AI
IN THE POST-MOORE ERA

Babak Falsafi

EcoCloud

parsa.epfl.ch

EPFL

OUR DIGITAL UNIVERSE

Fueled by:

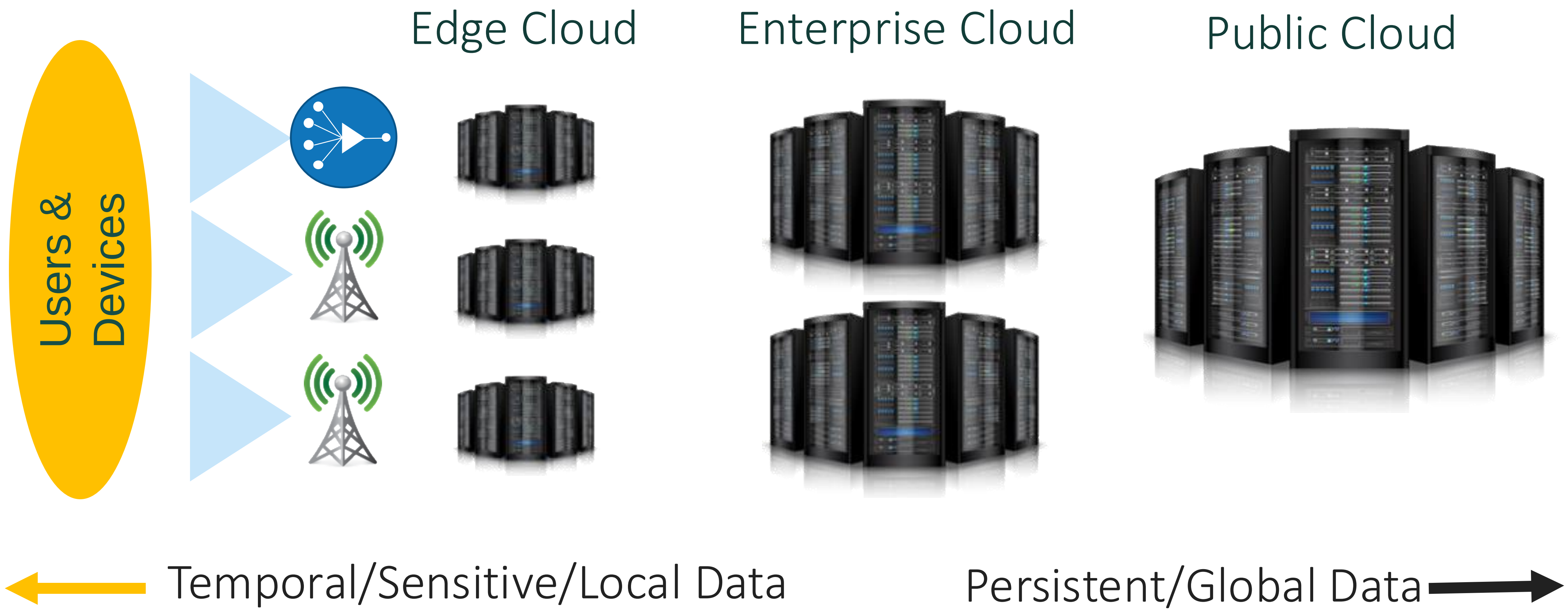
- Data volume
- Data growth rate
- Monetization of data
- Data's impact on GDP
- ….now AI

DATACENTERS:
THE BACKBONE OF OUR DIGITAL UNIVERSE

- 100s of thousands of commodity or home-brewed servers
  - Consuming 10s to 100s MW
- Centralized to exploit economies of scale
- Network fabric w/ μ-second connectivity
- Often limited by ingress
  - Electricity
  - Network
  - Cooling

