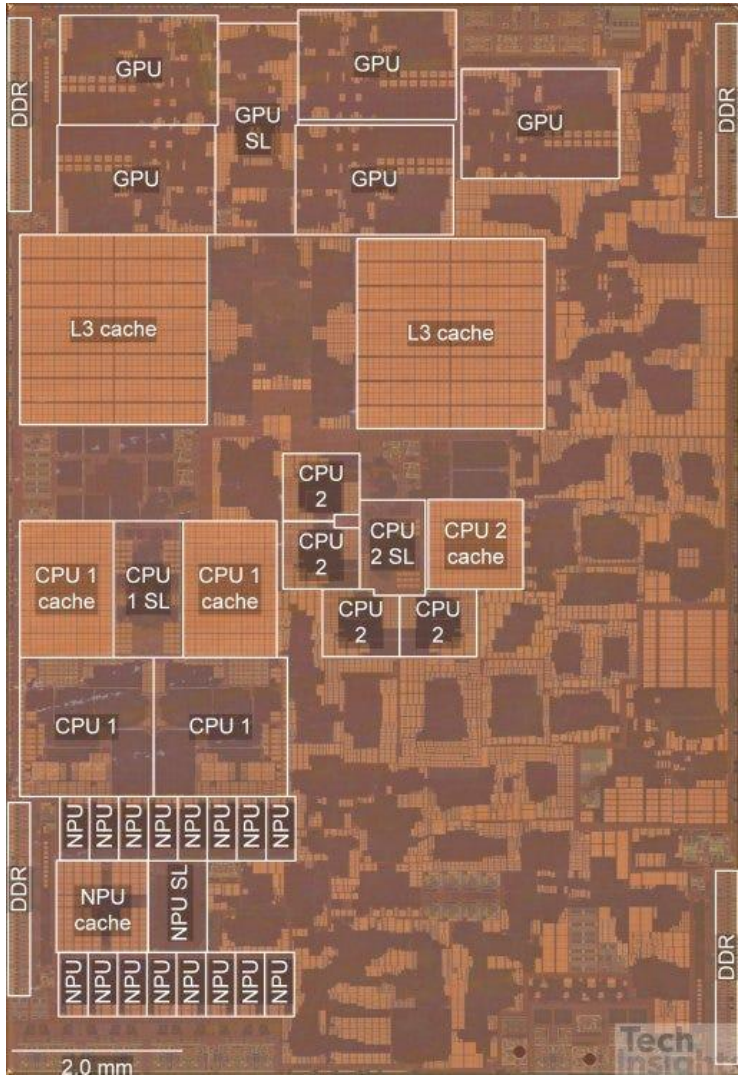


NVIDIA Tegra SoC

Source: NVIDIA

Domain-Specific Accelerators Dominate SoC

- Apple A15 Bionic Processor (2021)
  - Application Processor (AP)
  - TSMC 5nm technology
  - Die size: 107.68 mm^2
  - 15 billion transistors
  - 6-core CPU
  - 4- or 5-core GPU
  - 16-core Neural Engine
    - 15.8 trillion ops/s
  - Image processor
  - Video codec

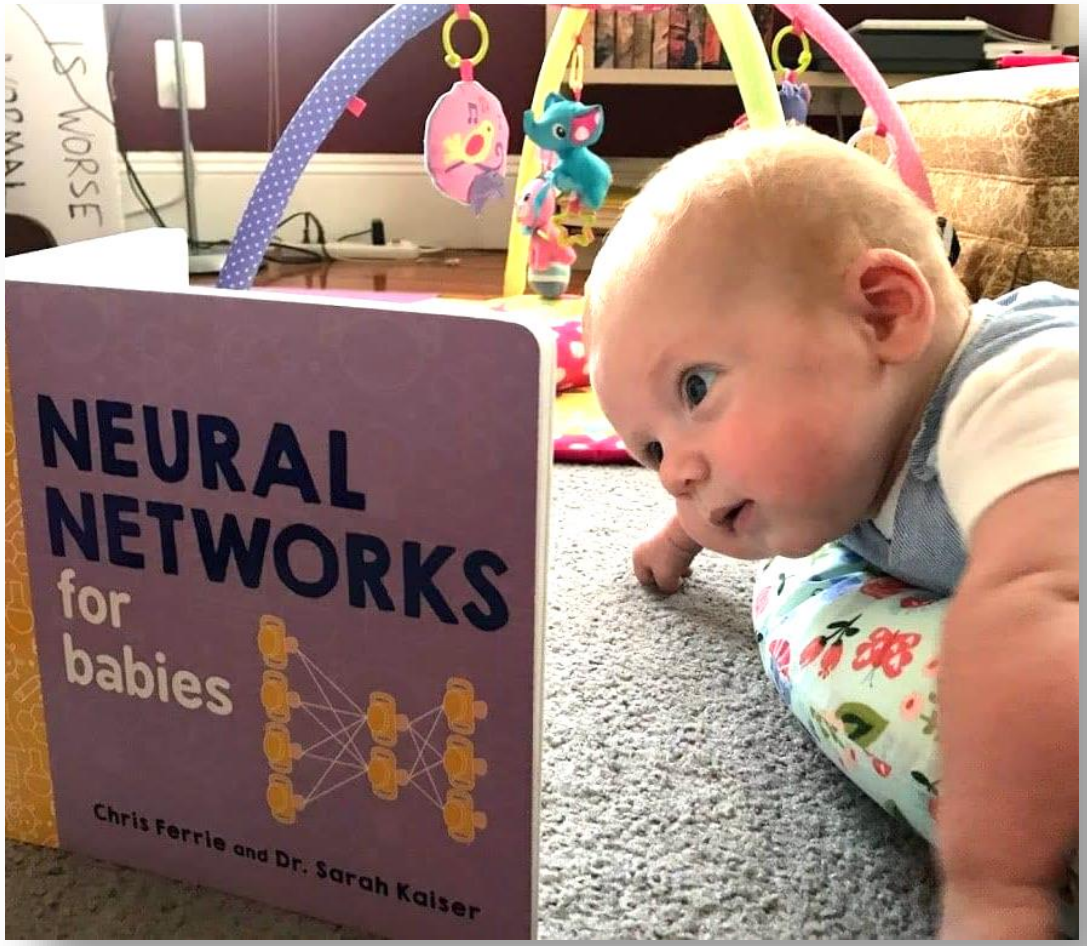


[Source: semianalysis.com]

Which specific domain are we going to discuss about?

AI! Of course!

