Sustainability of Edge to Cloud Computing

Adrian M. Ionescu, Nanolab, EPFL

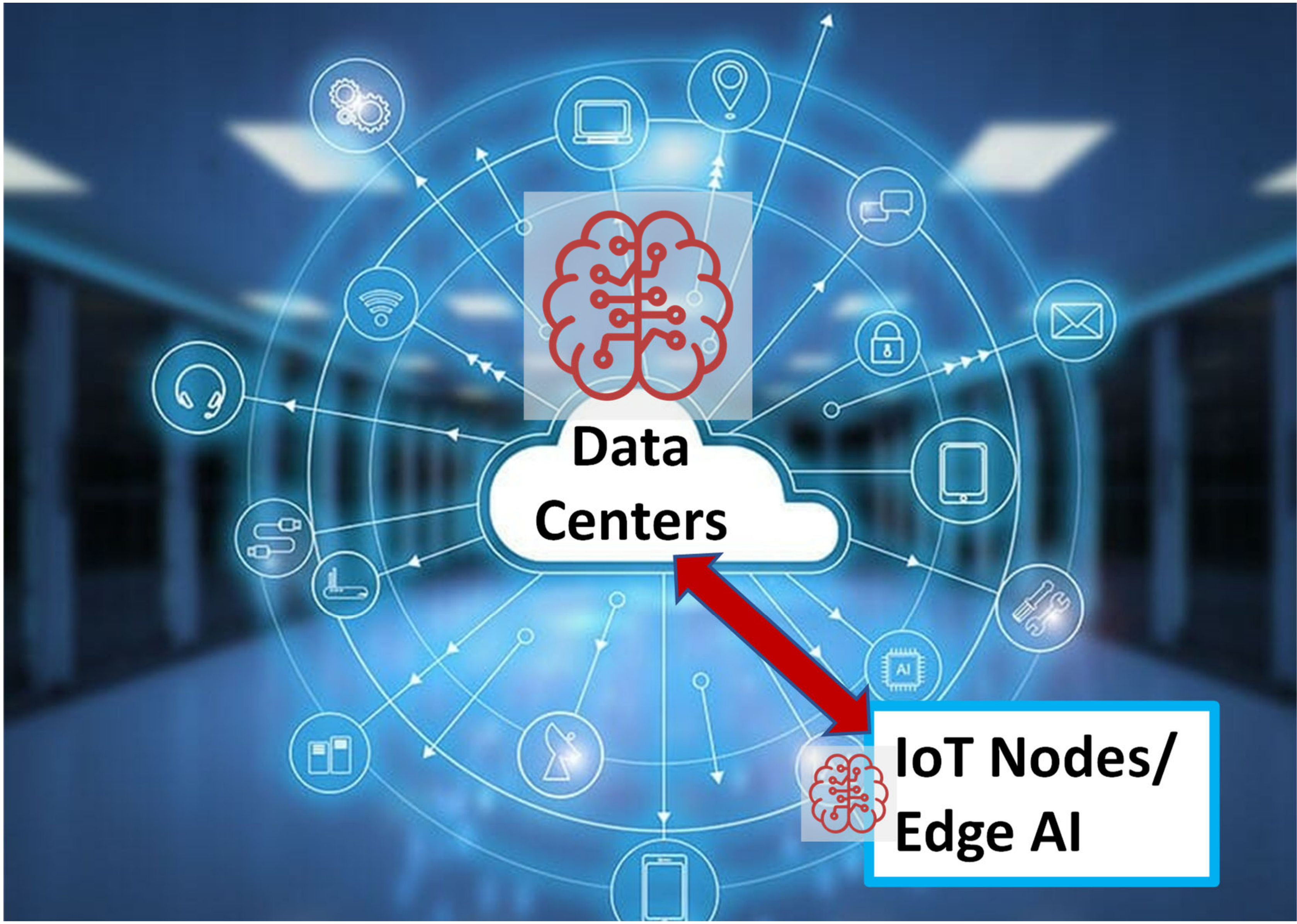
Outline

- Introduction & Learning Objectives
- **Part I – Cloud AI & Data Centers**
- **Part II – Edge AI, IoT & Data Proliferation**
- Wrap-up, Takeaways & Discussion

