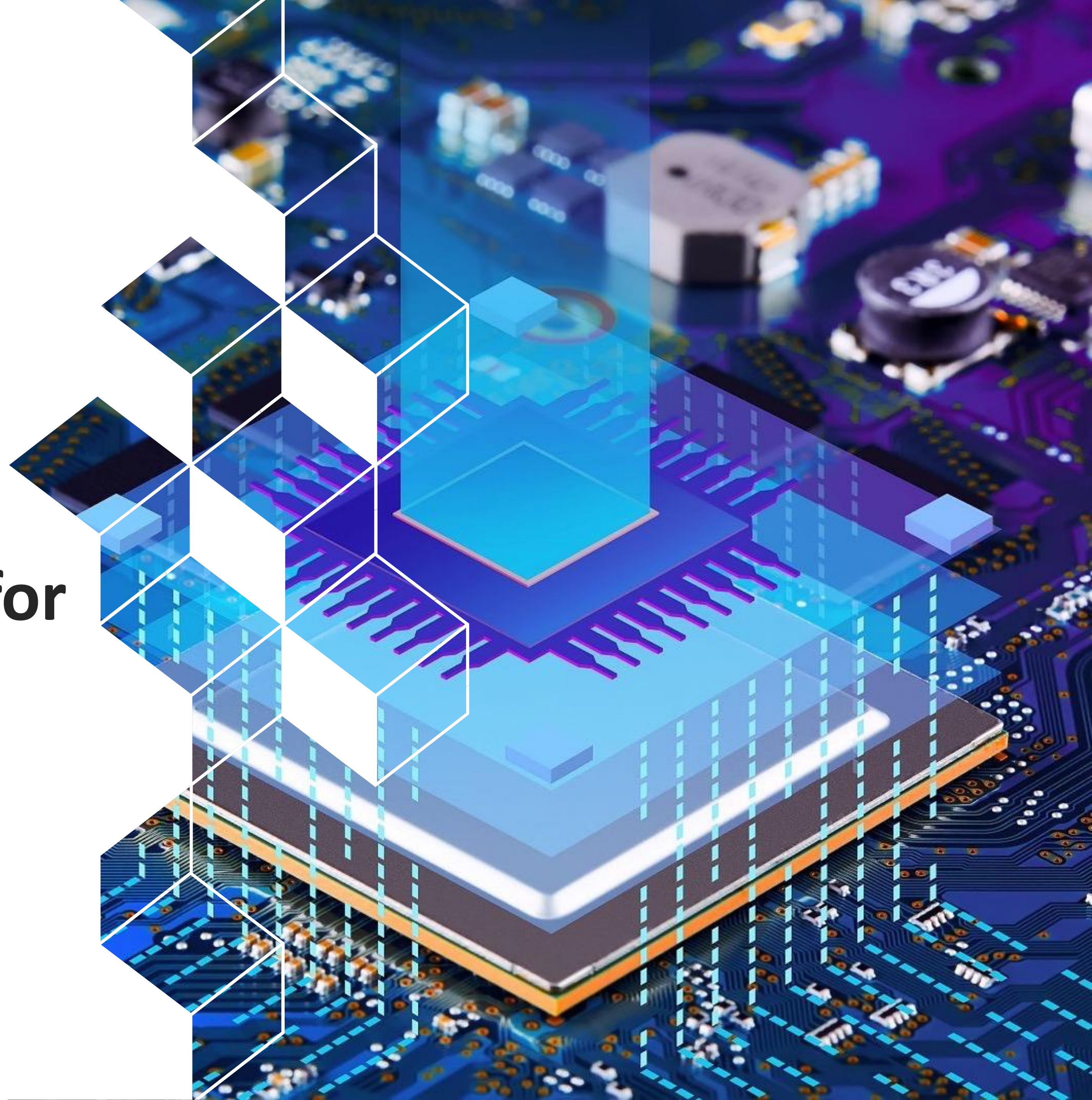
How to make Generative AI more sustainable ?

Dr. Marc Duranton
(マーク・デュラントン)
Senior Fellow – CEA (F)

November 13th, 2024

#Trilateral_AI

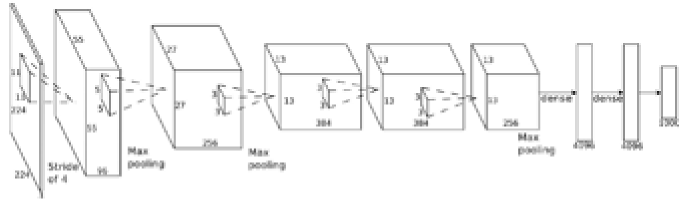
Short history of hardware for (g) Artificial Intelligence based on Neural Networks



2012: DEEP NEURAL NETWORKS RISE AGAIN

They give the state-of-the-art performance e.g. in image classification

- **ImageNet classification** (Hinton's team, hired by Google)
  - 14,197,122 images, 1,000 different classes
  - Top-5 17% error rate (huge improvement) in 2012 (now ~ 3.5%)



Year: 2012
650,000 neurons
60,000,000 parameters
630,000,000 synapses

- Facebook's 'DeepFace' Program (labs headed by Y. LeCun)
  - 4.4 million images, 4,030 identities
  - 97.35% accuracy, vs. 97.53% human performance

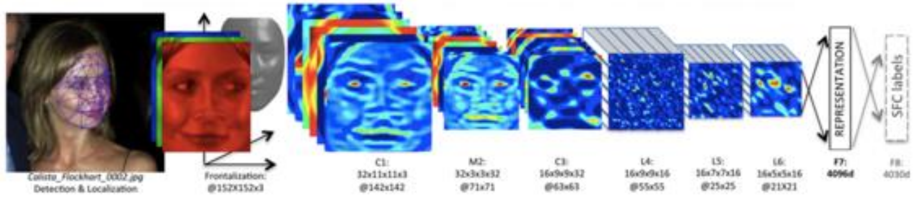
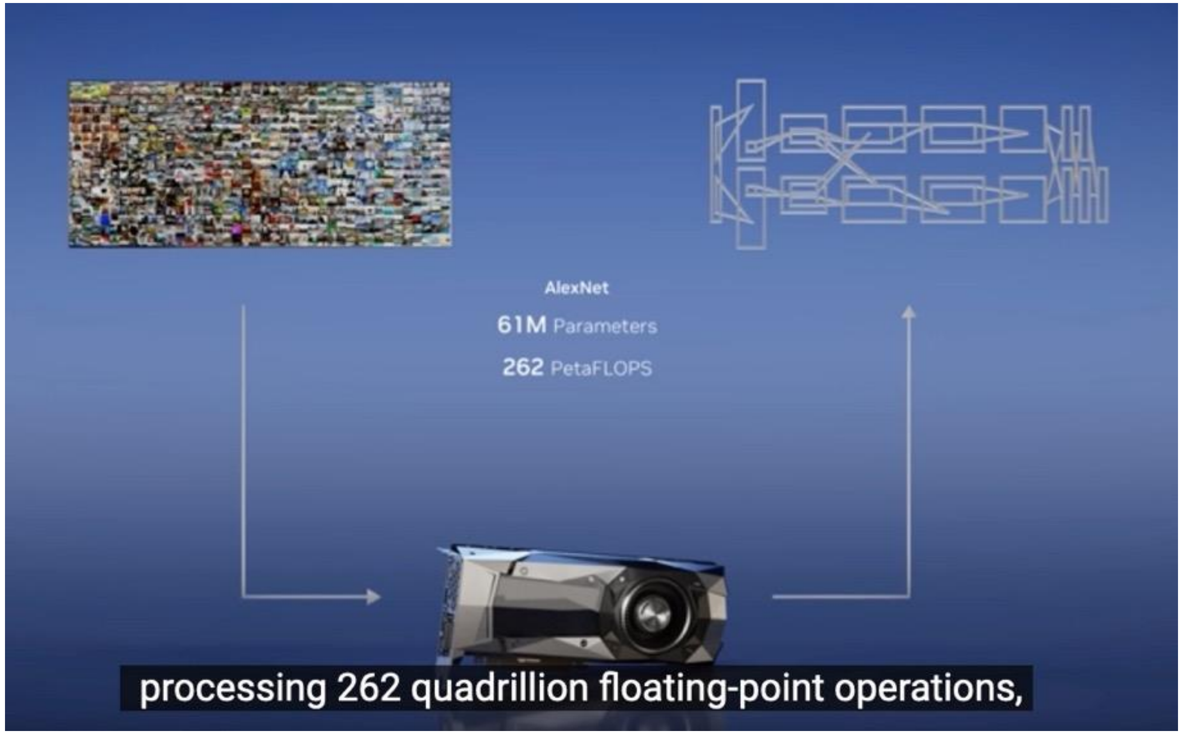


Figure 2. **Outline of the *DeepFace* architecture**. A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate feature maps produced at each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

From:Y. Taigman, M. Yang, M.A. Ranzato, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification"

Not possible with 1990's hardware
Philips L-Neuro 2.3 "NPU"

Computing power is driving the advance of AI



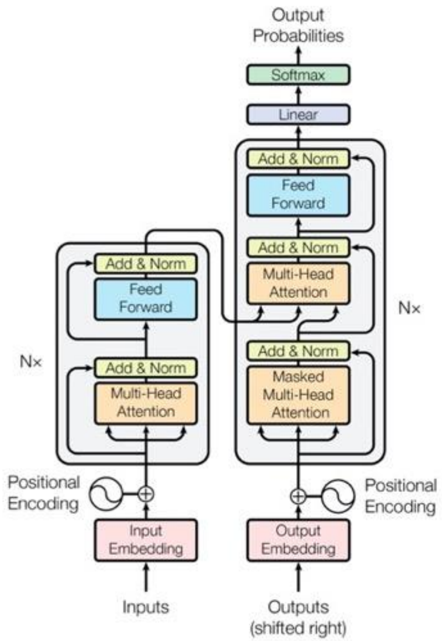
2012: AlexNet
GeForce GTX 580
Won ImageNet Challenge
262 x 10^{15} FLOPS

2020: GPT-3
323 x 10^{21} FLOPS
X 1 000 000 more floating point operations

From GTC 2023 Keynote with NVIDIA CEO Jensen Huang

The origin of LLMs: Transformers (2017)

We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Evolution of Generative Pre-trained Transformers (GPT) in OpenAI

Model	Architecture	Parameter count	Training data	Release date	Training cost
GPT-1	12-level, 12-headed Transformer decoder (no encoder), followed by linear-softmax.	117 million	BookCorpus: 4.5 GB of text, from 7000 unpublished books of various genres.	June 11, **2018**	"1 month on 8 GPUs", or 1.7e19 FLOP.
GPT-2	GPT-1, but with modified normalization	1.5 billion	WebText: 40 GB of text, 8 million documents, from 45 million webpages upvoted on Reddit.	February 14, **2019** (initial/limited version) and November 5, 2019 (full version)	"tens of petaflop/s-day", or 1.5e21 FLOP.
GPT-3	GPT-2, but with modification to allow larger scaling	175 billion	499 Billion tokens consisting of CommonCrawl (570 GB), WebText, English Wikipedia, and two books corpora (Books1 and Books2).	May 28, **2020**	3640 petaflop/s-day, or 3.2e23 FLOP.
GPT-3.5	Undisclosed	175 billion	Undisclosed	March 15, **2022**	Undisclosed
ChatGPT	Undisclosed	? (rumor 20M???)		**November 20, 2022**	
GPT-4	Also trained with both text prediction and RLHF; accepts both text and images as input. Further details are not public.	Undisclosed (1.8 trillon aka 1.8e12)	Undisclosed (13 trillon tokens, aka 1.3e13)	March 14, **2023**	Undisclosed. Estimated 2.1e25 FLOP.