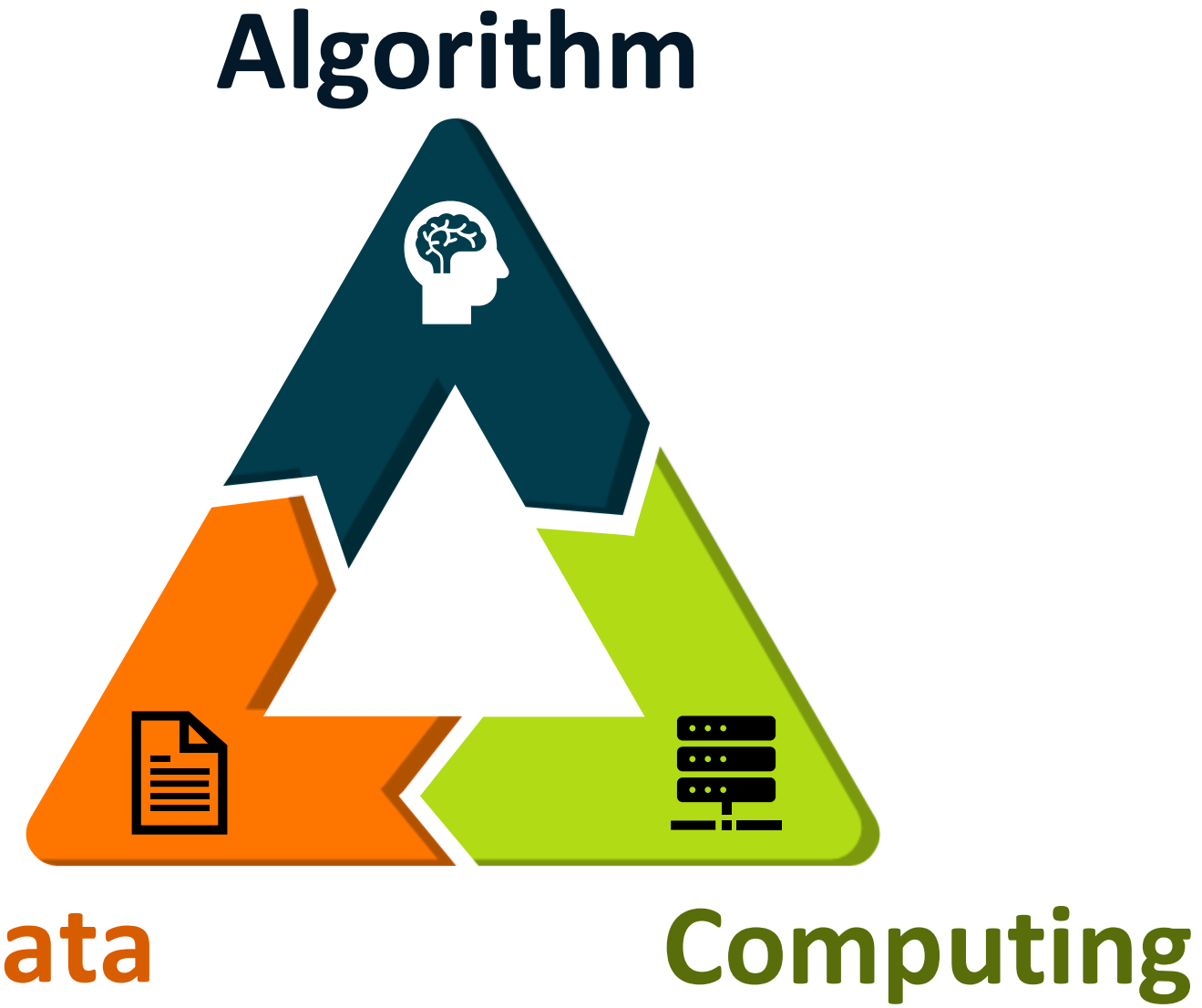
從小和AI做朋友？



https://market.cloud.edu.tw/list/ai.jsp
https://www.amazon.com/Neural-Networks-Babies-Baby-University/dp/1492671207

ABC of AI

- A: Algorithms
  - Improved learning techniques
- B: Big data
  - Significantly larger amounts of digital data
- C: Computing
  - Relatively inexpensive massively parallel computational capabilities

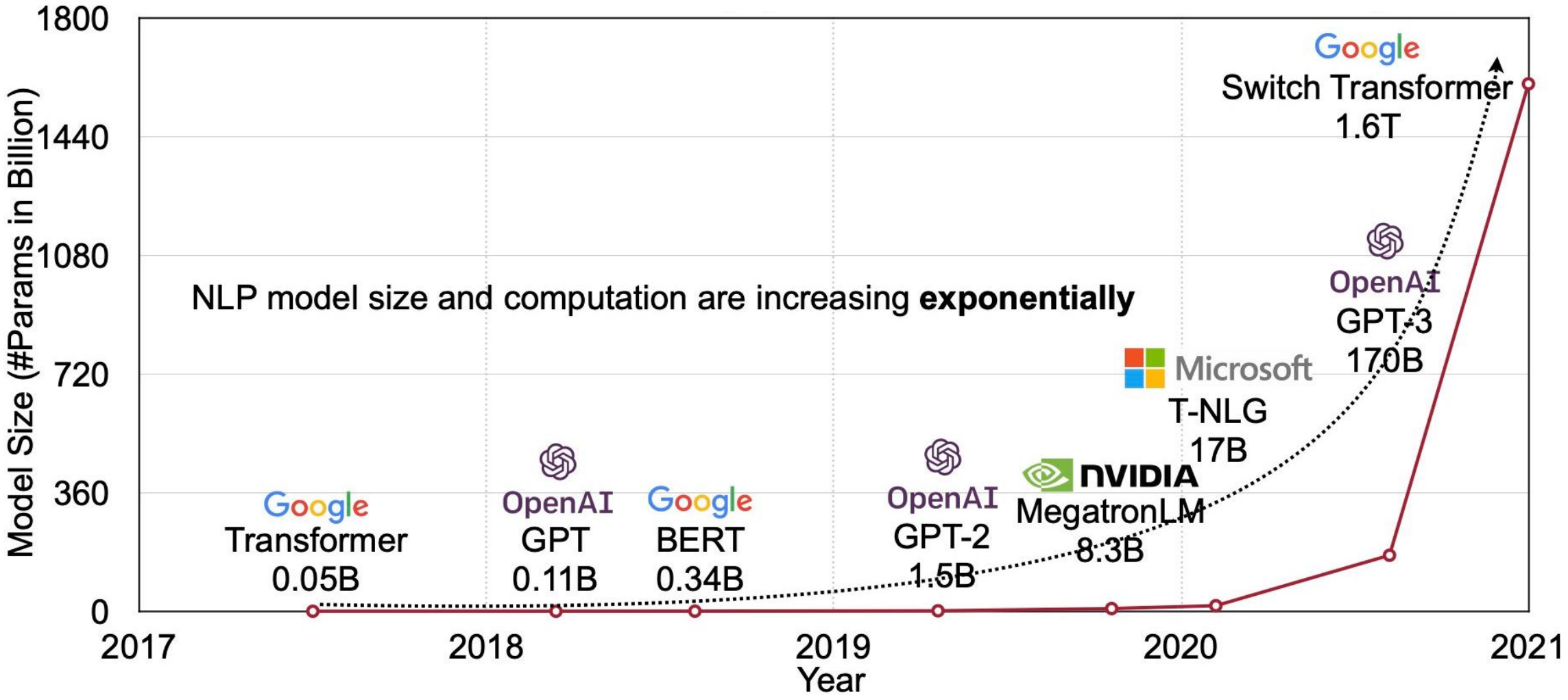


Lately, Software Giants Are Building Their Own AI Chips…

- Accelerators for Deep Learning
  - Google: Tensor Processing Unit (TPU)
    - Google Translate, image search
  - Microsoft: FPGA-based AI supercomputer (Intel/Altera)
  - Facebook: AI chip by Nvidia
  - Amazon: Inferentia
  - Tesla: Full-Self-Driving (FSD) chip
  - Qualcomm/ARM: creating chips to work with TPU
    - Qualcomm Neural Processing Engine for Snapdragon
  - Intel: creating chips optimized for Google's Tensor Flow software (machine learning and neural networks)
- Conventional von Neumann computer architecture is energy inefficient for AI
  - Datacenter: ~2-5 KW
  - 900-core systems: ~100 W
  - Digital hardware accelerator: ~100 mW
  - Neuromorphic processor: 1 mW

NLP's Moore's Law:

Model size increases by 10X every year



Source: MIT Han Lab

CPU vs. GPU

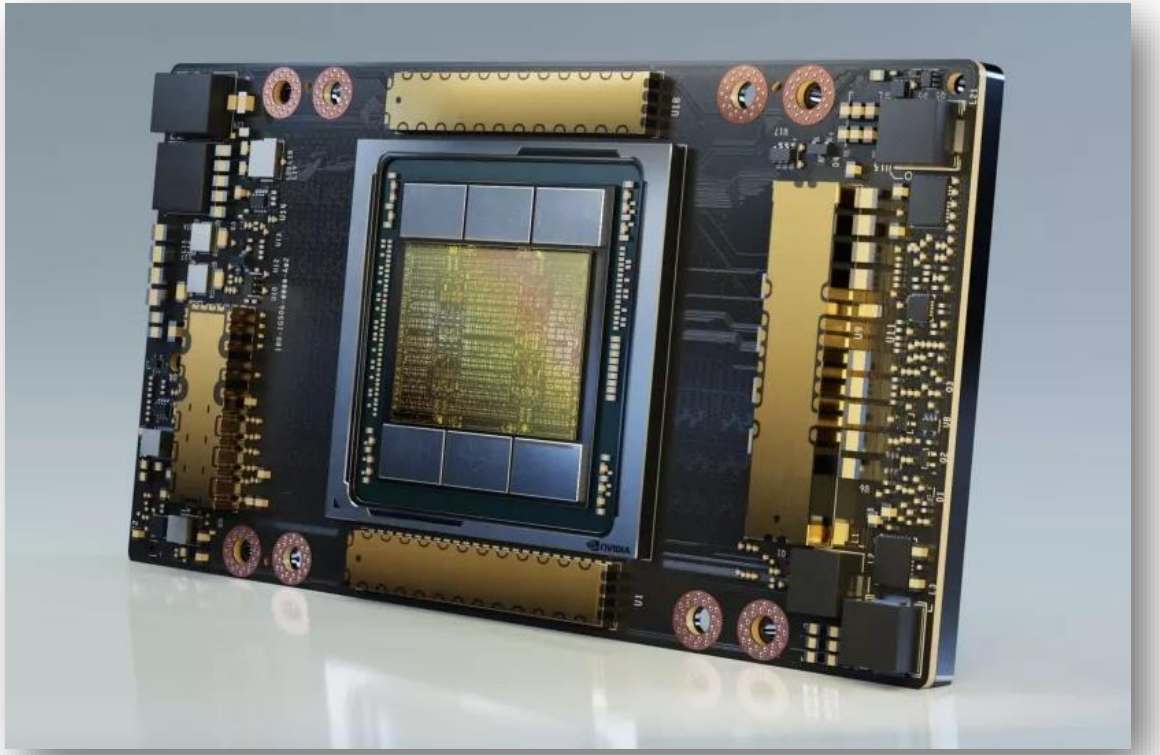
	Cores	Clock Speed	Memory (DRAM)	Price	Throughput
CPU (Intel Core i7 7700k)	4	4.2 GHz	DDR4	$385	~540 GFLOPs
GPU (Nvidia RTX 3090 Ti)	10496	1.7 GHz	DDR6 24GB	$1499	36 TFLOPs

6.7X

- CPU: a small number of complex cores
  - Clock speed of each core is high
  - Good for sequential, dynamic tasks
- GPU: a large number of simple cores
  - Clock speed of each core is low
  - Good for parallel, deterministic tasks

NVIDIA A100 GPU

- Ampere architecture
- TSMC 7 nm FinFET
- 54 billion transistors
- Die size 826 mm^2
- Stream multiprocessors: 108
- CUDA cores: 6,912
- Tensor cores: 432
- GPU memory: 80 GB HBM2
- GPU memory bandwidth: 2.39 TB/s
- Bus width: 5120
- FP32: 19.5 TFLOPS
- Tensor FP16 (sparsity): 312 (624) TFLOPS
- Max Thermal Design Power (TDP): 400W
- Launch date: May 2020
- Launch price:
  $199K for DXG A100 (with 8xA100)

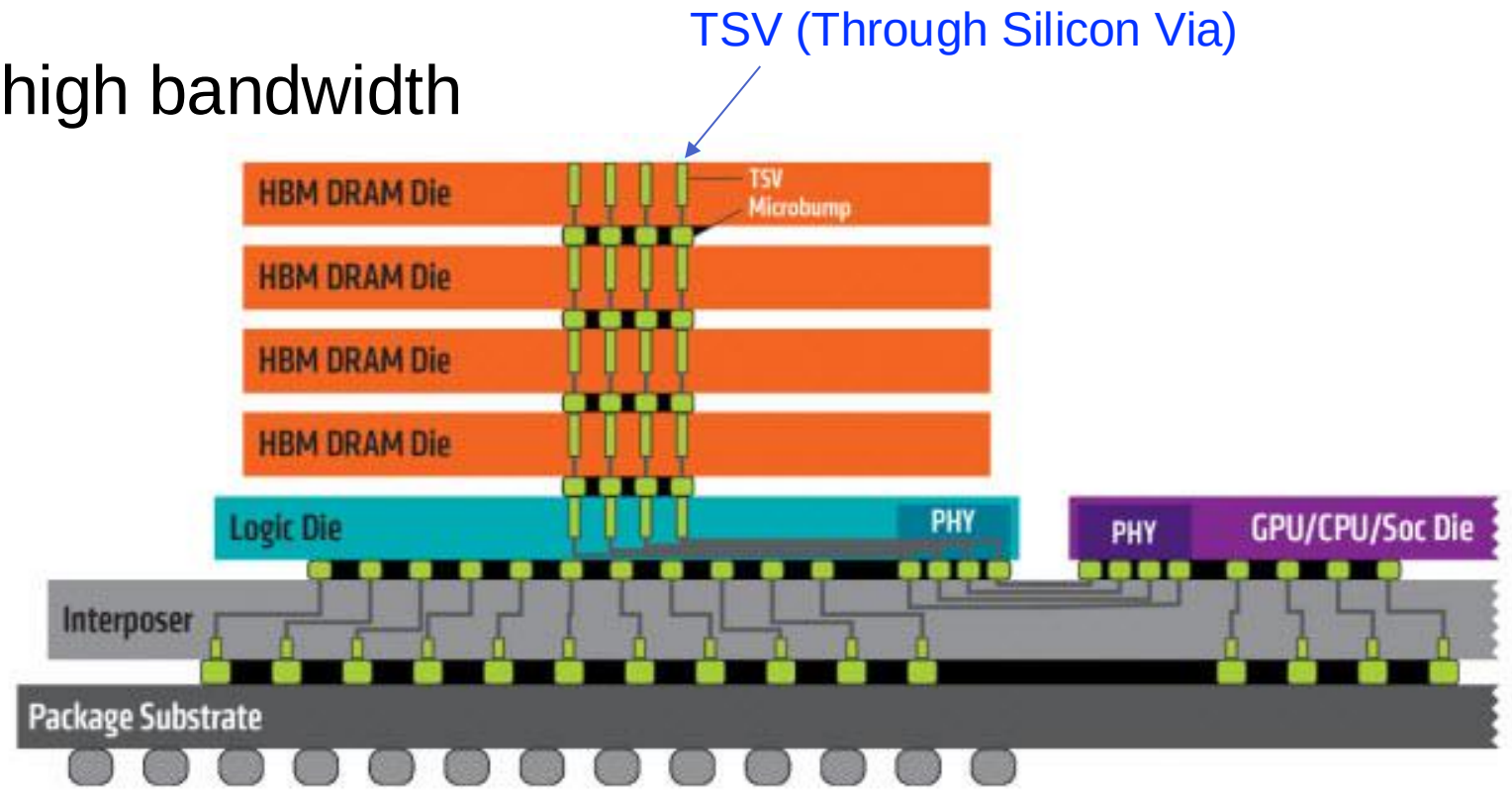


[Source: NVIDIA]