Introduction

- Computing today and the Zettabyte Era
- Today's sustainability paradox:

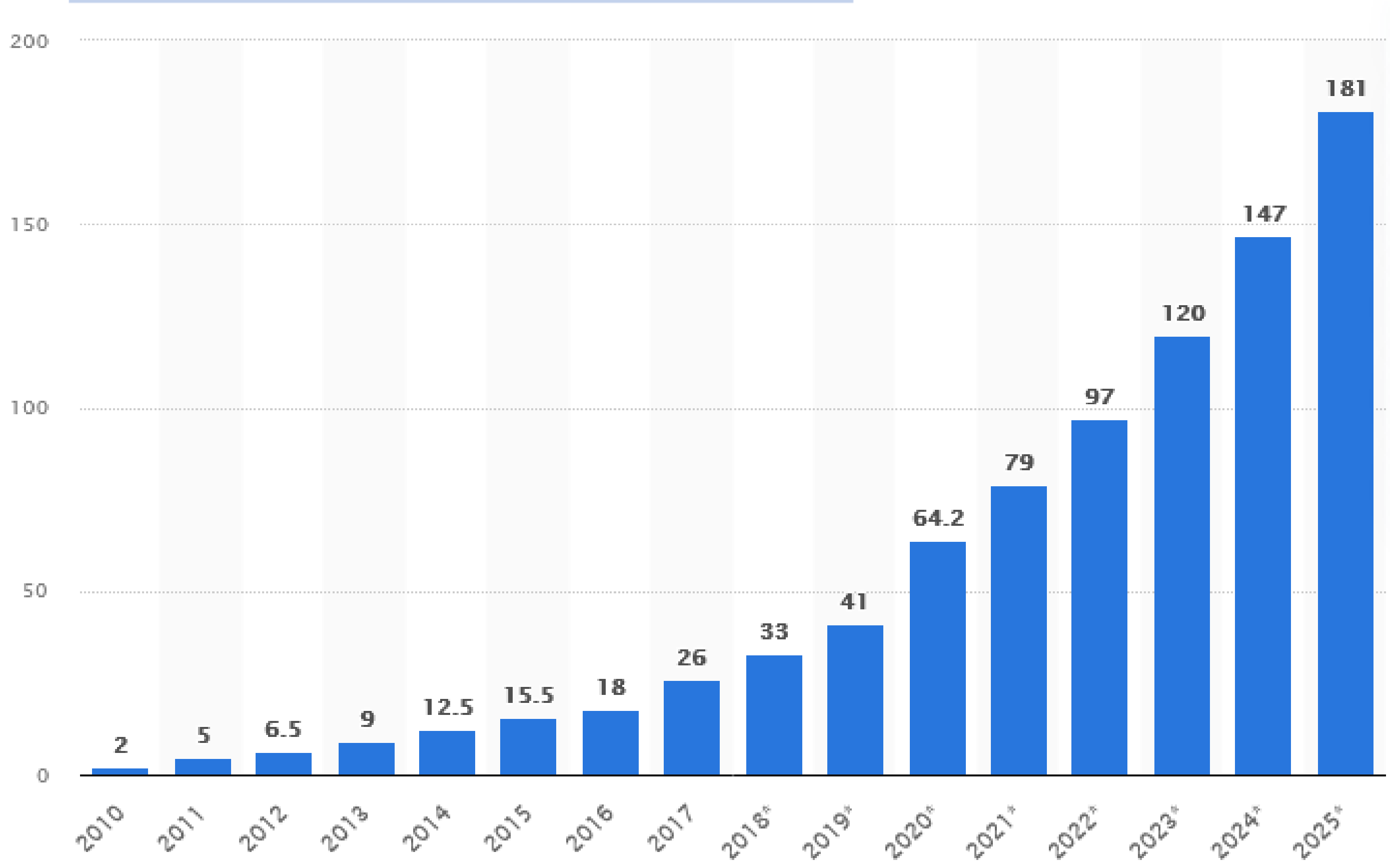
  AI enables efficiency but consumes unsustainable resources.
- Centralized (cloud) vs. decentralized (edge) AI systems.
- Key question: How can we make AI computing greener across compute hierarchies?