From https://en.wikipedia.org/wiki/Generative_pre-trained_transformer

Compute requirement

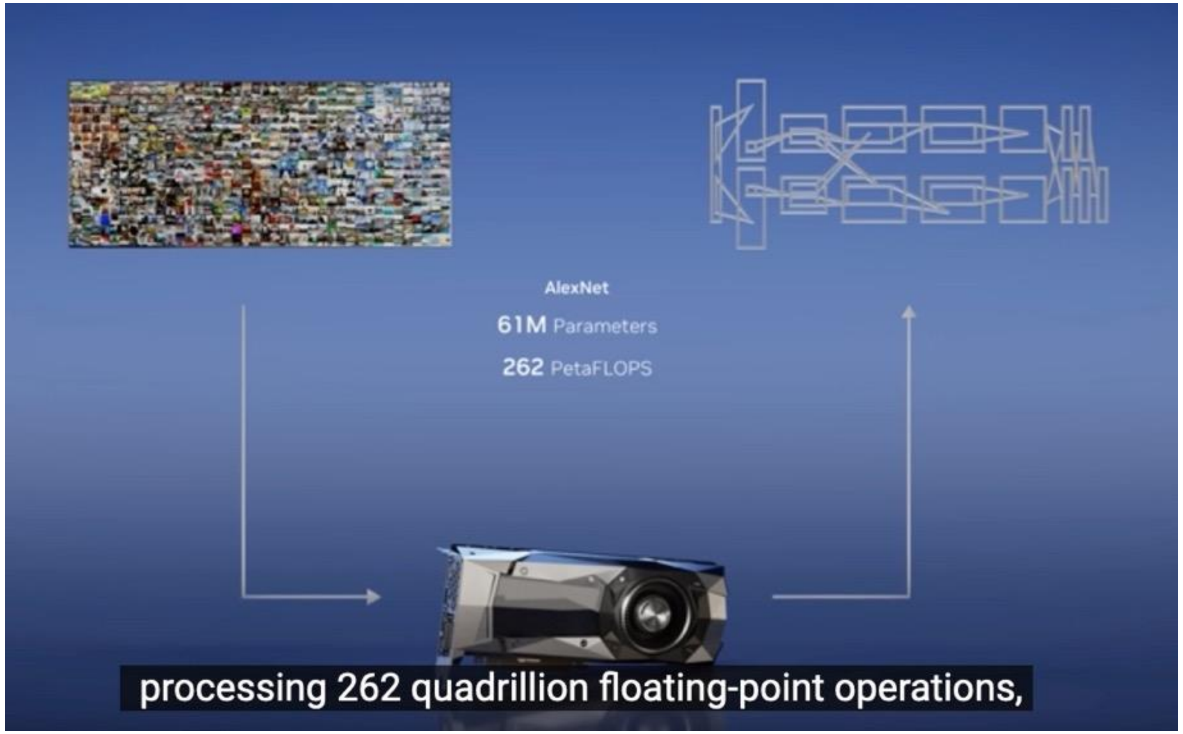
~ x 88

~ x 213

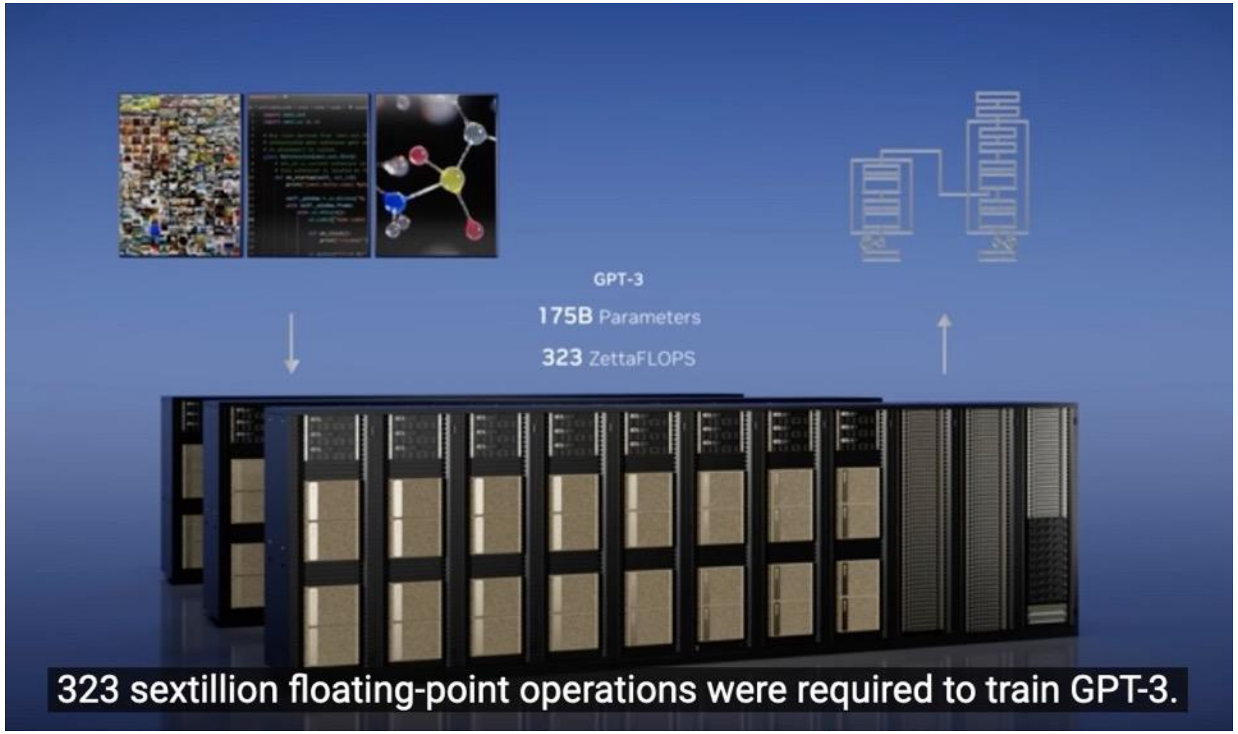
~ x 65

~ x 1 218 360

Computing power is driving the advance of AI



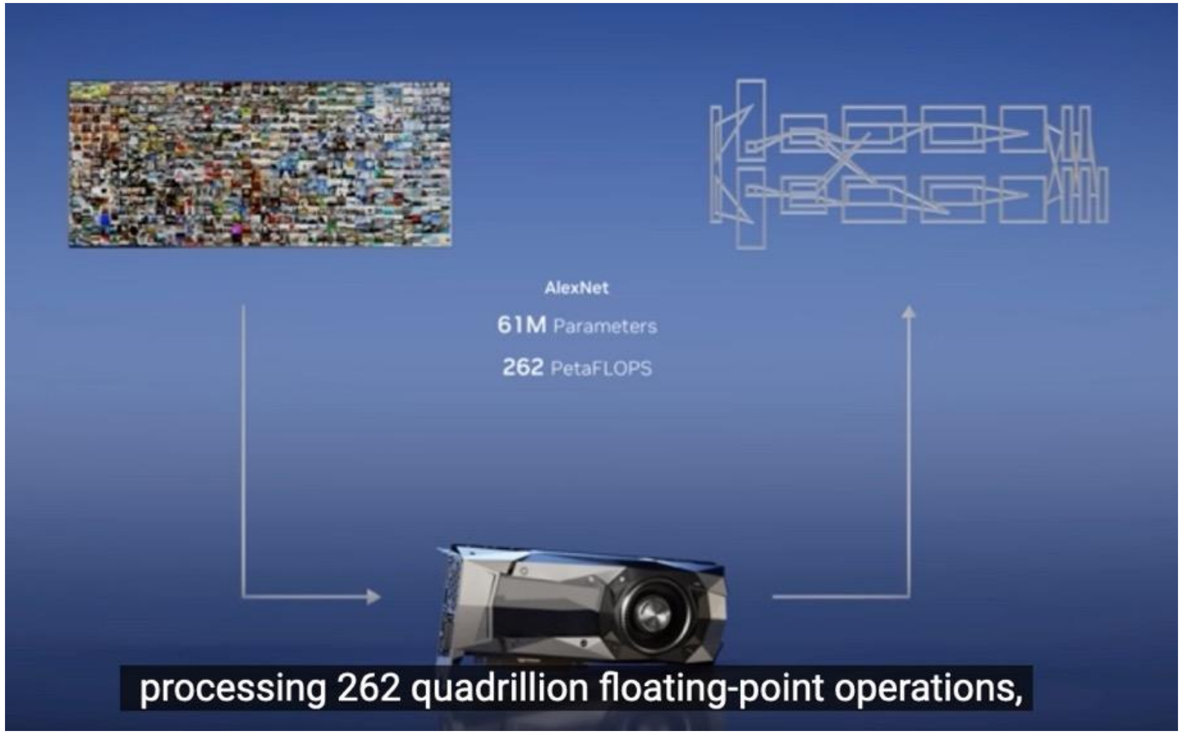
2012: AlexNet
GeForce GTX 580
Won ImageNet Challenge
262×10^{15} FLOPS