Boydton DC, 300MW

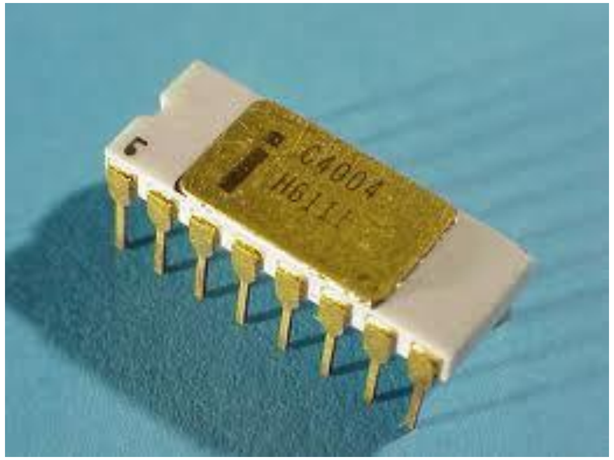
CLOUDS AT VARIOUS SCALES



UNIVERSE MADE POSSIBLE BY MOORE'S LAW

1971

Intel 4004

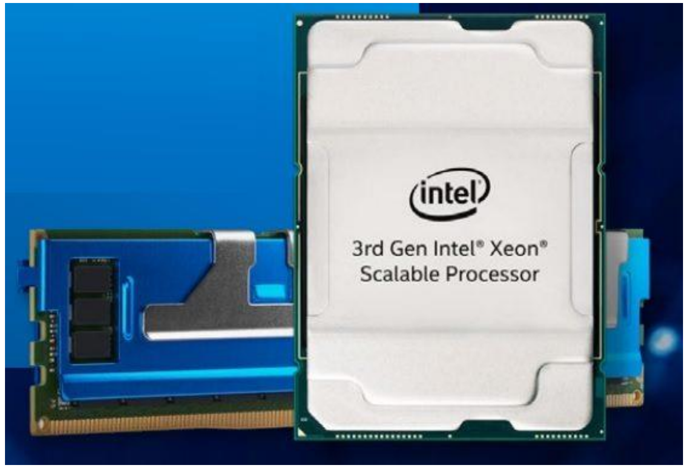


92,000 ops/s

1 Watt

2021

Intel Ice Lake



1,200,000,000,000 ops/s

270 Watts

MOORE'S LAW: EXPONENTIAL DENSITY & EFFICIENCY

1971	2021
Intel 4004	Intel Ice Lake

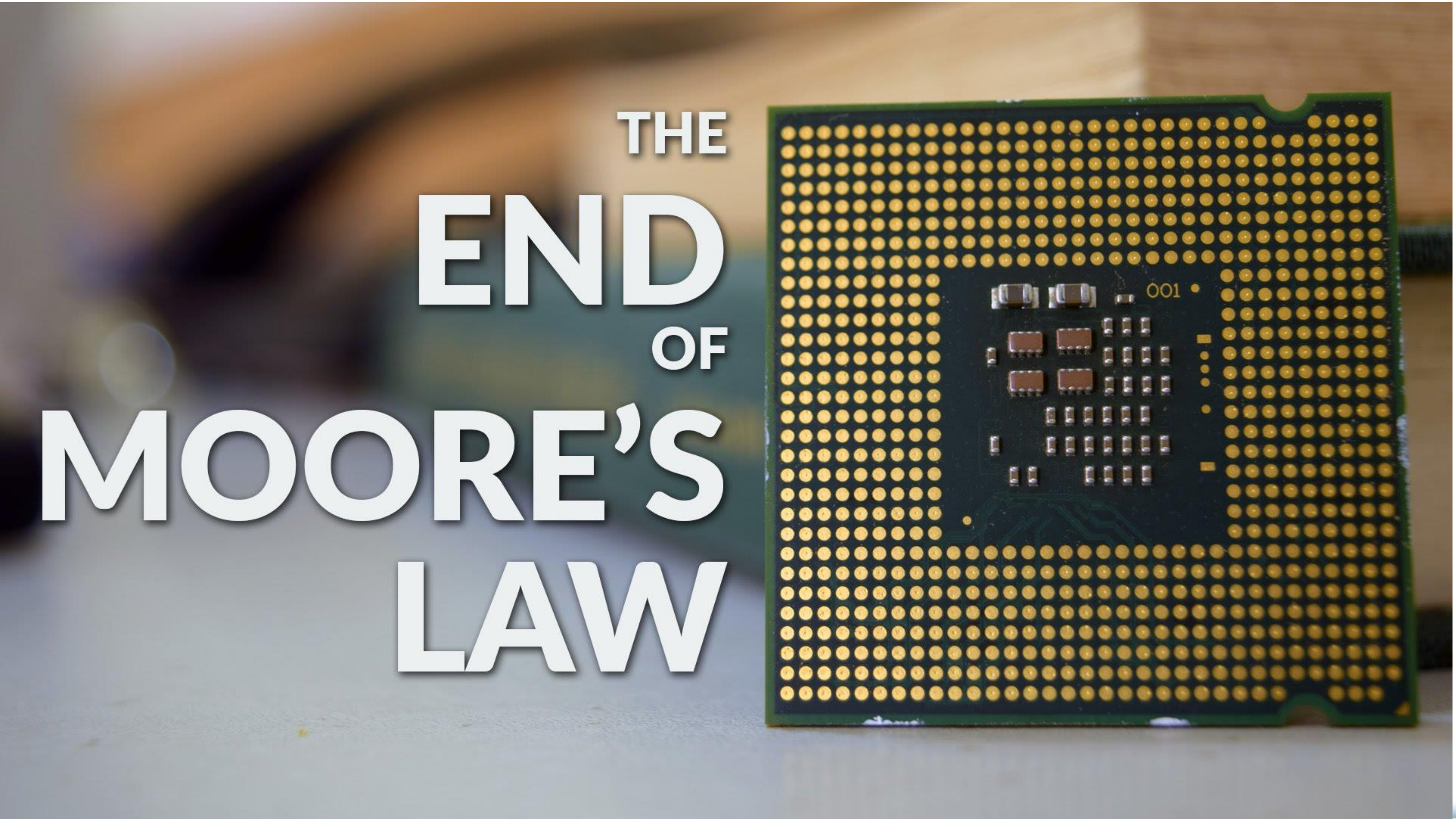
In 50 years:

13 million times faster

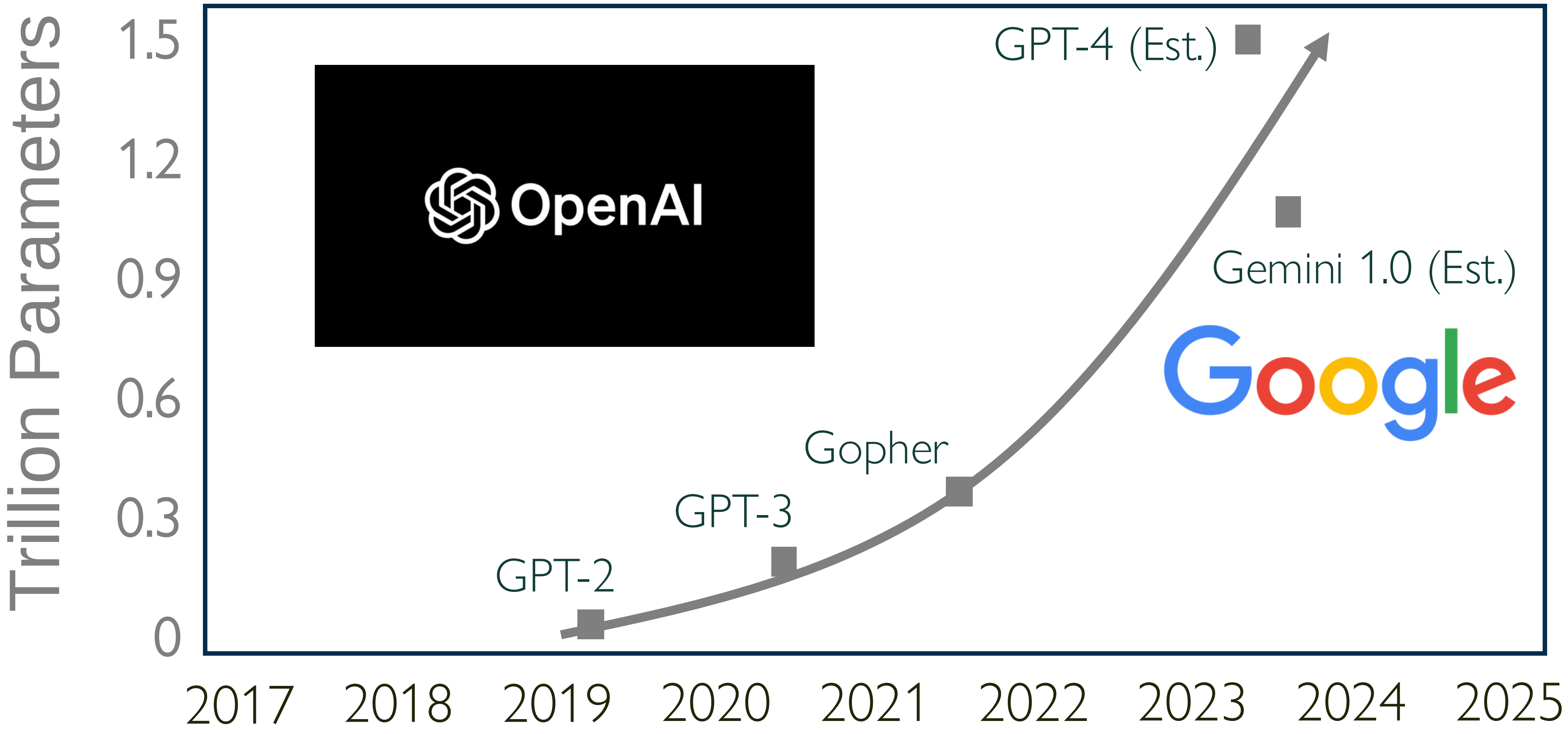
48 thousand times more efficient

1971	2021
92,000 ops/s	1,200,000,000,000 ops/s
1 Watt	270 Watts

LONG LIVE MOORE'S LAW

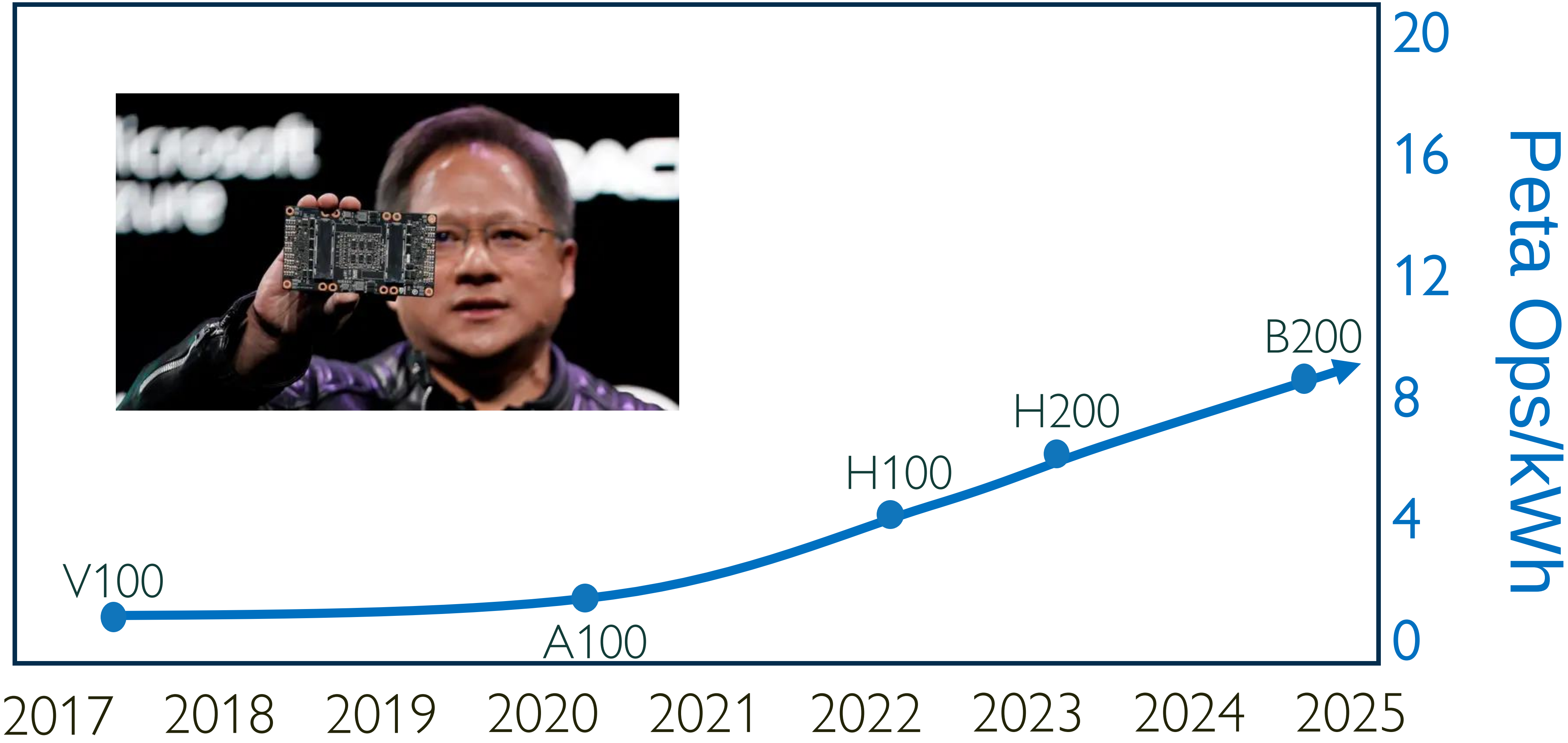