Data
Centers
IoT Nodes/
Edge AI

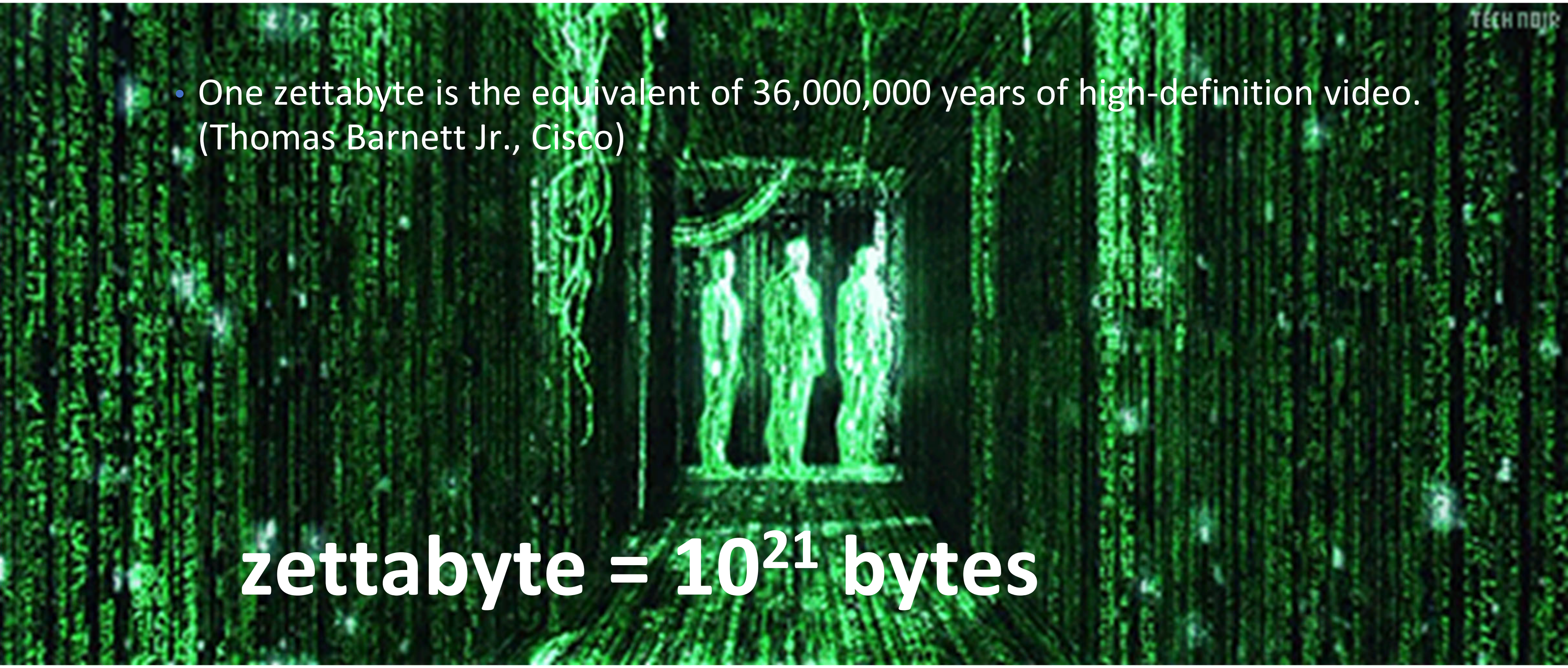
EPFL

Data volume is exploding

Data volume in Zettabytes



The Zettabyte Era… started in 2010!