2020: GPT-3
323×10^{21} FLOPS
X 1 000 000 more floating point operations

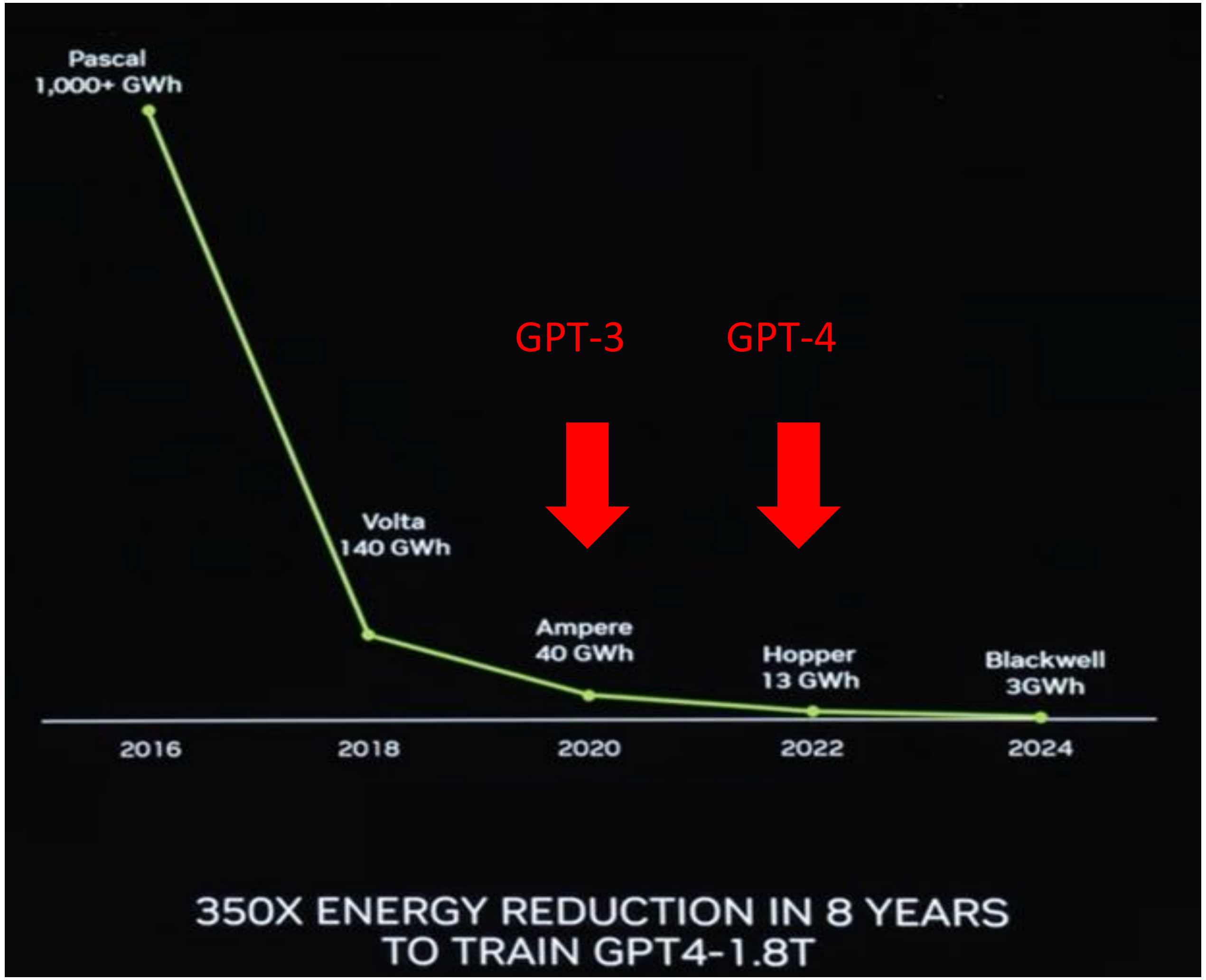
From GTC 2023 Keynote with NVIDIA CEO Jensen Huang

Computing power is driving the advance of AI



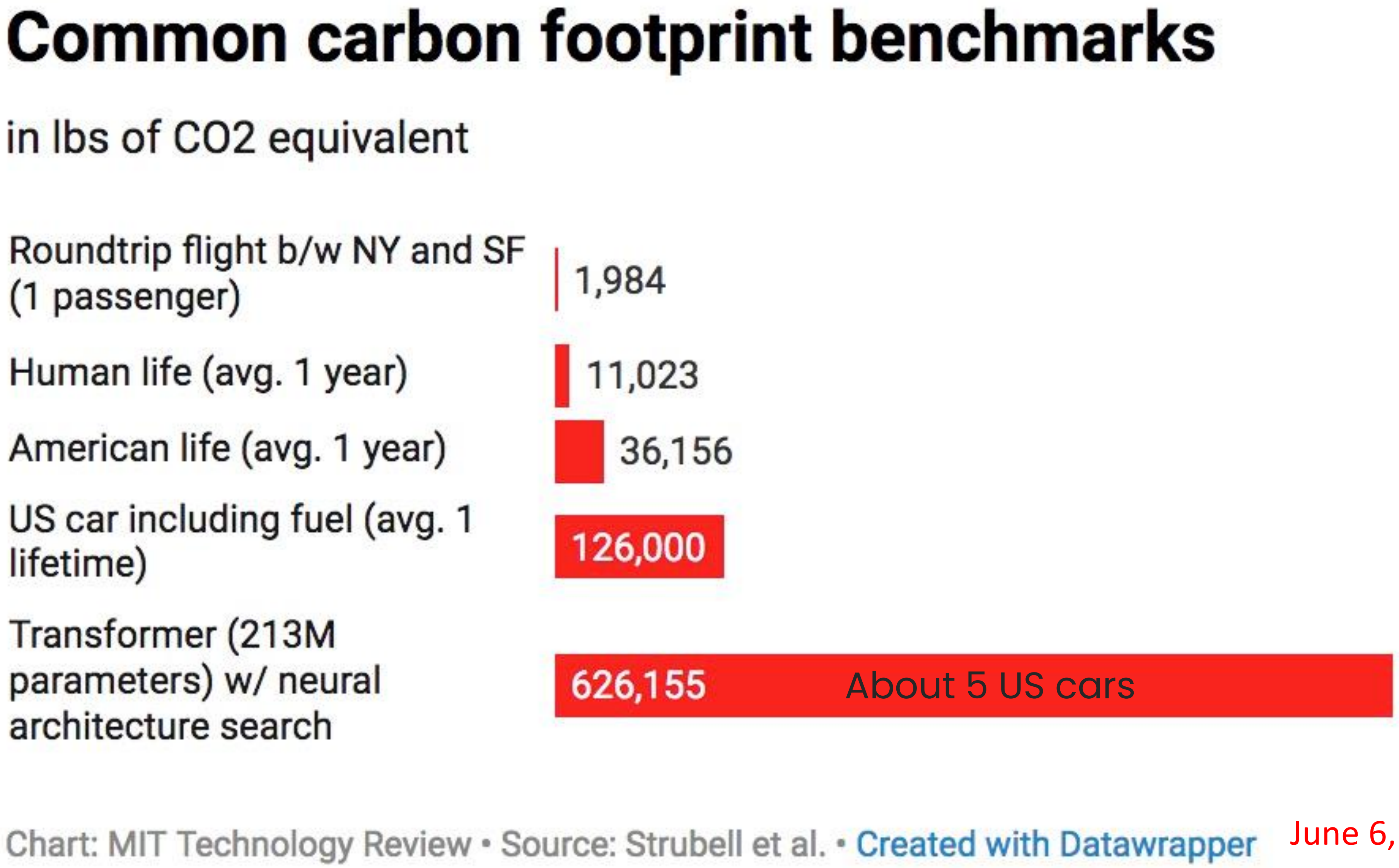
2012: AlexNet
GeForce GTX 580
Won ImageNet Challenge
262 x 10^{15} FLOPS

From GTC 2023 Keynote with NVIDIA CEO Jensen Huang



Cost of energy for training is a limiting factor!

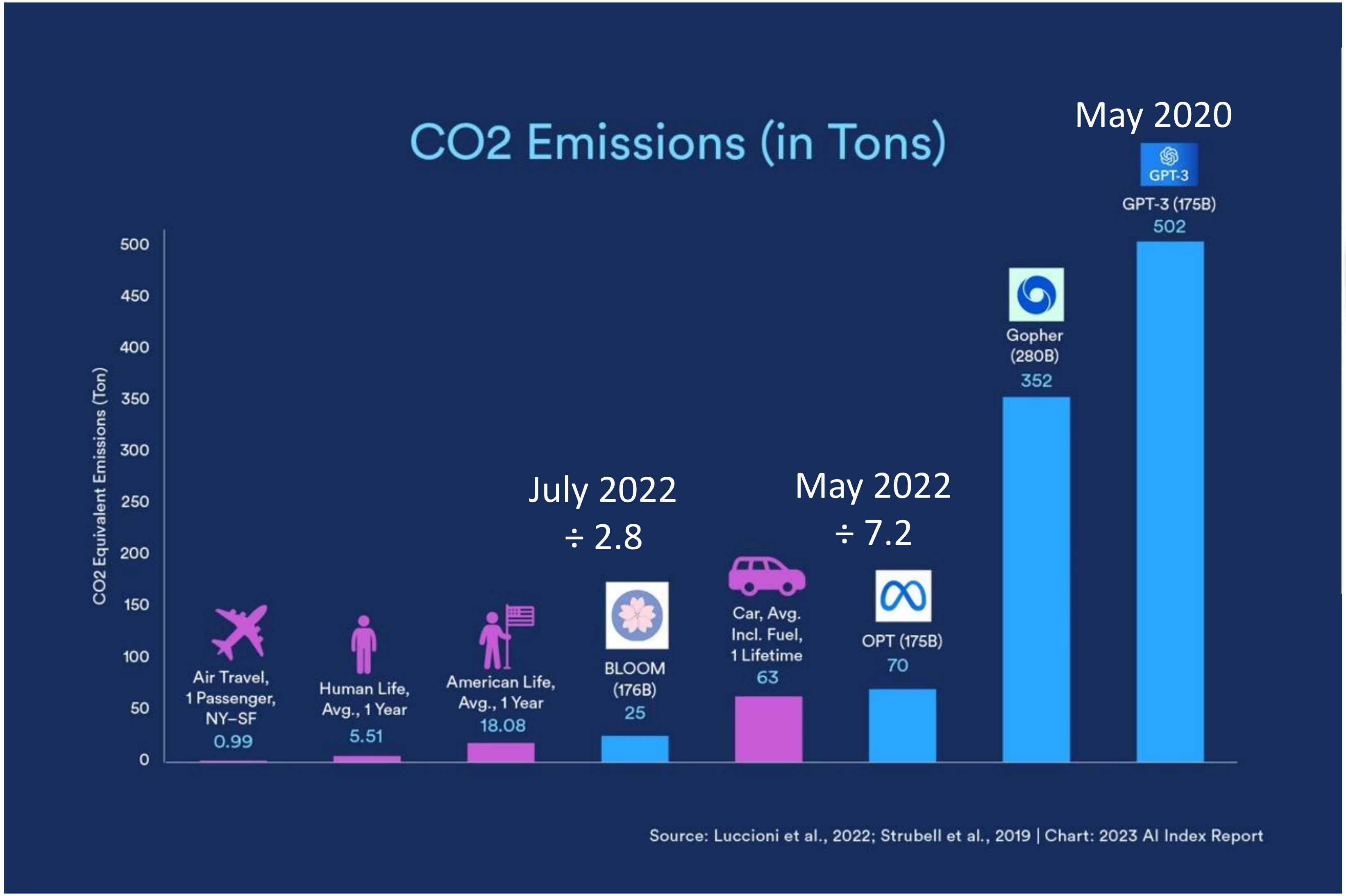
CO_2 impact of training Transformer based models



June 6, 2019

From https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Training Large Language Models has an ecological impact



From https://www.hipeac.net/vision/#/latest/

From “2023 State of AI in 14 Charts” available at https://hai.stanford.edu/news/2023-state-ai-14-charts

One of the early Open Source LLM (March-July 2022)