LLMS' GROWTH

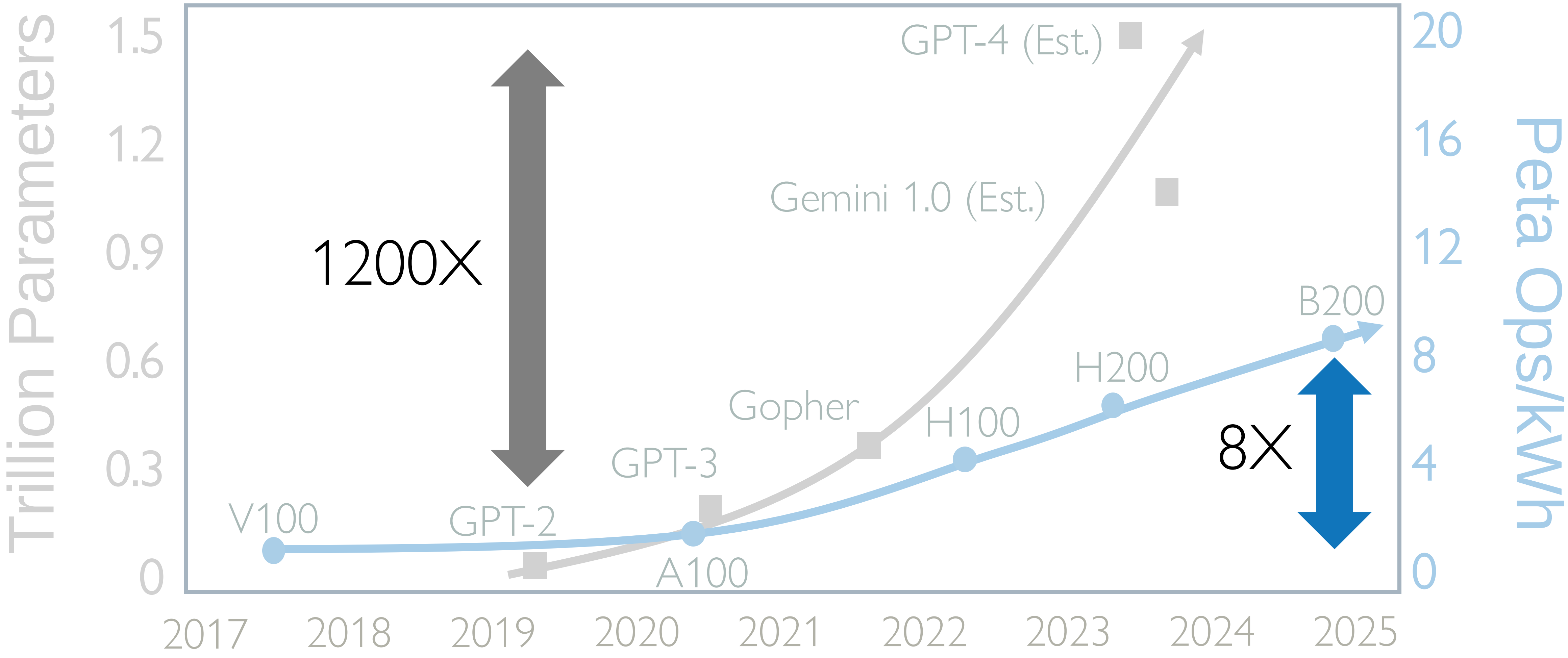


Trillion Parameters
1.5
1.2
0.9
0.6
0.3
0
OpenAI
GPT-4 (Est.)
Gemini 1.0 (Est.)
Google
Gopher
GPT-3
GPT-2
2017
2018
2019
2020
2021
2022
2023
2024
2025