- One zettabyte is the equivalent of 36,000,000 years of high-definition video. (Thomas Barnett Jr., Cisco)

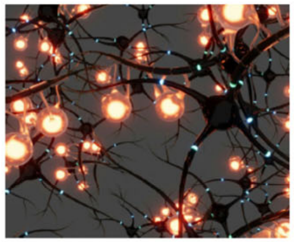
zettabyte = 10^{21} bytes

EPFL

Preliminary Framing
Computing specific domains

Quantum Computing

Quantum computing is an emergent field of cutting-edge computer science harnessing the unique qualities of quantum mechanics to solve problems beyond the ability of even the most powerful classical computers.[1]

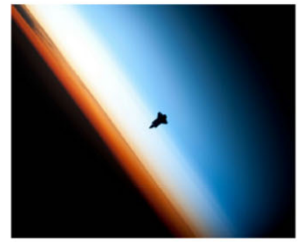


Neuromorphic Computing

Neuromorphic computing, also known as neuromorphic engineering, is an approach to computing that mimics the way the human brain works. It entails designing hardware and software that simulate the neural and synaptic structures and functions of the brain to process information.[2]

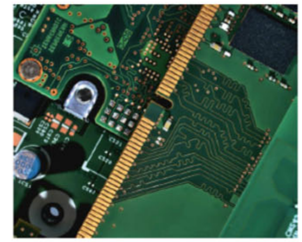
Biocomputing

Biocomputing uses molecular biology parts as the hardware to implement computational devices. By following pre-defined rules, often hard-coded into biological systems, these devices are able to process inputs and return outputs — thus computing information.[3]



In-orbit/space Computing

In-orbit or space-qualified computing is a technology that has been developed to address the most computationally-intensive part of a space mission. It can be deployed in flight systems, whether in space or the atmosphere, and will advance all types of future space missions.[4]



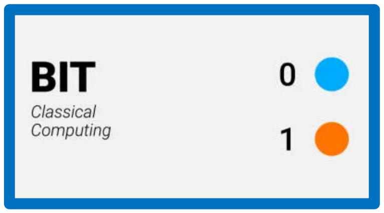
High-Performance Computing

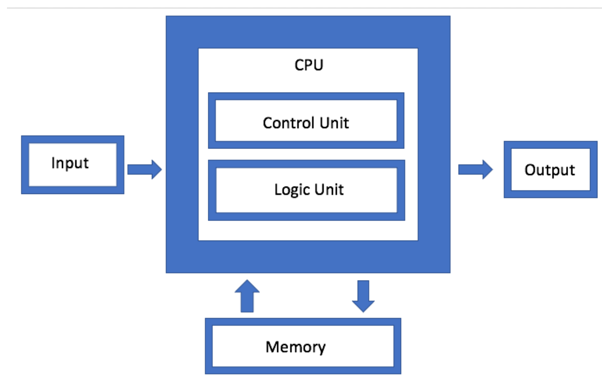
High-performance computing (HPC) is the art and science of using groups of cutting edge computer systems to perform complex simulations, computations, and data analysis out of reach cutting-edged commercial compute systems available.[5]