BigScience

a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

BLOOM: open-source alternative to GPT-3

https://bigscience.huggingface.co

https://huggingface.co/bigscience/bloom

1.5TB of text, 350B tokens

43 languages, 16 programming languages

118 days of training on 384 A100 GPUs

More details at https://huggingface.co/blog/bloom-megatron-deepspeed

Smaller versions are available : 560M, 1.1B, 1.7B, 3B, 7.1B

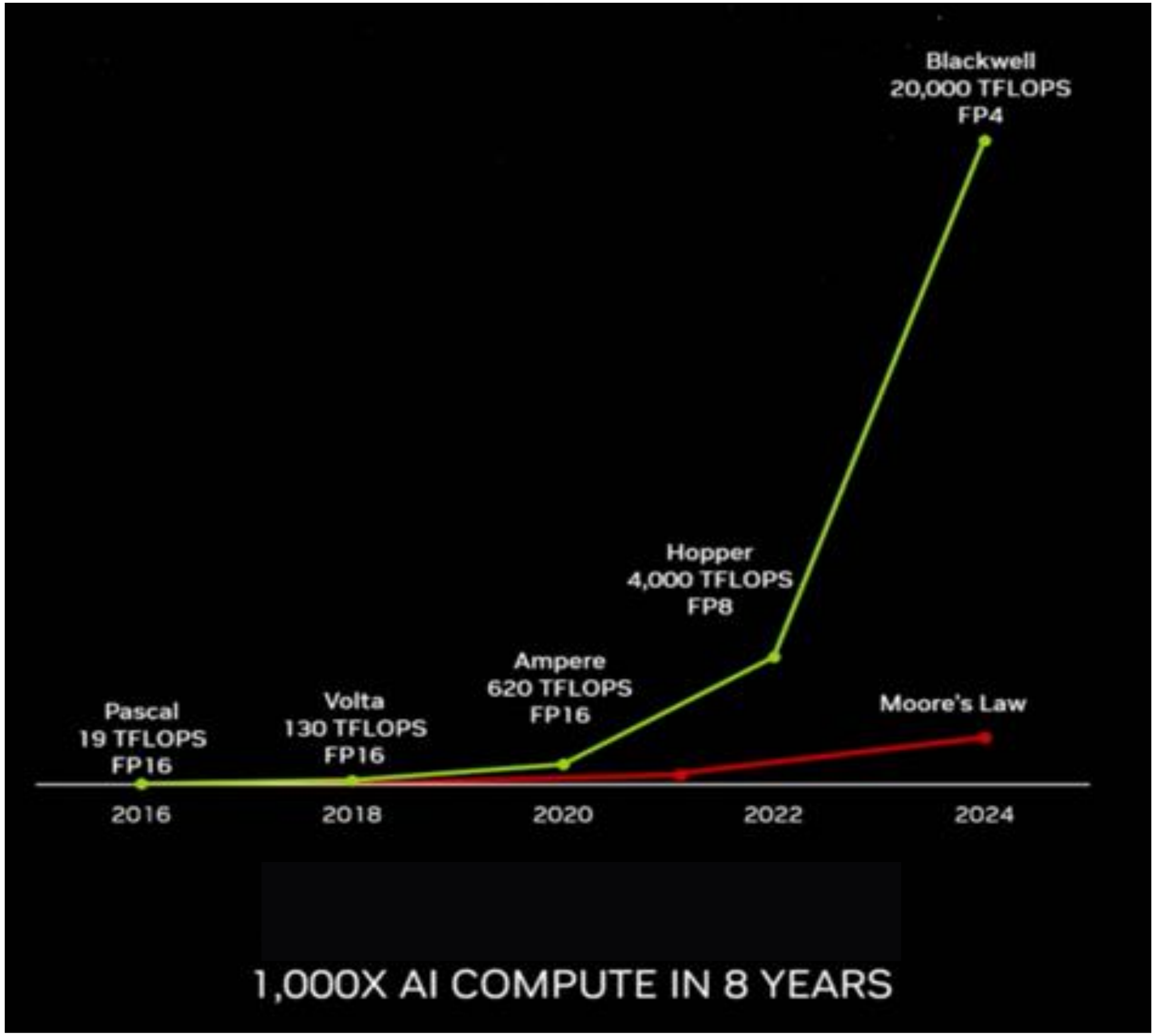
BLOOMZ models (same sizes) are fine-tuned for **instruction following**
https://huggingface.co/bigscience/bloomz

Estimated cost of training: Equivalent of $2-5M in cloud
Server training location: Île-de-France, France
Environmental Impact: The training supercomputer, Jean Zay, uses mostly nuclear energy. The heat generated by it is reused for heating campus housing.

… Then what can be done to have a more sustainable generative Artificial Intelligence?

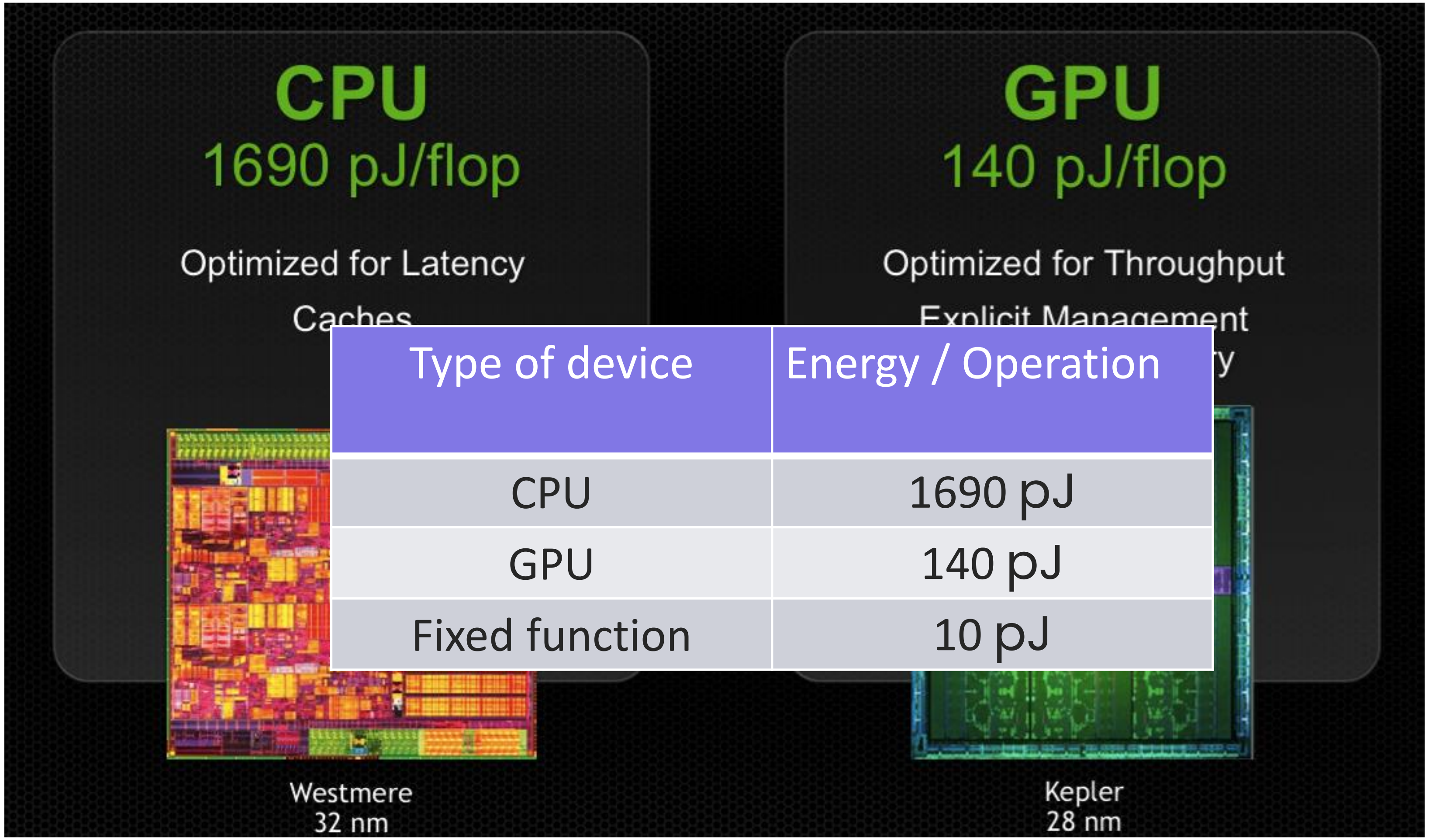
Exponential increase of AI performances

Thanks to advances in architecture and data coding



From Nvidia, Computex 2024

Specialized architectures leads to more efficiency



Type of device	Energy / Operation
CPU	1690 pJ
GPU	140 pJ
Fixed function	10 pJ

From Bill Dally (nVidia) « Challenges for Future Computing Systems »
HiPEAC conference 2015

Gain ~ 150

Deep learning and voice recognition form Google: drive for the TPU design

" The need for TPUs really emerged about six *(13*)* years ago, when we started using computationally expensive deep learning models in more and more places throughout our products. The computational expense of using these models had us worried. If we considered a scenario where **people use Google voice search for just three minutes a day** and we ran deep neural nets for our speech recognition system on the processing units we were using, **we would have had to *double the number of Google data centers*!"**

[https://cloudplatform.googleblog.com/2017/04/quantifying-the-performance-of-the-TPU-our-first-machine-learning-chip.html]

* In 2024

2017: Google's Customized hardware…

… required to increase energy efficiency

with accuracy adapted to the use (e.g. float 16)

Google's TPU2 : 11.5 petaflops$_{16}$ of machine learning number crunching
(and guessing about 400+ KW…, 100+ GFlops$_{16}$/W)

From Google

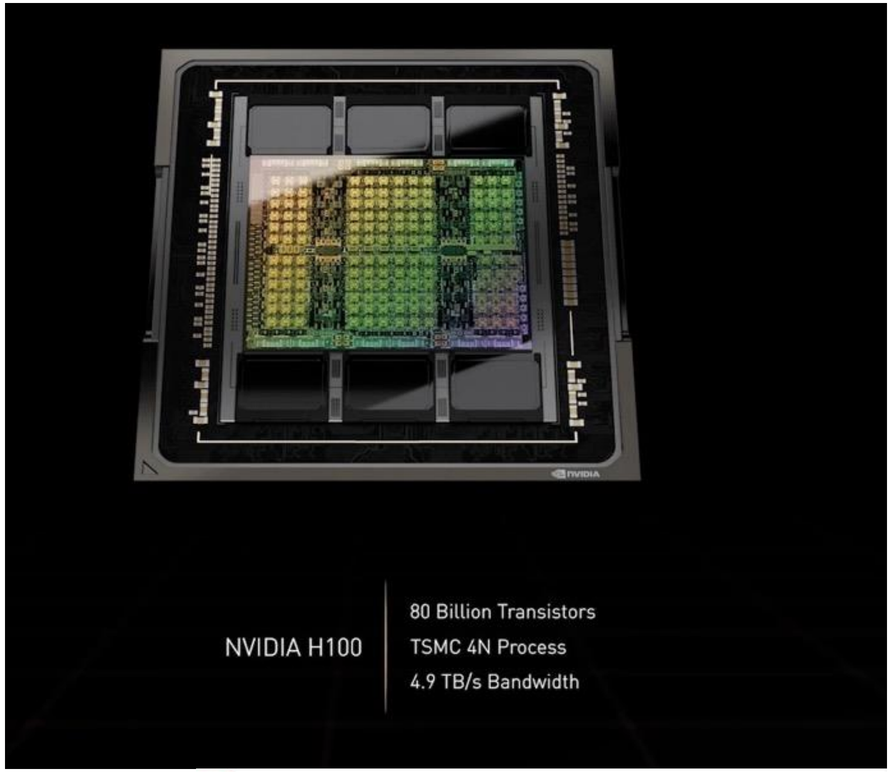
Peta = 10^{15} = million of milliard

2022: NVIDIA H100 GPU

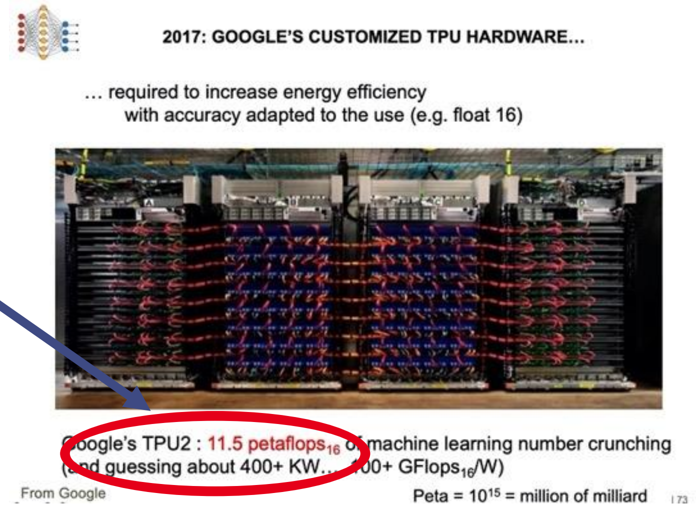
NVIDIA H100 Tensor Core GPU Preliminary Performance Specs