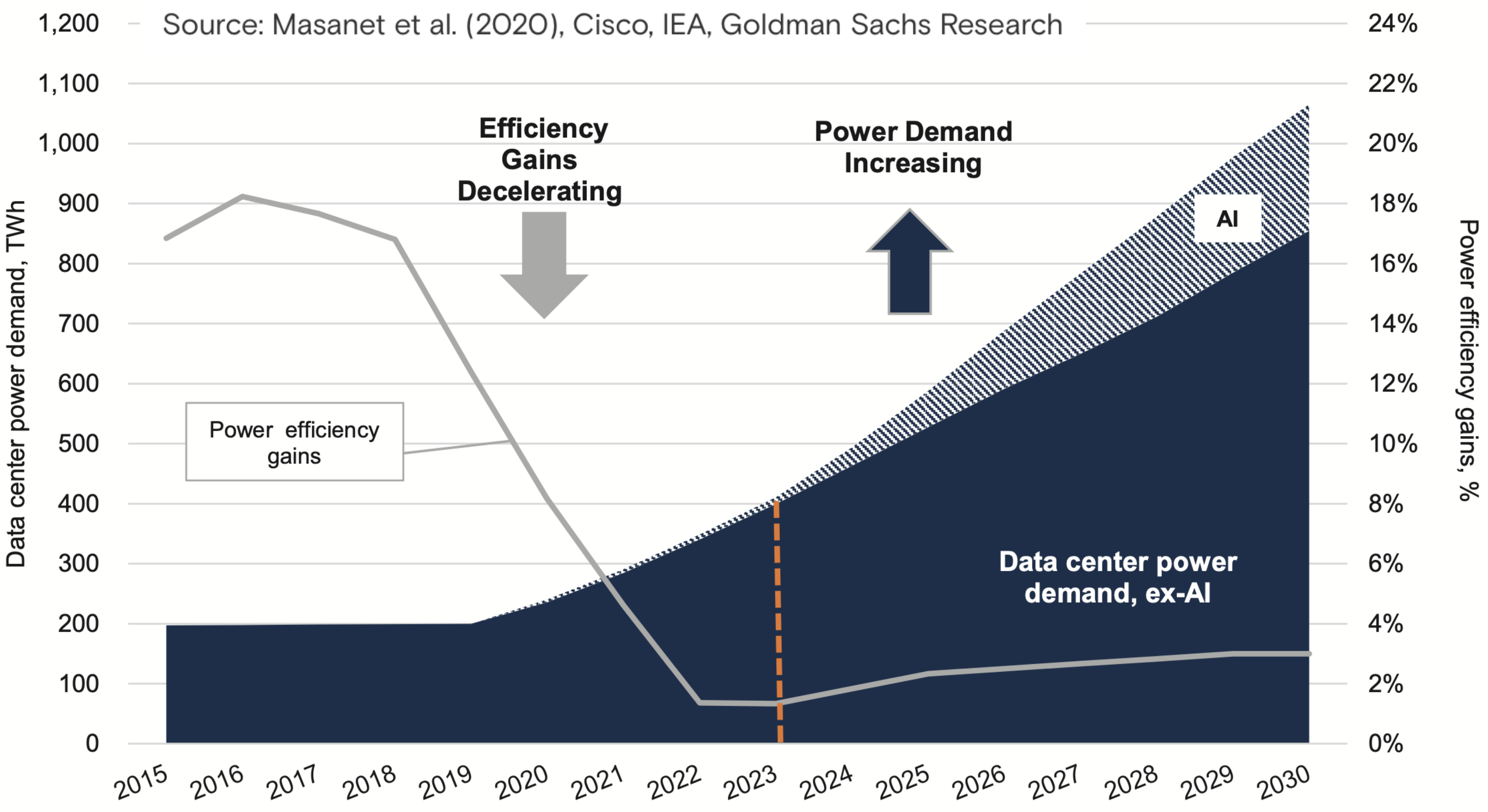
NVIDIA CHIP EFFICIENCY



CATCH ME IF YOU CAN!



GROWTH IN DATACENTER ENERGY



OPERATIONAL VS. EMBODIED EMISSIONS

> "The **use stage** GHG emissions in 2020 relating to electricity use account for the **majority of total GHG emissions**."

- *Malmodin et al. (2020)*

OPERATIONAL EMISSIONS
Scope 1 & Scope 2

95 million tons CO_2

EMBODIED EMISSIONS
Scope 3

31 million tons CO_2

76% | 24%

POST-MOORE DATACENTERS

Design for "ISA"

- Integration
  - reduce data movement
- Specialization
  - cut resources to analyze data
- Approximation
  - compress data & computation

From algorithms to infrastructure

Algorithm

Infrastructure

CENTER @ EPFL SINCE 2011

Mission

- Sustainable computing
- IT for sustainability
- Best practices, metrics & methodologies

Impact

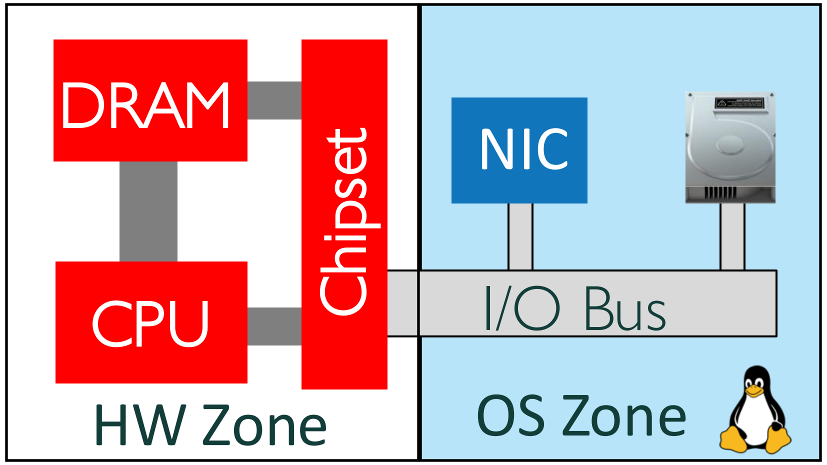
- Server-grade ARM CPU
- Cloud-native network/database stacks
- Liquid-cooling from chip to rack

ecocloud.ch

OUTLINE

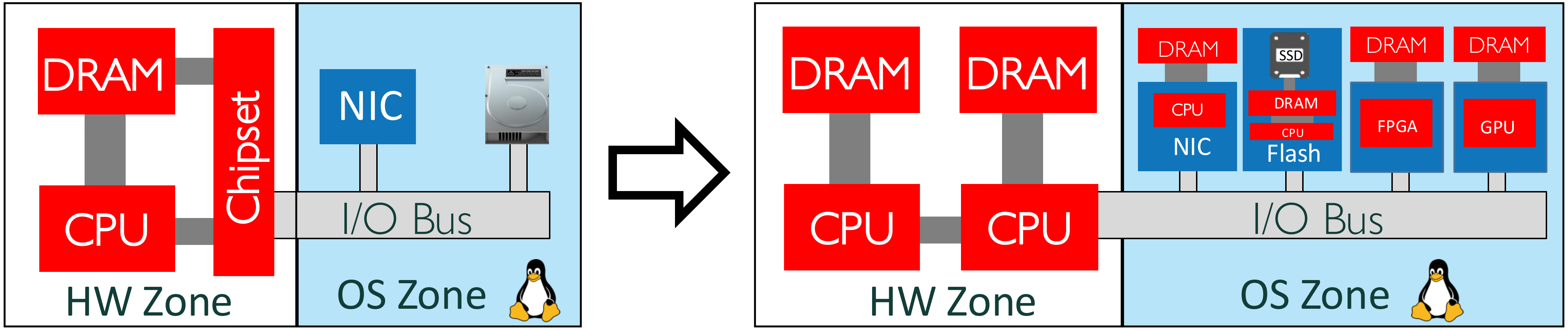
- Overview
- Post-Moore Computing
  - Compute infrastructure
  - AI runtime stack
  - Metrics & methodologies
- Summary