High-Performance Graphics Memory

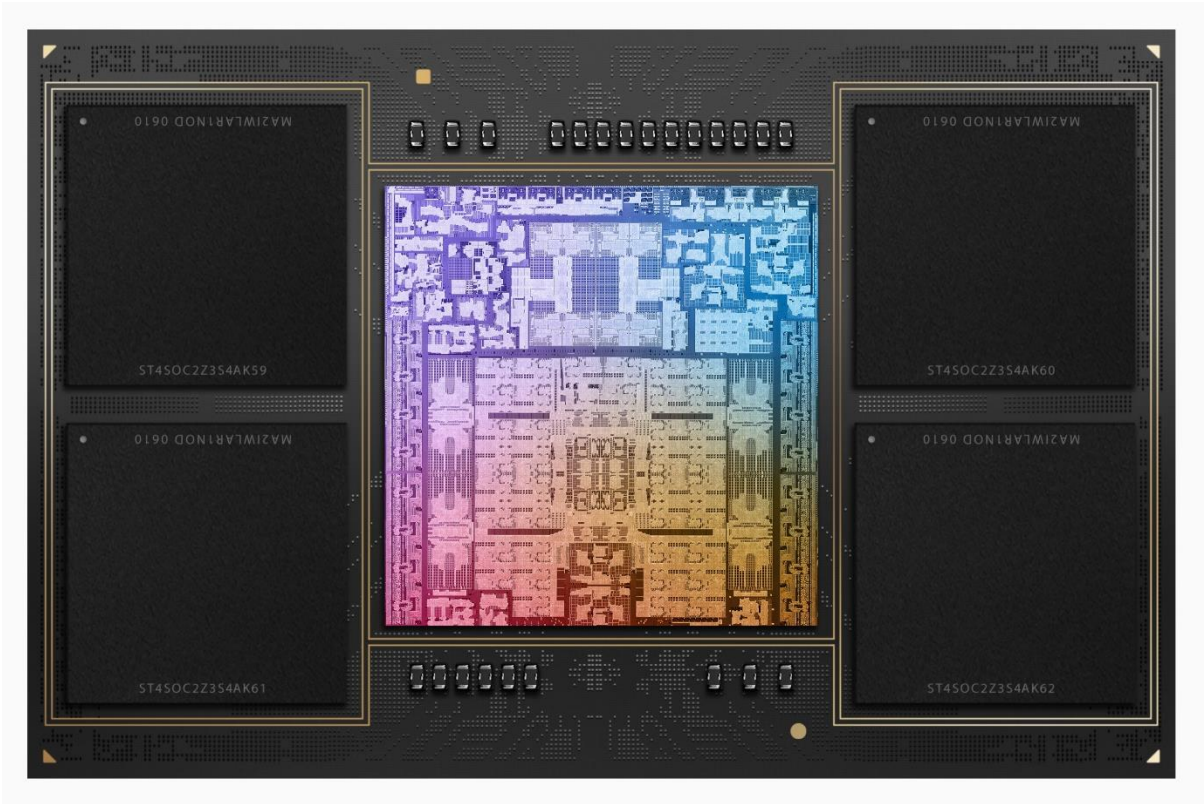
- Modern GPUs even employing 3D-stacked memory via silicon interposer
  - Very wide bus, very high bandwidth
  - E.g., HBM2 in Volta

    (High Bandwidth Memory)



Graphics Card Hub, “GDDR5 vs GDDR5X vs HBM vs HBM2 vs GDDR6 Memory Comparison,” 2019