	NVIDIA H100 SXM5[1]	NVIDIA H100 PCIe[1]
Peak FP64[1]	30 TFLOPS	24 TFLOPS
Peak FP64 Tensor Core[1]	60 TFLOPS	48 TFLOPS
Peak FP32[1]	60 TFLOPS	48 TFLOPS
Peak FP16[1]	120 TFLOPS	96 TFLOPS
Peak BF16[1]	120 TFLOPS	96 TFLOPS
Peak TF32 Tensor Core[1]	500 TFLOPS | 1000 TFLOPS[2]	400 TFLOPS | 800 TFLOPS[2]
Peak FP16 Tensor Core[1]	1000 TFLOPS | 2000 TFLOPS[2]	800 TFLOPS | 1600 TFLOPS[2]
Peak BF16 Tensor Core[1]	1000 TFLOPS | 2000 TFLOPS[2]	800 TFLOPS | 1600 TFLOPS[2]
Peak FP8 Tensor Core[1]	2000 TFLOPS | 4000 TFLOPS[2]	1600 TFLOPS | 3200 TFLOPS[2]
Peak INT8 Tensor Core[1]	2000 TOPS | 4000 TOPS[2]	1600 TOPS | 3200 TOPS[2]

1. **Preliminary performance estimates for H100 based on current expectations and subject to change in the shipping products**
2. **Effective TFLOPS / TOPS using the Sparsity feature**

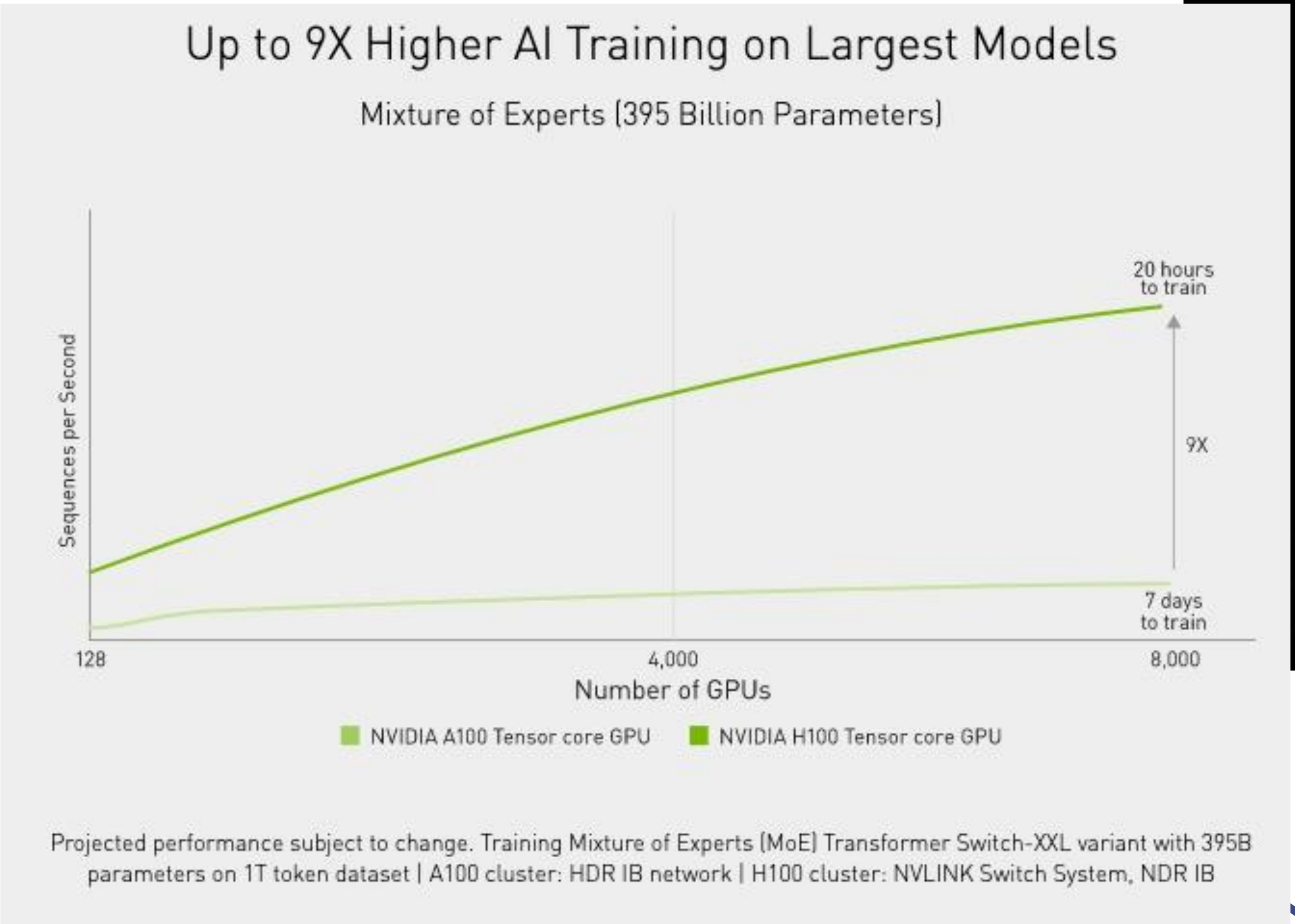


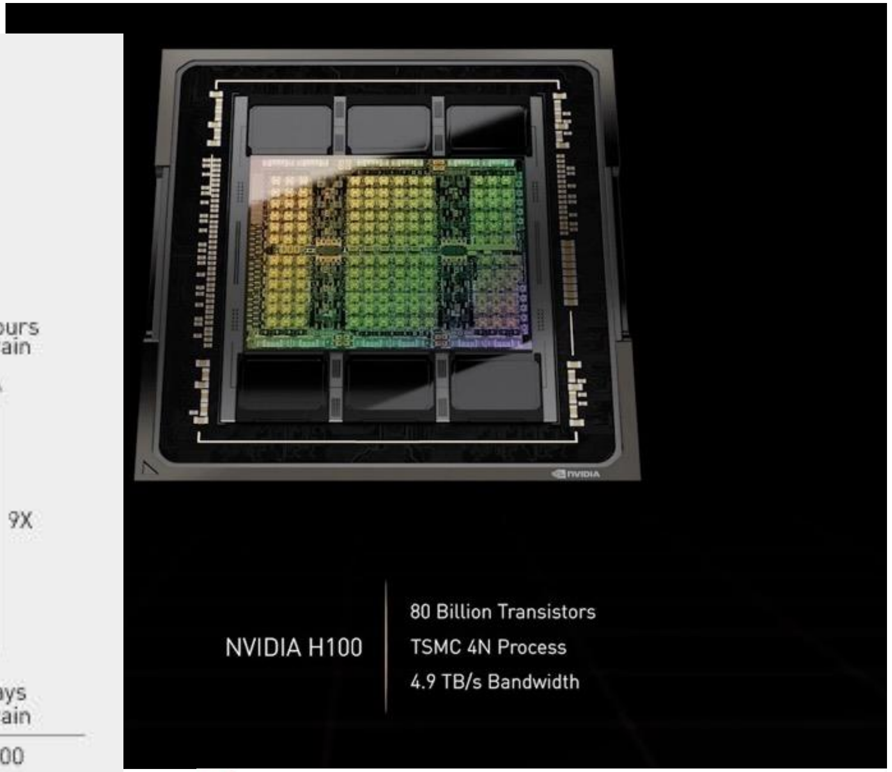
In 5 years: 6 chips!

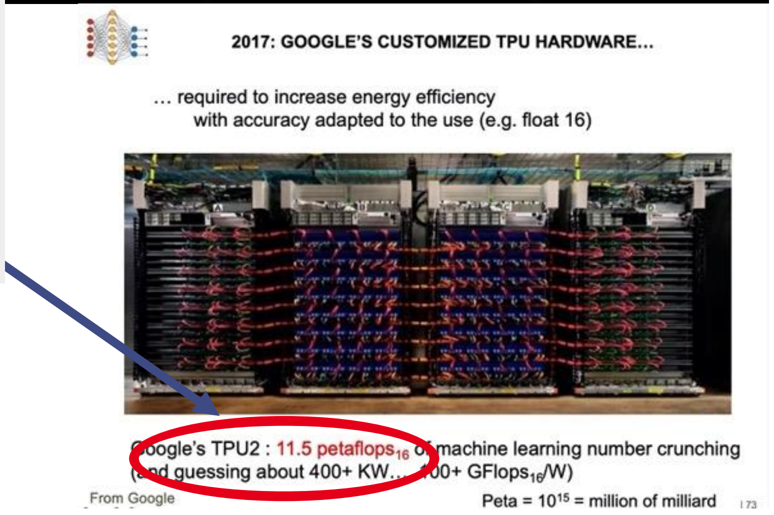


For similar loads, far less hardware, so lower ecological impact

2022: NVIDIA H100 GPU







Smaller LLM models get more powerful, ready for smart devices

- The competition is high to get « small » LLMs with best performances, and there are new techniques emerging everyday.
- Current models of about 10B parameters have similar performances of the original ChatGPT (2 years ago)

Model name	Announced	MMLU benchmark*
ChatGPT (gtp-3.5-turbo)	November 2022	70
GPT-4 (gpt-4-0314)	March 2023	86.4
GPT-4o	May 2024	88.7
o1	September 2024	92.3
Gemma 2 9B	September 2024	71.3
Pixtral-12B	September 2024	69.2

- Fine tuning « small » LLMs to specialize them can lead to very good performances