TODAY'S SERVER = 90'S DESKTOP PC



90's Desktop PC

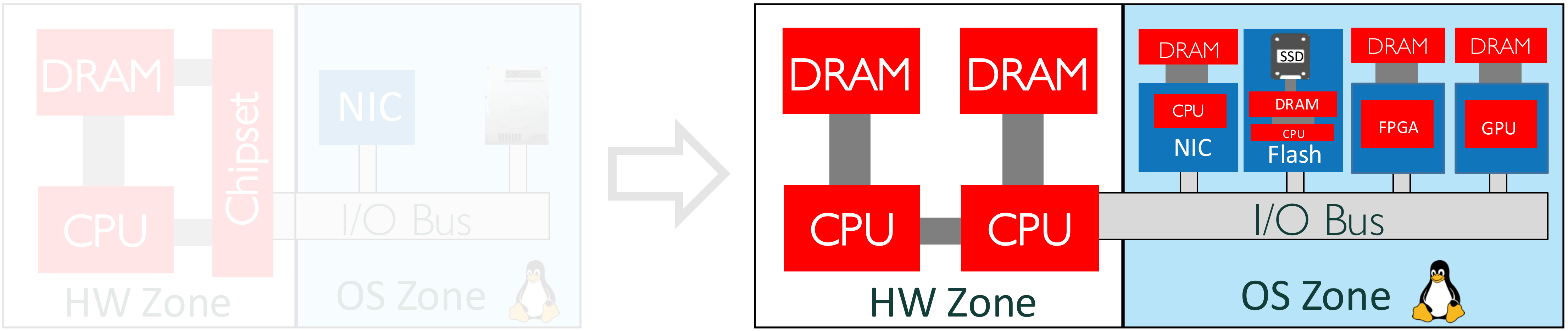
TODAY'S SERVER = 90'S DESKTOP PC



90's Desktop PC

Today's Server

TODAY'S SERVER = 90'S DESKTOP PC



90's Desktop PC　　Today's Server

- Focused on minimizing cost (Google c.a. 2000)
- CPU, memory = nanosecond timescale, OS, I/O = millisecond timescale
- OS follows legacy interfaces (PCIe) and abstractions (POSIX)
- Silicon fragmented across legacy interfaces

EFFICIENCY PROBLEMS IN COMPUTING

Hardware/workload mismatch (EPFL, Meta, Google)

Datacenter tax ~ 20% (Google)

- 20,000 threads running per CPU
- Virtualization/containerization
- RPC

Memory wasted (Microsoft)

- 50% of containers do not use their memory
- 20% of memory is stranded

GPU utilization for deep learning < 50% (Microsoft)

POST-MOORE SERVERS [IEEE Micro'24]

Server-centric CPU design

- Exploit massive request-level parallelism per service
- Maximize efficiency: throughput/area, throughput/Watt

Tight integration of CPU, GPU, memory, NIC

- Emerging chip-to-chip standards (UCIe)
- High-bandwidth memory for AI

Rack-level fabrics

- NVLink, CXL

Liquid cooling at chip level

- 2x higher power density

POST-MOORE CPUS [ISCA'12]

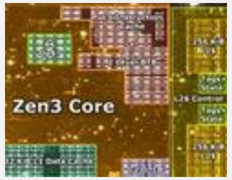
Today's server CPUs

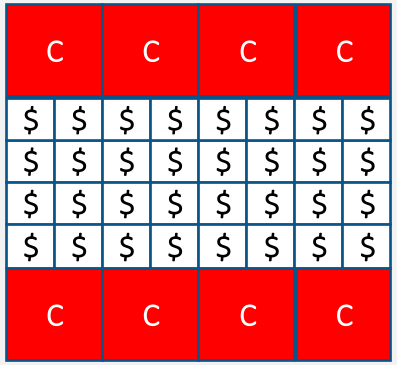
- ✗ Designed for single-core performance
- ✗ Power-bound → ½ big cores + ½ memory
- ✗ Run at high frequency
  (power ~ superlinear w/ performance)

AMD Zen 3

4.0 mm^2

3.7W @ 3 GHz





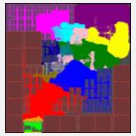
Cloud-native CPUs

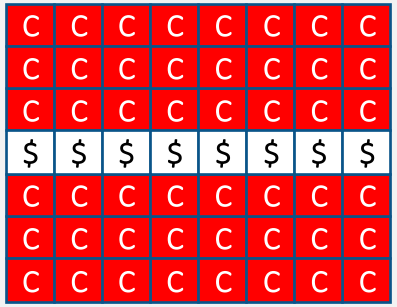
- ✓ Custom cores for max area density
- ✓ Higher throughput/Watt at lower frequency
- ✓ Need only memory for per-core working set