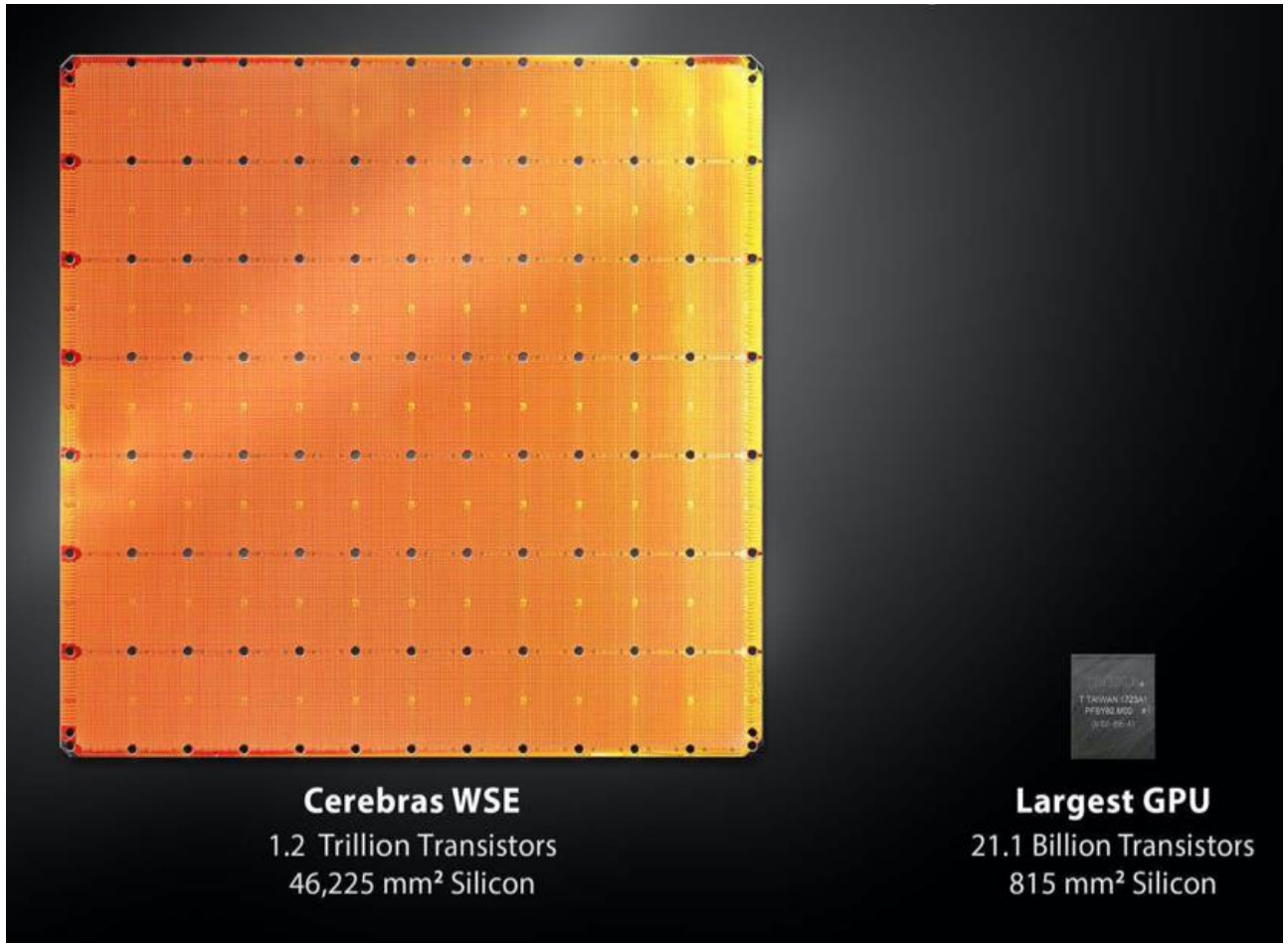
Apple M2 MAX



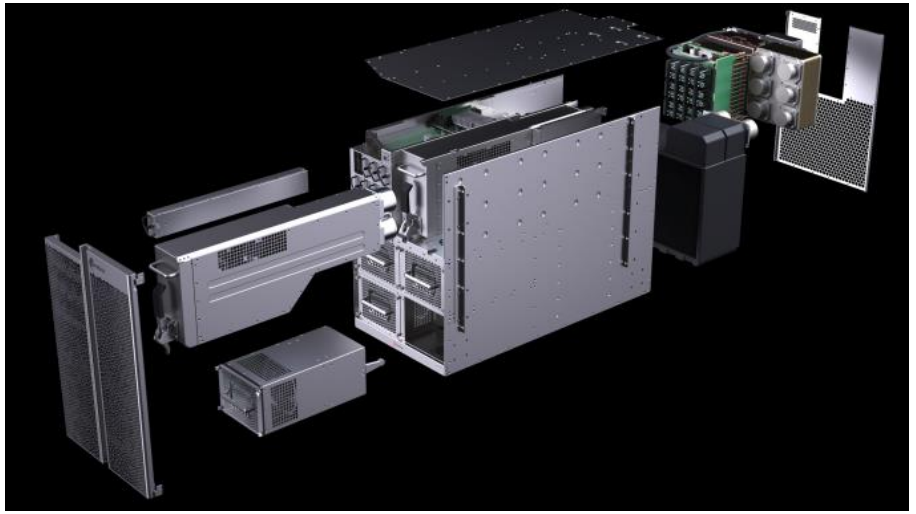
[Source: Apple]

- 5nm technology
- 67 billion transistors
- 12-core CPU
- 38-core GPU
- 16-core Neural engine
  - 15.8 trillion ops/s
- 96GB unified memory
  - LPDDR5
  - Memory bandwidth: 400GB/s

Largest Chip Ever Built: Cerebras Wafer Scale AI Engine



- 46,225 mm^2 silicon
- 1.2 trillion transistors
- 400,000 AI optimized cores
- 18 Gigabytes of On-chip Memory
- 9 PByte/s memory bandwidth
- 100 Pbit/s fabric bandwidth
- TSMC 16nm FinFET process

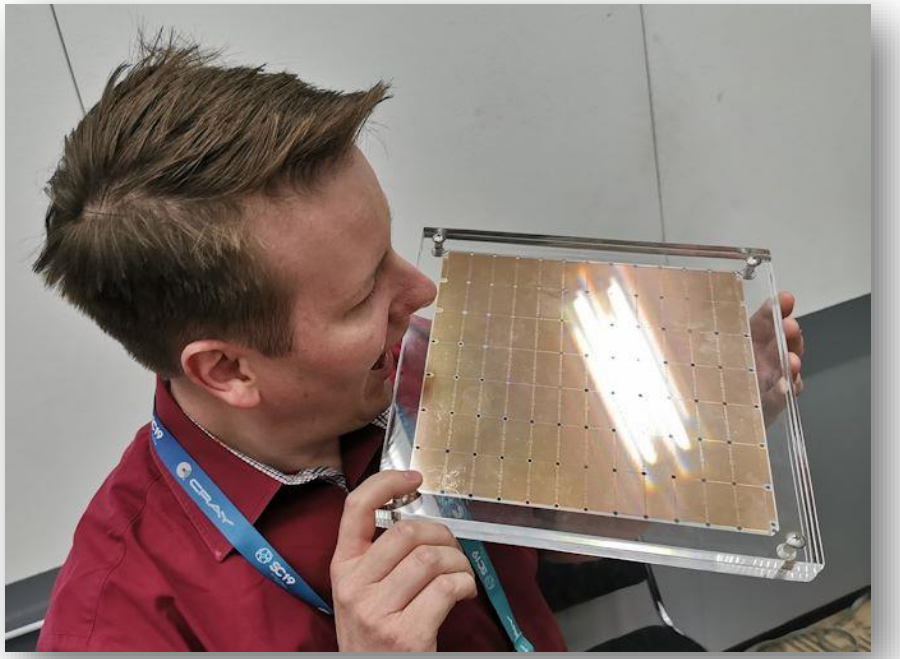


[Source: Cerebras]

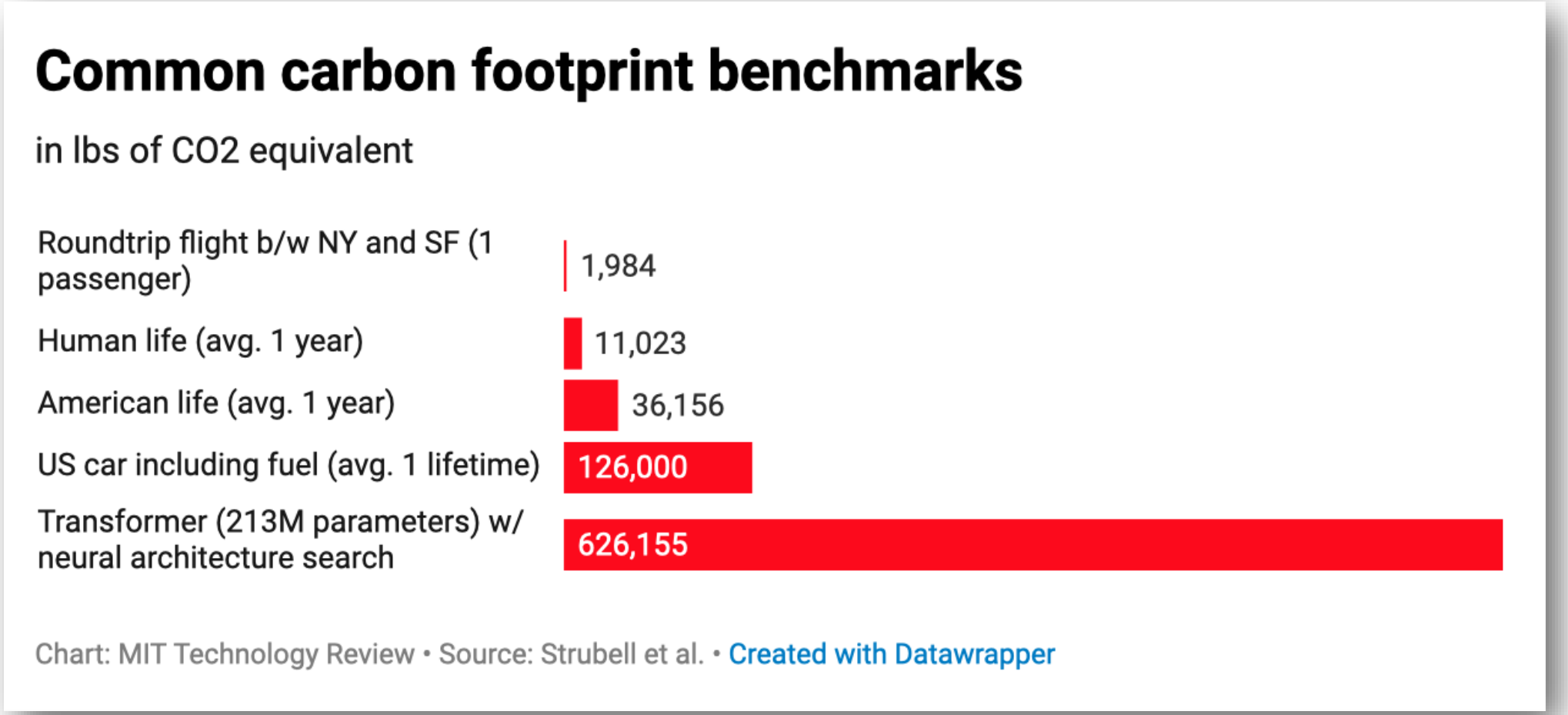
Cerebras Wafer Scale Engine

- Cerebras' Wafer Scale Engine Scores a Sale:
  $5M Buys Two for the Pittsburgh Supercomputing Center
  [AnandTech, June 9, 2020]

https://www.anandtech.com/show/15838/cerebras-wafer-scale-engine-scores-a-sale-5m-buys-two-for-the-pittsburgh-supercomputing-center



AI Consumes Too Much Energy!!!