*Massive Multitask Language Understanding

Structure of Apple Intelligence: already a continuum of computing

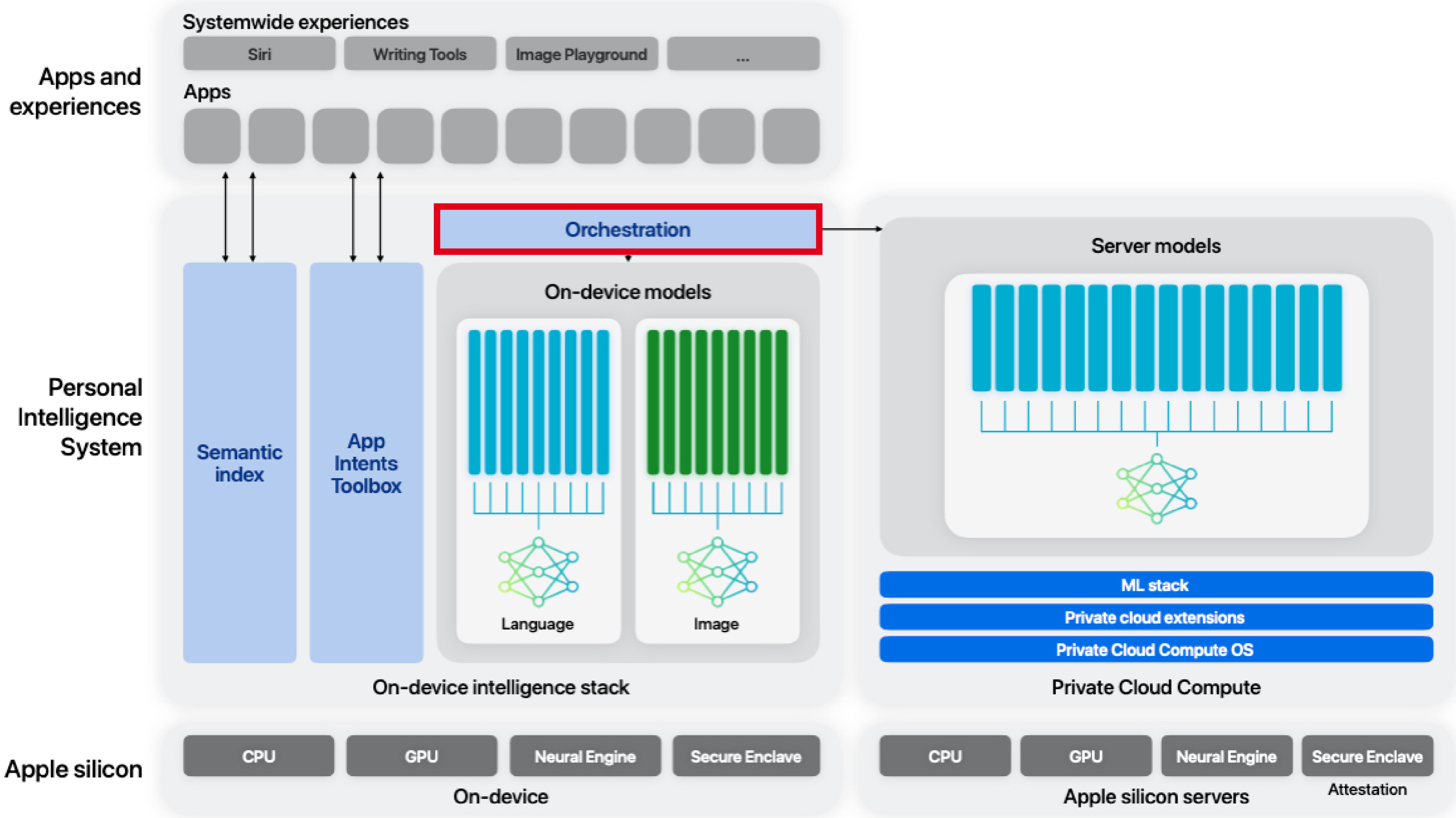


Figure 2: Architecture of Apple Intelligence with adapters for the language on-device and server models and the image models. In this report we are only describing the text models.

From: https://arxiv.org/abs/2407.21075

NPU for LLM running locally on your smartphone (2024)

From https://www.mediatek.com/products/smartphones-2/mediatek-dimensity-9300

November 6th, 2023

« the APU 790 can run a 7 billion parameter LLM at 20 tokens per second, which is fast enough for real-time use. …For comparison, Qualcomm says its Snapdragon 8 Gen 3 can run a 10 billion parameter LLM at almost 15 tokens per second, which seems fairly comparable. The Dimensity 9300 can extend this to run a 13 billion LLM within 16GB of RAM, right up to 33 billion parameters with 24GB RAM, albeit with a much slower 3-4 tokens per second processing rate. »*

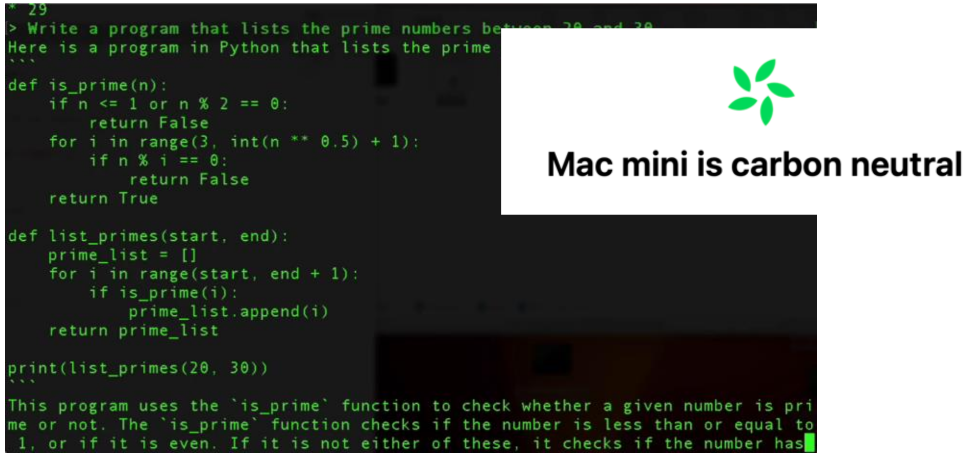
« the APU 790 supports INT4 (A16W4) to run smaller quantized models and a dedicated hardware memory decompression block that feeds the APU. In MediaTek's example, a 13GB INT8 model can be pre-compressed to just 5GB to fit into RAM and then decompressed in hardware on its way to the APU. »*

Support for NeuroPilot Fusion, which can continuously perform LoRA low-rank adaptation

* From https://www.androidauthority.com/mediatek-dimensity-9300-explained-3381678/

Processing at the edge with less and less energy

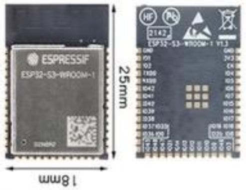
LLM running locally on Mac mini: about 20W



Consumer devices: about 2 to 5W*

Amazon

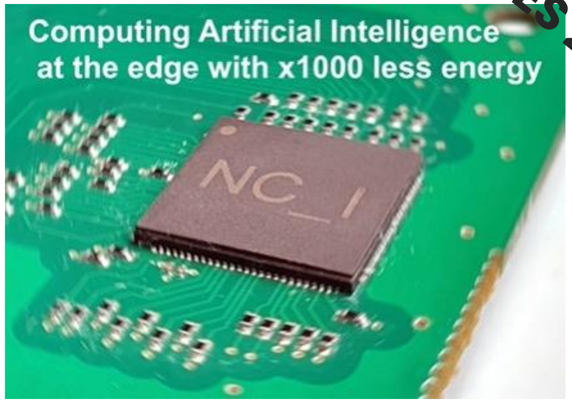
Amazon unveiled the AZ1 Neural Edge processor, a silicon module that will speed up Alexa's ability to answer your queries and commands by hundreds of milliseconds per response. The company built this module alongside MediaTek, and it will allow for on-device neural speech recognition for new products.



About 60 mW

Low cost microcontroller able
to recognize 200 sentences for about 5€

CES 2024



CEA's object detection on
HD images at 30FPS for 23.2mW

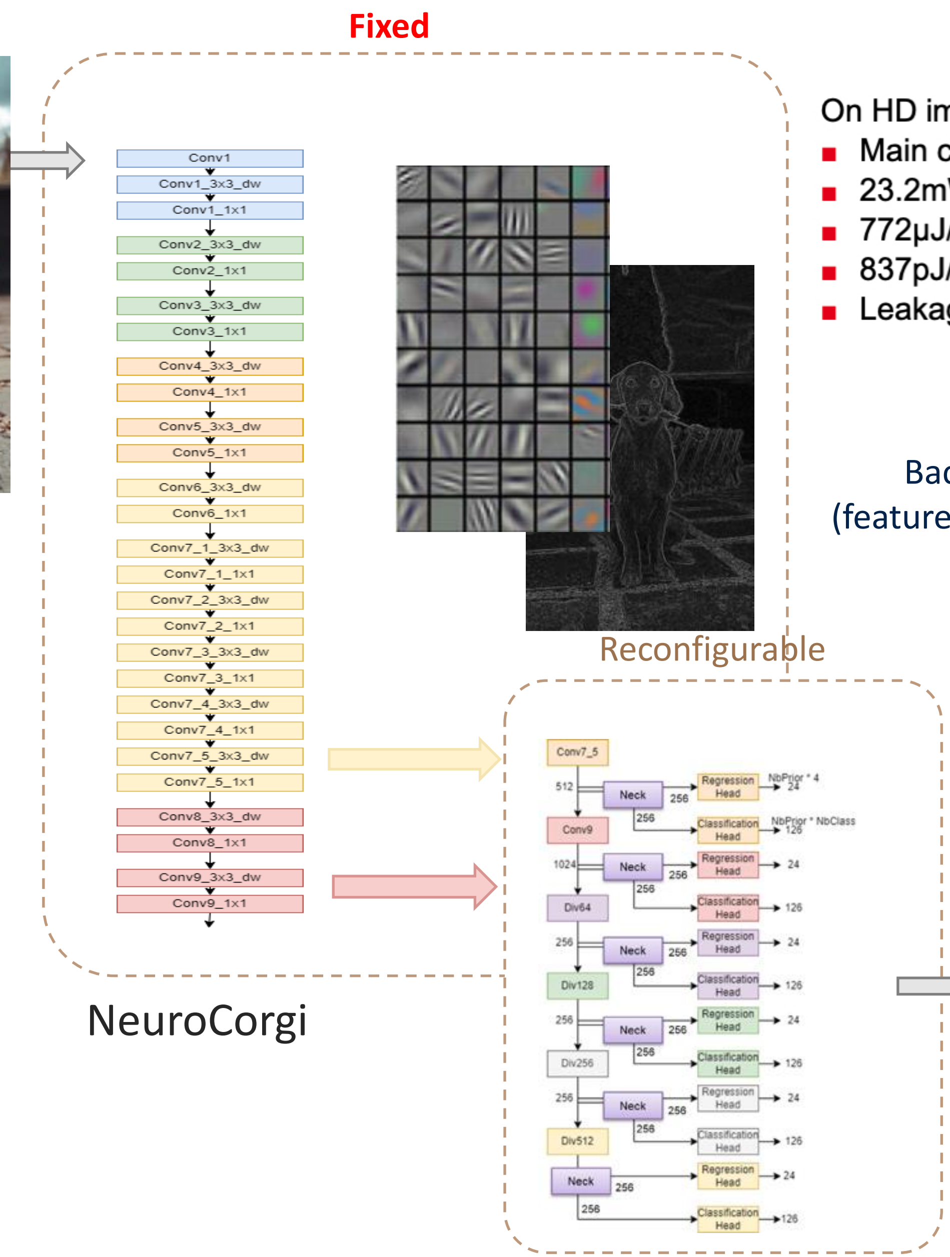
CEA's Ultra-Low Power Neural Processing Unit
ML Commons benchmarks (2022) best in class:

- Keyword spotting: 12 uJ
- Visual Wake Words : 32 uJ (6.4mW)

* Google assistant and Apple Siri can do the same

NeuroCorgi concept and demonstration at CES 2024

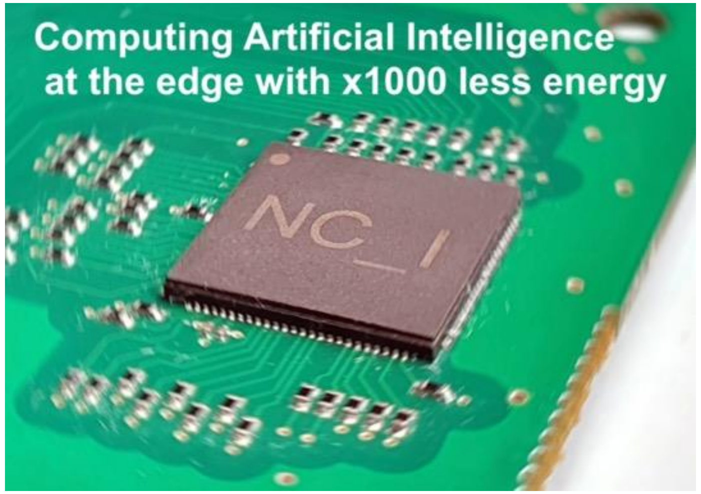
Input



On HD images (1280x720) at 30FPS, 0.76V

- Main clock 59MHz
- 23.2mW @ 30FPS
- 772µJ/frame
- 837pJ/pixel/frame
- Leakage 0.96%: 224µW

Backbone
(features extractor)



Output

Use Case 2.1: Autonomous Weeding System

Use Case 3.1: Drones/USV

Use Case 2.2: Tomato pests and diseases forecast

Use Case 3.2: Underwater Acoustic Signal Classification

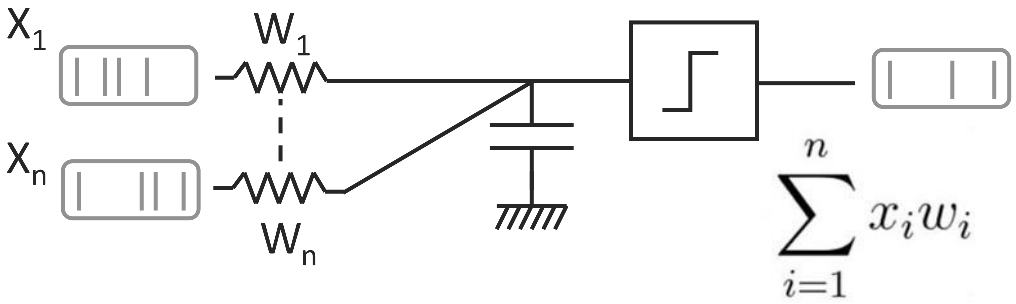
Use Case 3.3: 3D Object Detection and Classification of Road Users

« Neuromorphic » computing

Further energy gain can be achieved thanks to:

- Sparsity of data representation
    - Derivative of the signal : information coding (temporal, transmit and compute only when there is a change in the data)
    - Sparsity of coding: quantization down to 1 bit
- Changing the information coding can simplify operations (no multiply)
    - Exemple: coding in formation in "spikes"
- Using physics to make computations ("analog" computing)
    - Can use electronics, optics, …

- **RRAM synapses**
    - Weighted input thanks to Ohm's law
- **Analog neurons**
    - Inputs summation thanks to Kirchhoff's law



Using physics to compute: development of memristors

PRESS INFORMATION

TDK

Corporate

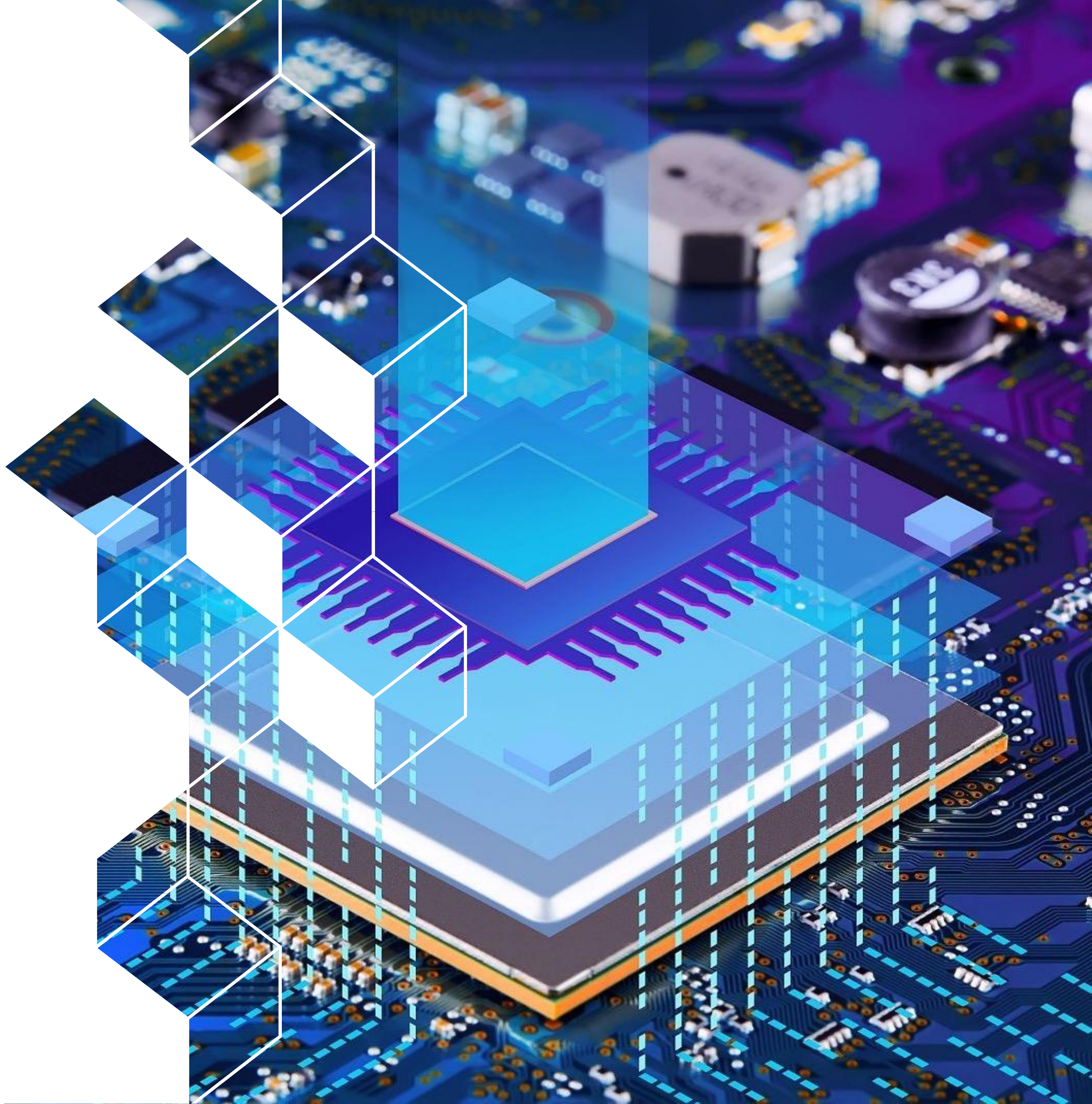
TDK develops "spin-memristor" for neuromorphic devices, and collaborates with CEA and Tohoku University to achieve practical application of neuromorphic devices able to reduce power consumption of AI down to 1/100

- TDK has developed a "spin-memristor" with immunity to environmental influences and long-term data storage
- Collaborating with CEA, the “spin-memristor” has been demonstrated to function as the basic element of a neuromorphic device
- For practical development of the technology, TDK is collaborating with the Center for Innovative Integrated Electronic Systems (CIES) at Tohoku University on prototyping in the semiconductor process
- Development is being pursued through an international collaboration between CEA and Tohoku University

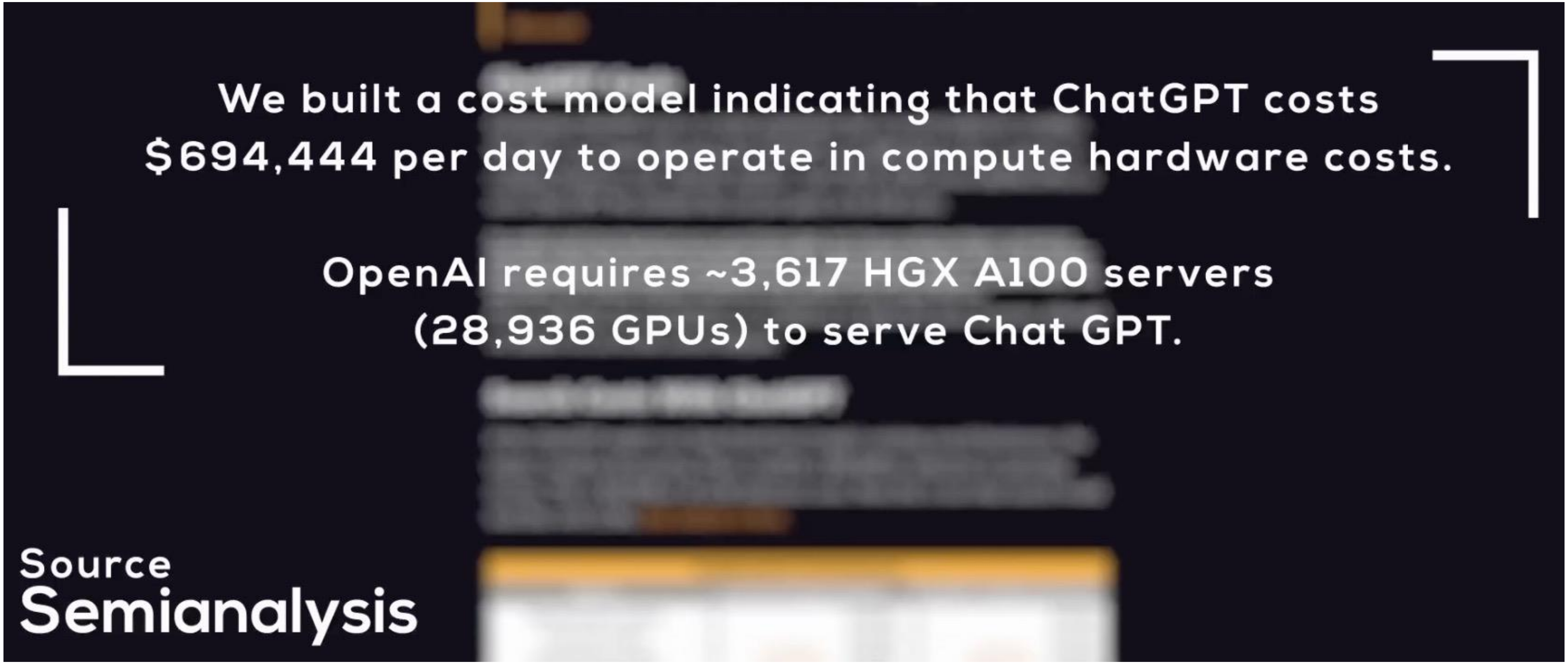
October 2, 2024

Innovation Category Award at CEATEC 2024

Distributed AI, Agentic AI, and the possible next steps…



Inference of LLM is also very demanding (200 M users for OpenAI ?)