ARM N1

1.4 mm^2

0.7W @ 2 GHz

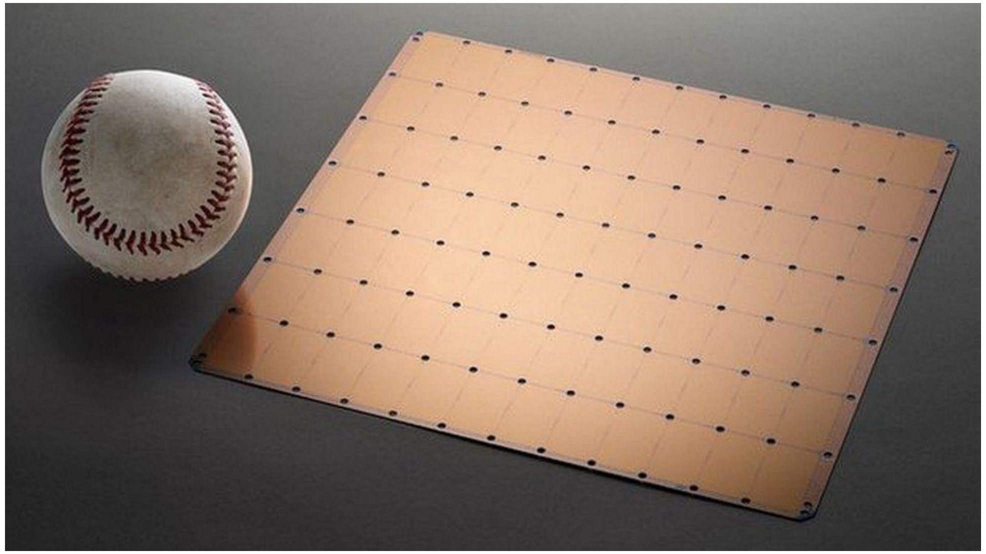




10X higher throughput with SLO!

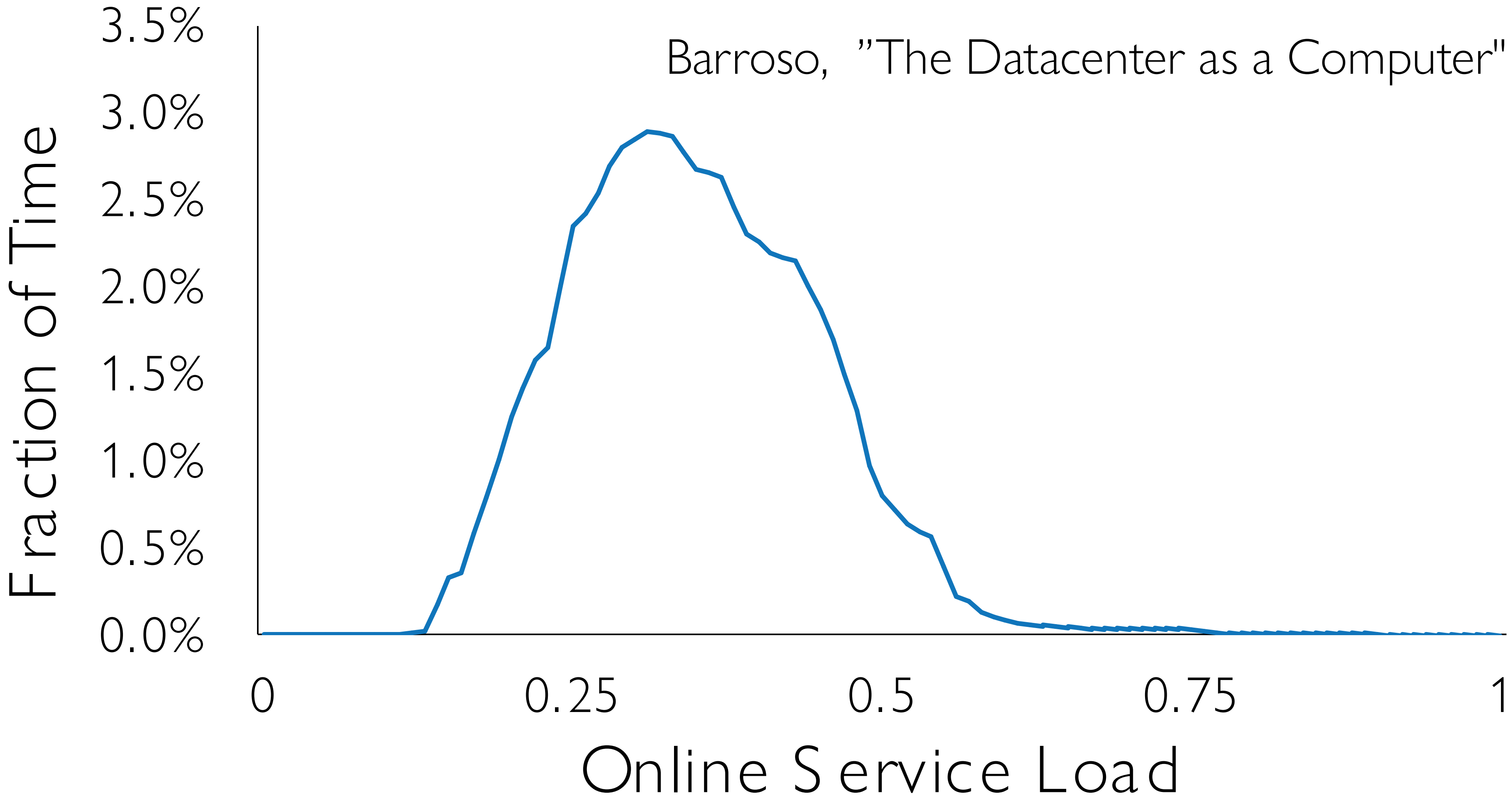
AI ACCELERATORS

Inference workloads:
- Online
- Tight latency constraints
- Rely on low-precision arithmetic



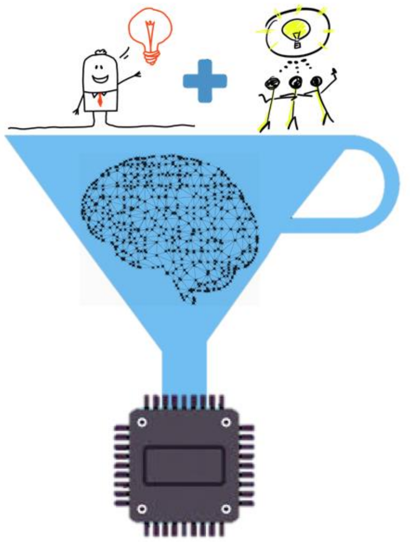
Training workloads:
- Offline
- Throughput optimized
- Need high-precision arithmetic