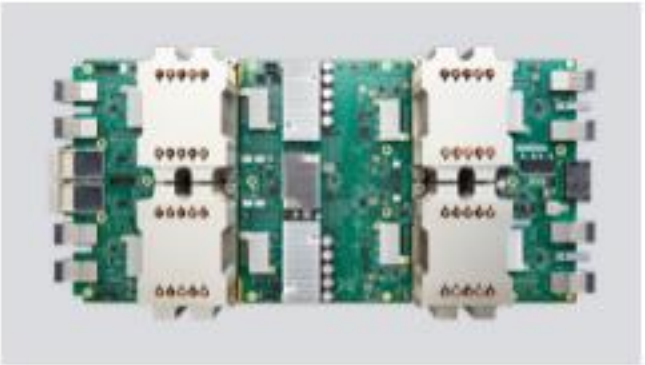


Global Shortage of Energy Supply

- In 2014, cloud data centers consumed about 1.62% of global energy
- CNBC interviewed David Patterson
  [www.cnbc.com, 5/6/2017]
  - “Four years ago, Google worried that if every Android user had 3 minutes of conversation translated a day using machine learning, they'd have to double their data centers.”
  - Alphabet spend about $10B each year on Google data center equipment
    - New data centers or improved equipment
  - Google TPU outperforms CPU by 15-30X
    - 30-80X in energy efficiency

Google TPU (Tensor Processing Unit)

- Systolic MAC array
  - V1: 8-bit Inference
  - V2: Training with bfloat
  - V3: 2x powerful over V2
  - AlphaGo V1 to V3
    - 1000X lower power
- Edge TPU
  - Coral Dev Board
  - 4 TOPS
  - 2 TOPS/Watt
  - Supports TensorFlow Lite



Cloud TPU v2

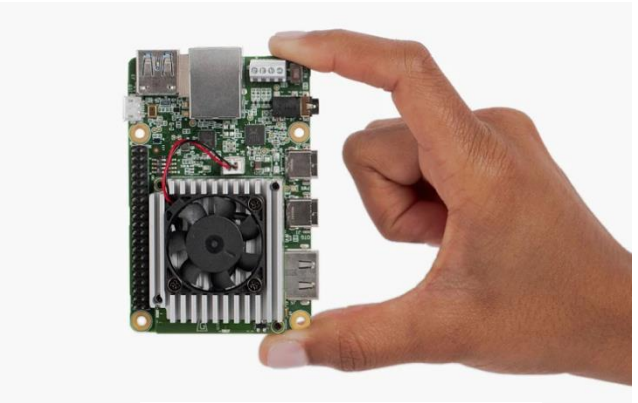
180 teraflops

64 GB High Bandwidth Memory (HBM)

Cloud TPU v3

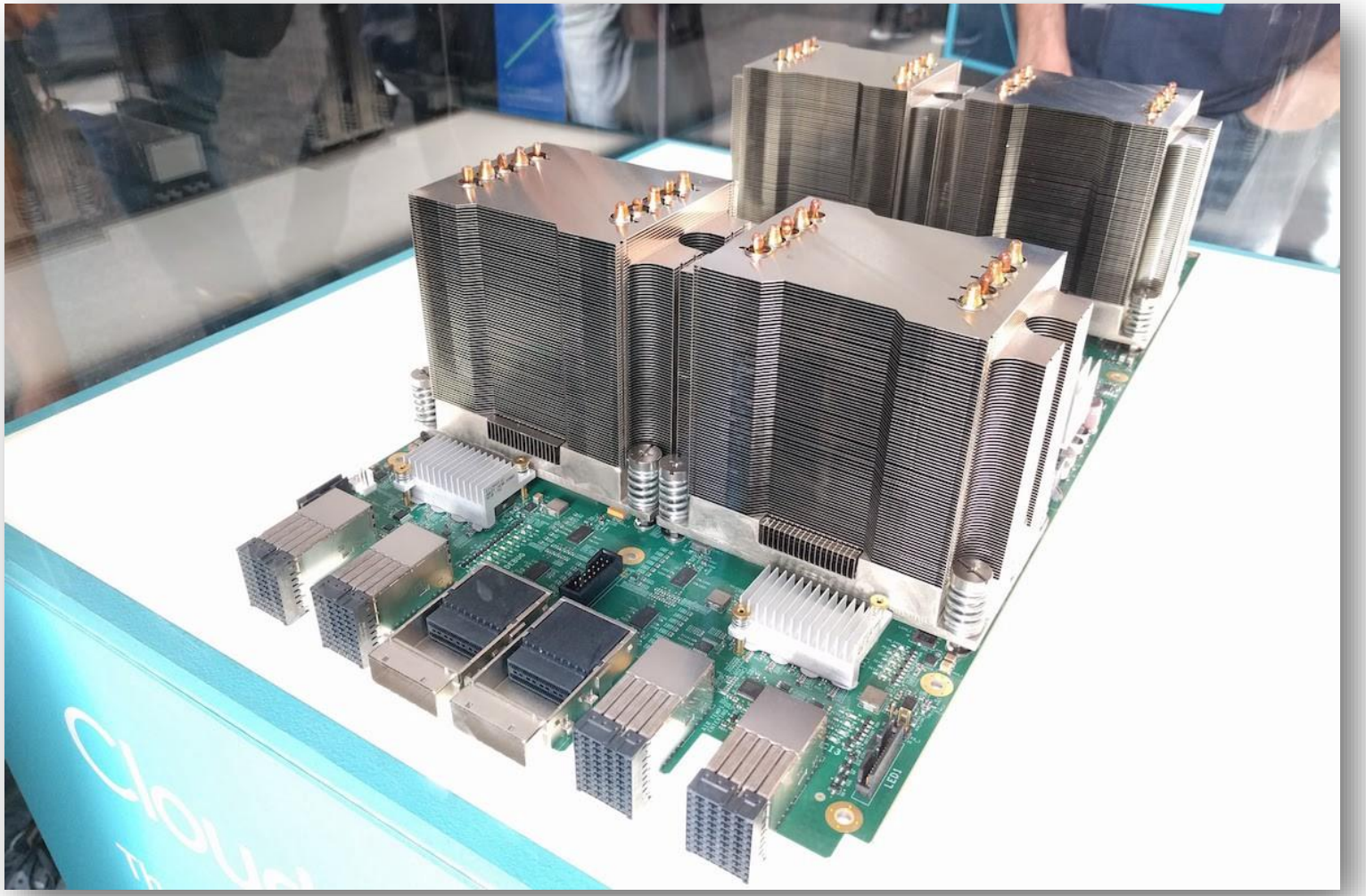
420 teraflops

128 GB HBM



[Source: Google]