Slide from High Yield – Everything Silicon

6.5x more GPUs than for the learning phase

How to reduce the inference cost?

- Inference (using generative AI) is becoming more demanding due to the large number of users
  - "OpenAI says ChatGPT's weekly users have grown to 200 million"*

- Inference is more used by the end user

- **Specialization of hardware for inference** (e.g Groq chip, AWS Inferentia vs AWS Trainium chips, etc)

- Approaches that **don't need to use all the "neurons"** of a LLM:
  - **Mixture of Experts**: MoE architectures create specialized "experts" within a large model, where each expert is optimized to handle certain types of inputs or tasks. **Only a subset** of these experts are activated for each task, promoting a modular structure within a single model.
  - **Agentic AI:** Agentic AI often operates with multiple, distinct agents, each responsible for specialized tasks or competencies. Like MoE, it's structured to have these agents work in tandem or **selectively engage** to perform complex tasks.

* From https://www.reuters.com/technology/artificial-intelligence/openai-says-chatgpts-weekly-users-have-grown-200-million-2024-08-29/

GPT-4 exceptional performances due to its new structure

- **GPT-4's Scale**: GPT-4 has ~1.8 trillion parameters across 120 layers, which is over 10 times larger than GPT-3.
- **Mixture Of Experts (MoE)**: OpenAI utilizes 16 experts within their model, each **with ~111B parameters** for MLP. **Two of these experts are routed per forward pass**, which contributes to keeping costs manageable. (NB: **1/8 of the computation**)
- **Dataset**: GPT-4 is trained on ~13T tokens, including both text-based and code-based data, with some fine-tuning data from ScaleAI and internally.
- **Dataset Mixture**: The training data included CommonCrawl & RefinedWeb, totaling 13T tokens. Speculation suggests additional sources like Twitter, Reddit, YouTube, and a large collection of textbooks.
- **Training Cost**: The training costs for GPT-4 was around $63 million, taking into account the computational power required and the time of training.
- **Inference Cost**: GPT-4 costs 3 times more than the 175B parameter Davinci, due to the larger clusters required and lower utilization rates.
- **Inference Architecture**: The inference runs on a cluster of 128 GPUs, using 8-way tensor parallelism and 16-way pipeline parallelism.
- **Vision Multi-Modal**: GPT-4 includes a vision encoder for autonomous agents to read web pages and transcribe images and videos. The architecture is similar to Flamingo. This adds more parameters on top and it is fine-tuned with another ~2 trillion tokens.

From SemiAnalysis, July 2023

Agentic AI the future of AI?

- Using a set of small specialized LLMs can have similar performances than of a large LLM
- Only a subset of the SLM are activated simultaneously

Agentic AI the future of AI?

- Using a set of small specialized LLMs can have similar performances than of a large LLM
- Only a subset of the SLM are activated simultaneously

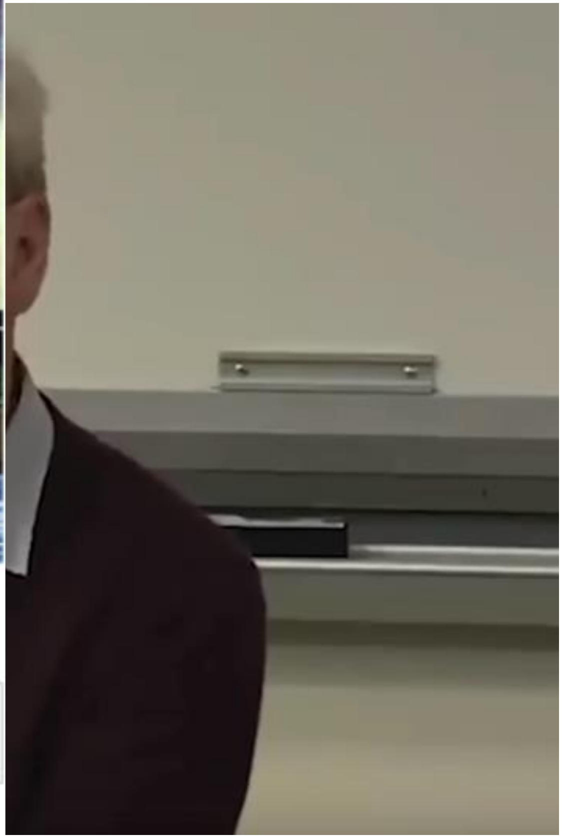
Agentic AI the future of AI?

- Using a set of small specialized LLMs can have similar performances than of a large LLM
- Only a subset of the SLM are activated simultaneously

(Image credit: Getty Images / Justin Sullivan)

Jump to: Read more

Bringing AI agents into the workforce will soon be as common as onboarding human employees, as they work together to make businesses smarter and more efficient, Nvidia CEO Jensen Huang has predicted.



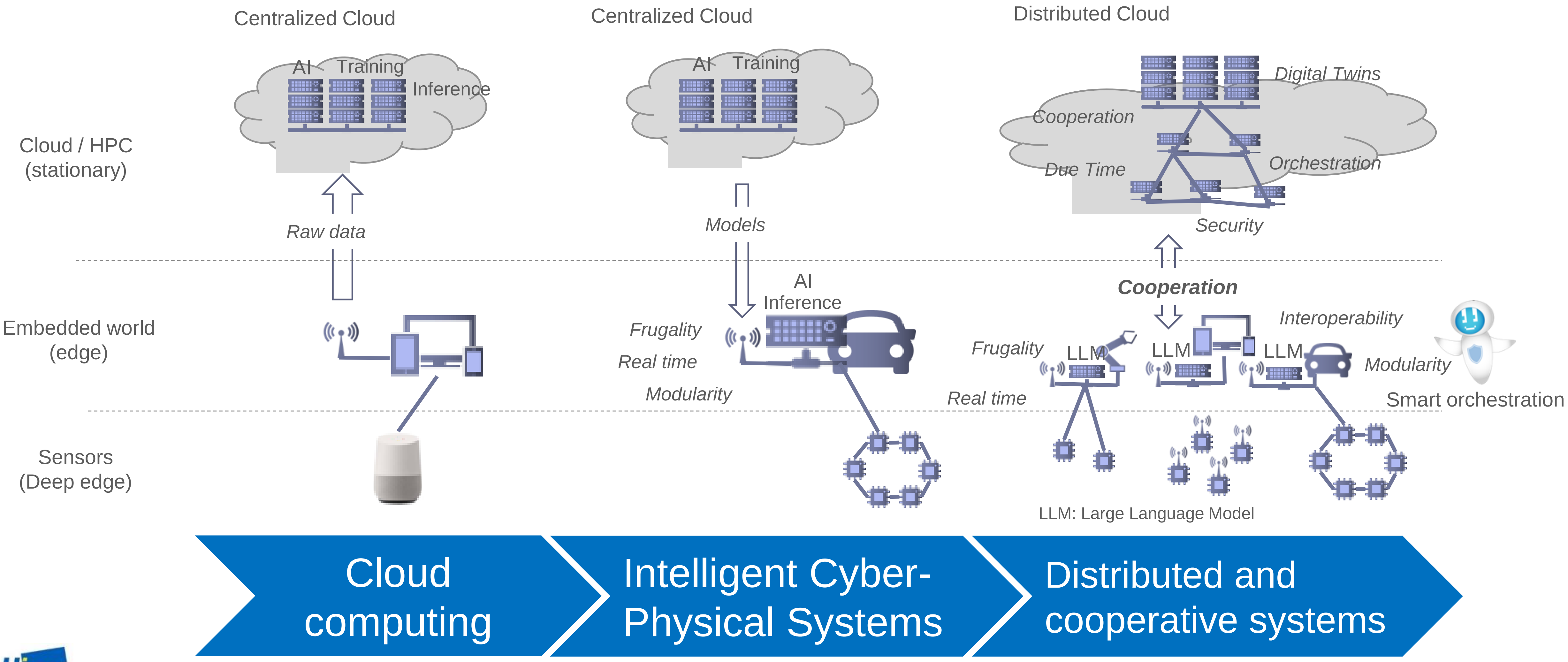
What is the key element of Agentic AI?

The key element in both approaches is the “**router**”, or “**orchestrator**”

- **MoE:** The MoE **router** selects the most appropriate experts based on the input context, enhancing the model’s ability to adapt dynamically to various types of inputs. This routing mechanism is foundational in allowing a large model to focus on the right areas at the right time.
- **Agentic AI**: Similarly, Agentic AI involves a decision-making layer or **"agent manager”, or “orchestrator”** that allocates tasks to the best-suited agents. The manager dynamically routes requests to different agents based on the context or goal, enabling the system to adaptively respond to complex, changing inputs.

Agents can be centralized, or ***distributed***:
- Agents can run on different devices, even “old” ones, increasing lifetime of devices
- If a device is not powerful enough, it can delegate to other devices

Evolution of computing: Cloud, CPS, IoT, AI → Next Computing Paradigm



From Denis Dutoit, inspired from HiPEAC 2023

The next computing paradigm in HiPEAC

https://vision.hipeac.net/the-next-computing-paradigm-ncp--introduction.html

HiPEAC Vision

Search ⌘ K

Introduction

Introduction

The next computing paradigm (NCP)

Introduction
The Spatial Web
Bridging CPS communities
NCP societal aspects

Artificial intelligence

New hardware

Cybersecurity

Sustainability

DOI 10.5281/zenodo.10874349

> Revolutionizing the Digital World: Unveiling the Future of Computing with Spatial Awareness, Generative AI, and Dynamic Web Integrations!

The Next Computing Paradigm: an Introduction

by Tullio Vardanega and Marc Duranton

What will be the future of computing systems (infrastructure, software and hardware)? HiPEAC envisions *the next computing paradigm* (NCP), focusing on a seamless integration of key ingredients from various digital elements like the Web, the Cloud, Cyber-Physical Systems, the Internet of Things, digital twins, the metaverse, and Artificial Intelligence. Envisioning the NCP emphasizes the evolution towards a spatial dimension in computing, a coherent continuum of computing, intertwining the real world with the cyberworld, incorporating Generative AI, and dynamic orchestrations of resources. The aim is to create a seamless, networked cooperative structure where resources are accessed and manipulated with streamlined Web-type protocols, where programs (in fact services) and data flow smoothly onto computing resources that cooperate with each other enhancing context-awareness and efficiency in digital interactions.

More detail in HiPEAC Vision 2024:
https://www.hipeac.net/vision/#/latest/

Thank you for your attention

どうもありがとうございます