INFERENCE UTILIZATION



UNIVERSAL AI ACCELERATORS

- Scaled numeric formats (HBFP [NeurIPS'18], MX)
  - Quantize while maintaining accuracy
  - Use for both for inference and training
  - Inference → helps with outliers
  - Training → don't need FP precision for dot products
  - Work well with sparsification [Harma, ICLR'25]
  - 4-Bit training for transformers [Harma, arXiv'24]



parsa.epfl.ch/coltrain

- Accelerators with prioritized schedulers [Drumond, MICRO'21]
  - Piggy-back fine-tuning jobs on an inference accelerator

AI RUNTIME

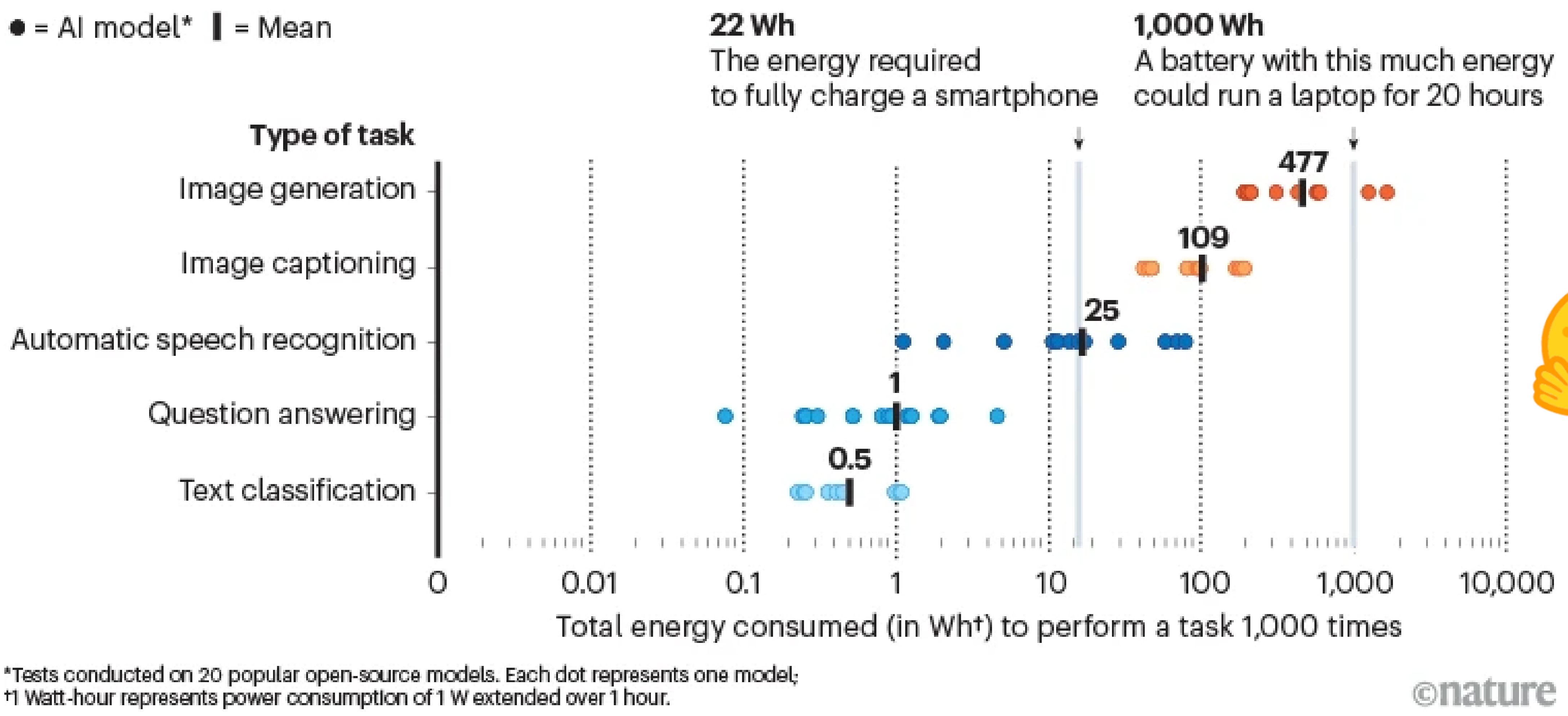
- Enhance utilization [Gao, ICSE'24]
  - Proper batching
  - Overlap CPU-centric tasks
  - Hide data transfer between CPU/GPU
  - Minimize or overlap communication

- Need elasticity (Ana's talk)
  - Today's runtimes are too static (container-based deployment)
  - Allow software to scale up/down GPU, memory, network resources

METRICS & METHODOLOGIES

- Design metrics
  - Post-Moore means more accelerators
  - Take into account both operational & embodied emissions
- Operational metrics
  - How much energy do we need for training/inference?
- Operational methodologies
  - How do we measure/monitor our efficiency?

AI SUSTAINABILITY CLASSIFICATION



MEASURE FULL-STACK EFFICIENCY

DC EFFICIENCY
- electricity w/ renewables, cooling, heat recycling

IT EFFICIENCY
- compute, storage, network and workloads

CARBON FOOTPRINT
- emissions from input electricity sources

DATA CENTER

EFFICIENCY

sdea.ch

EPFL EcoCloud Hewlett Packard Enterprise asut Schneider Electric SWISS DATACENTER ASSOCIATION energieschweiz HOCHSCHULE LUZERN

SUMMARY

AI's energy requirements grow exponentially

Moore's Law of silicon is dead

Need post-Moore technologies, metrics, best practices

Post-Moore computing:

- Integration + Specialization + Approximation

THANK YOU!

For more information, please visit us at

parsa.epfl.ch