A Closer View of Google TPU V2

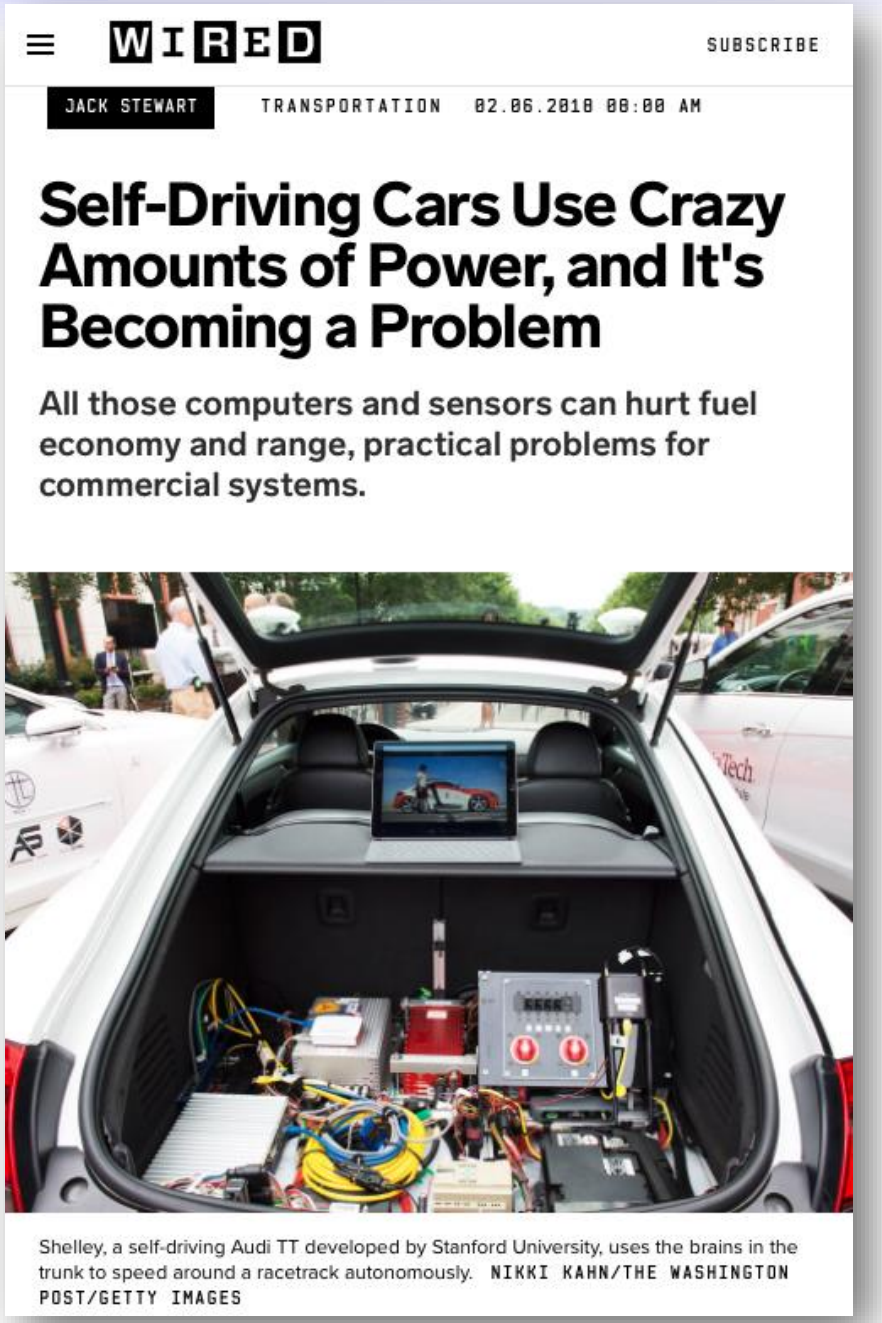


Source: https://medium.com/@antonpaquin/whats-inside-a-tpu-c013eb51973e

AI Consumes Too Much Energy!!!

WIRED [Feb 2018]

- Cameras/radar generate about **6 gigabytes** of data **every 30 seconds**.
- Self-driving car prototypes use approximately **2,500 watts** of computing power!



WIRED SUBSCRIBE

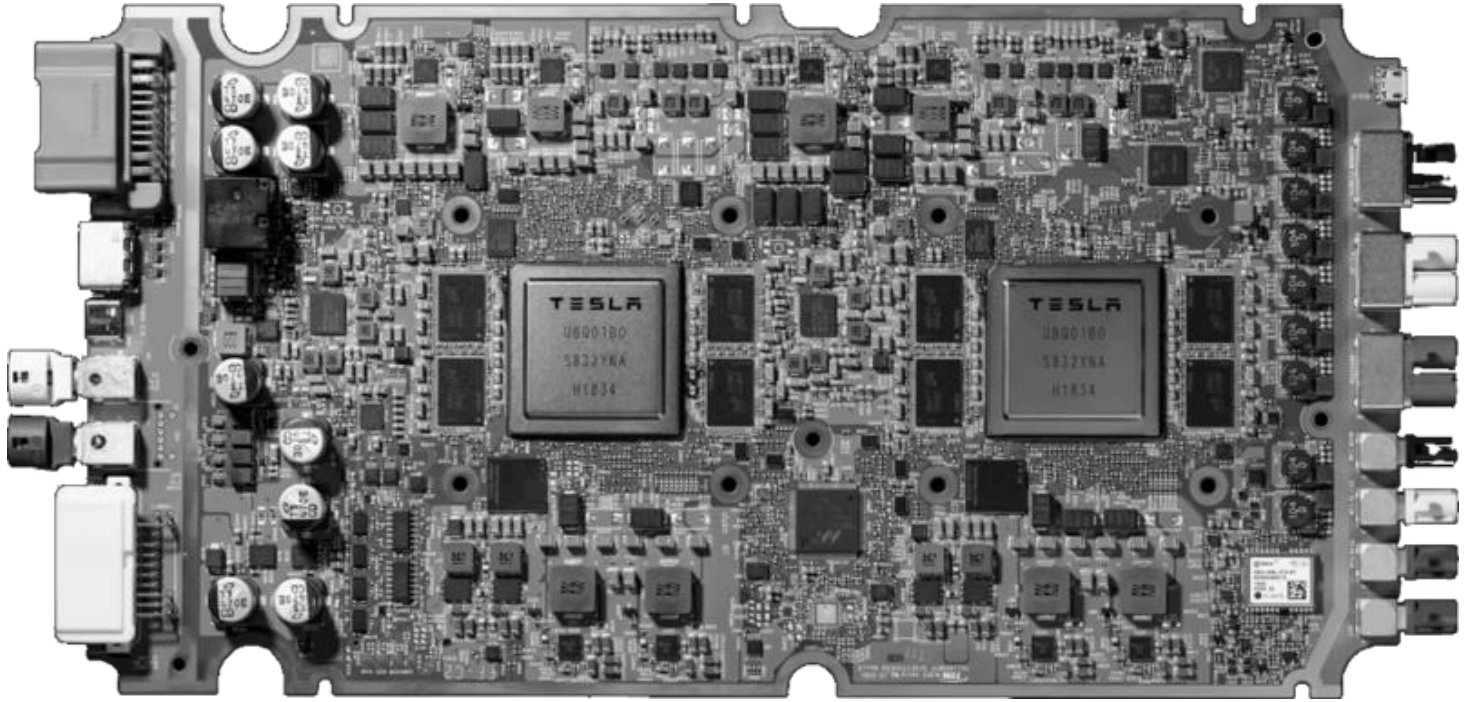
JACK STEWART TRANSPORTATION 02.06.2018 08:00 AM

Self-Driving Cars Use Crazy Amounts of Power, and It's Becoming a Problem

All those computers and sensors can hurt fuel economy and range, practical problems for commercial systems.

Shelley, a self-driving Audi TT developed by Stanford University, uses the brains in the trunk to speed around a racetrack autonomously. NIKKI KAHN/THE WASHINGTON POST/GETTY IMAGES

Tesla Full Self-Driving Computer



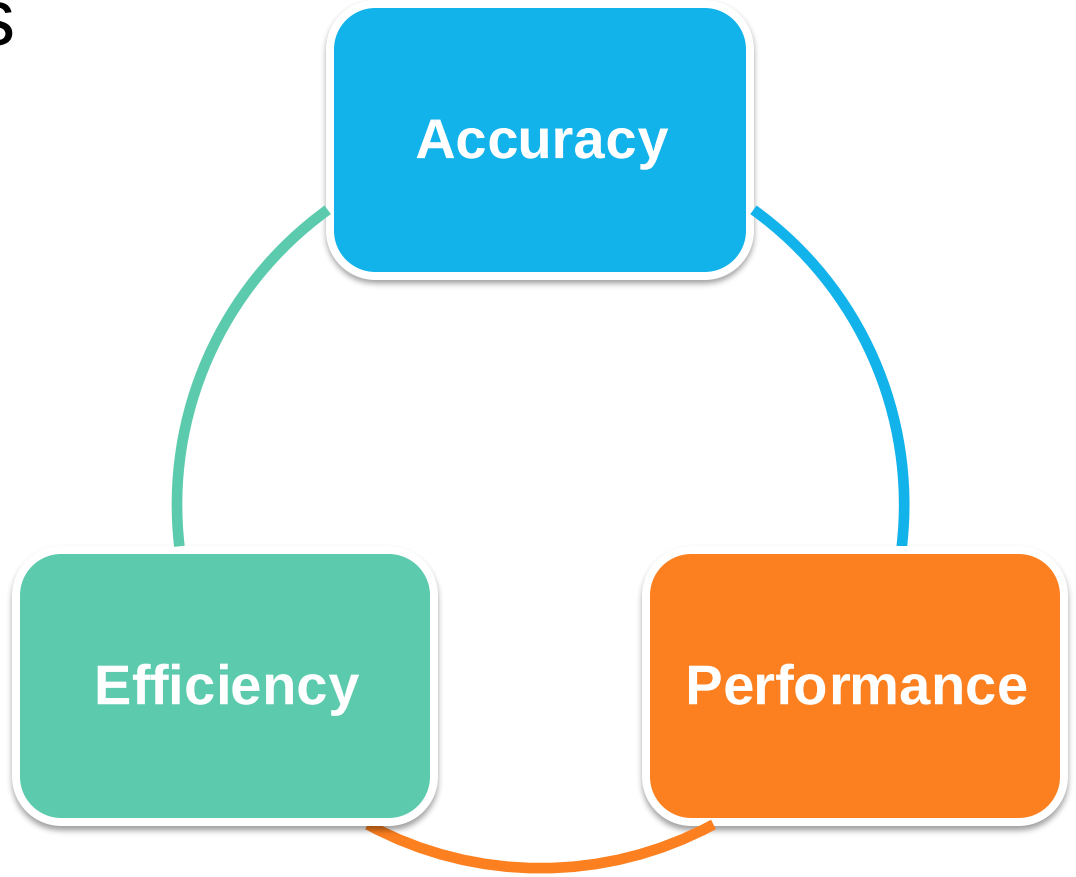
[Source: Tesla]

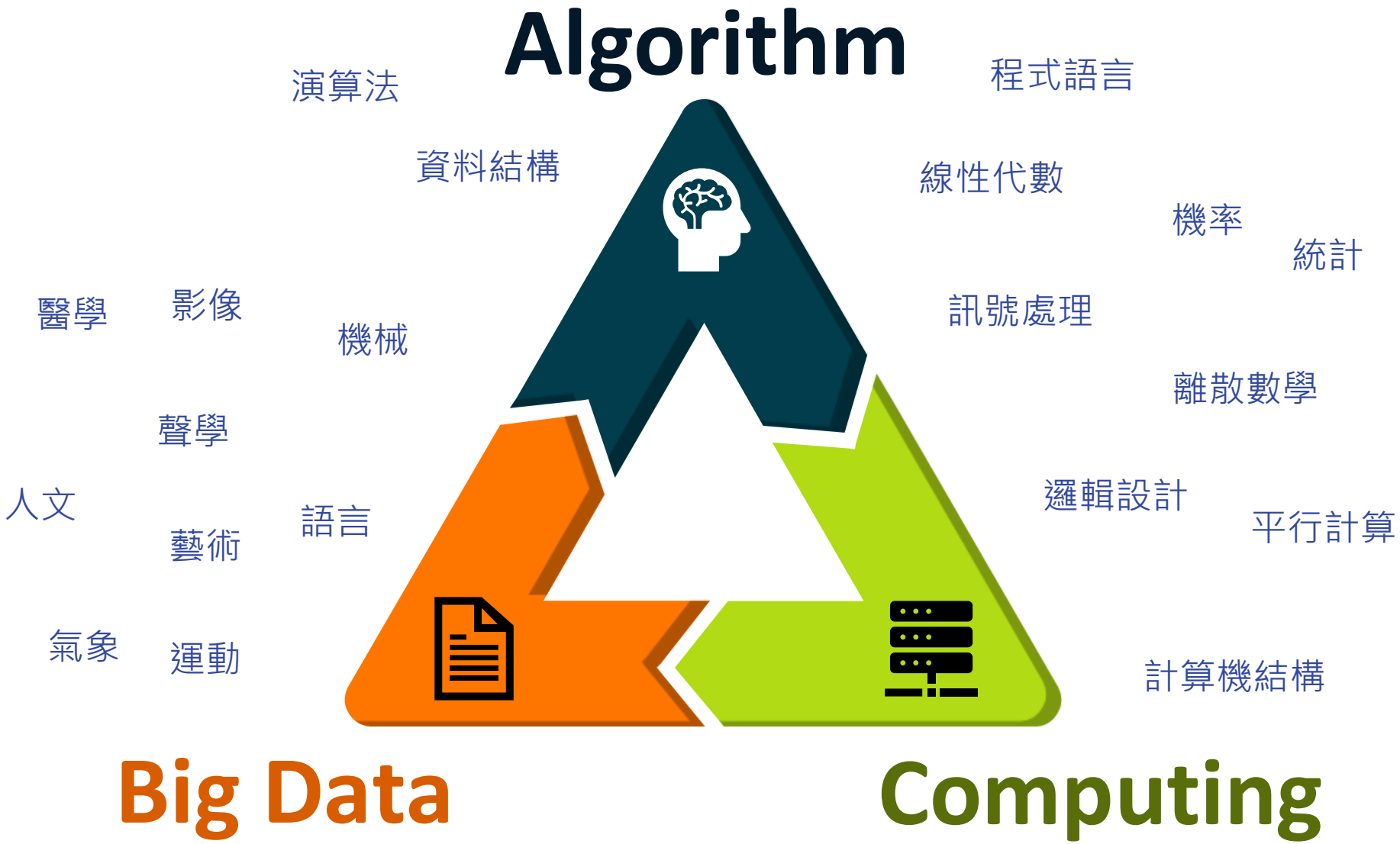
Tesla FSD chip
- ~74 TOPS
- 36 Watts

NVIDIA DGX-1
- 130-170 TFLOPS
- 3200 Watts

Efficient System Design for Deep Learning

- How do we identify the hot spots of the problem?
  - Profiling the bottleneck
- How do we optimize the system
  - Hardware/software co-design
  - Trade-off among
    - Accuracy
    - Performance
    - Efficiency




Algorithm
演算法
資料結構
程式語言
線性代數
機率
統計
訊號處理
離散數學
邏輯設計
平行計算
計算機結構
醫學
影像
機械
聲學
人文
語言
藝術
氣象
運動
Big Data
Computing


Algorithm
成為跨領域、
多領域的專家
必須具備紮實
的基礎知識
Big Data
Computing
醫學
影像
聲學
人文
藝術
氣象
運動
機率
統計
離散數學
設計
平行計算
計算機結構

“People who are really serious about software should make their own hardware.”

– Alan Kay

“Design is not just what it looks like and feels like. Design is how it works.”

– Steve Jobs

“We choose to go to the Moon in this decade and do the other things, not because they are easy, but because they are hard.”

– JFK, 1962