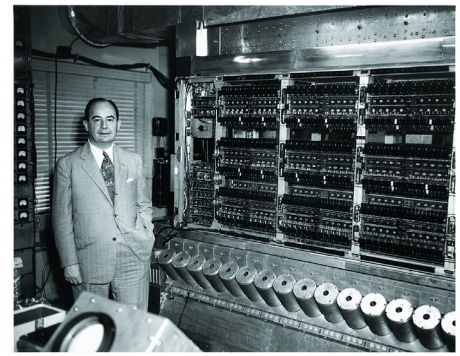
WORLD ECONOMIC FORUM

Global Future Councils

Von Neumann computing and beyond

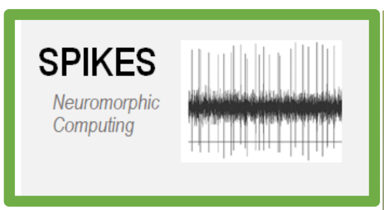


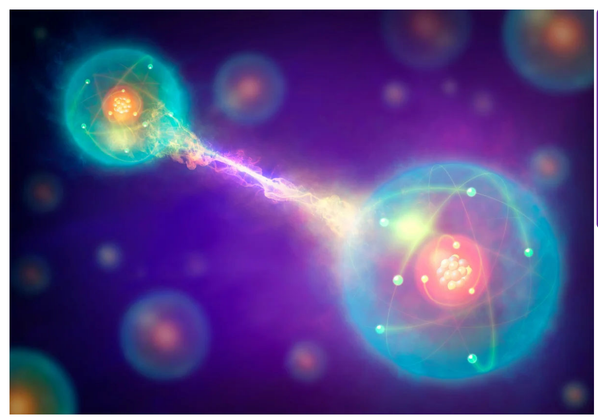




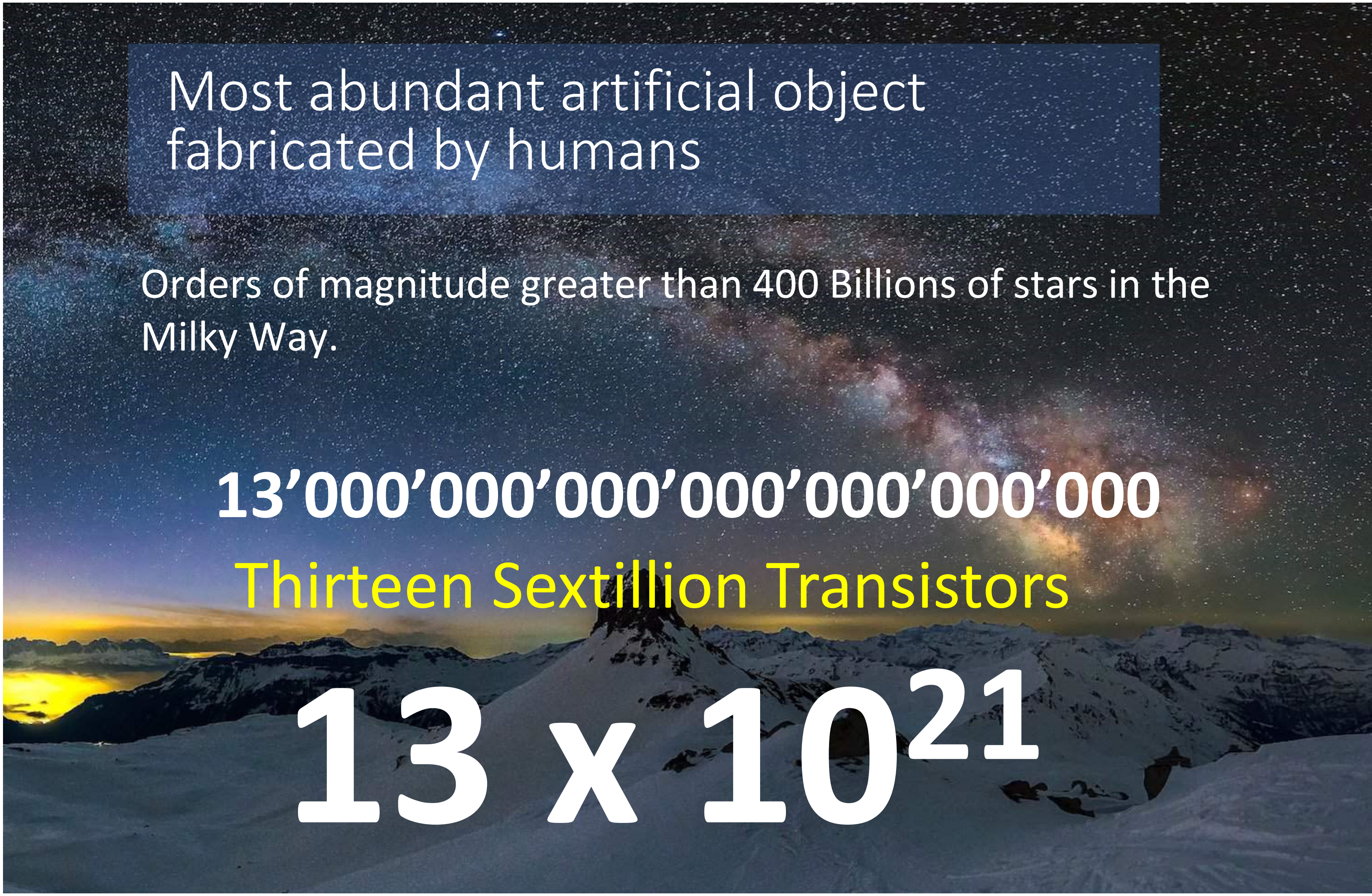
von Neumann in the 40's

*"The future of computing will not be based on ever-increasing processing power… it will rely on **understanding and drawing inferences from massive collections of data**."*





EPFL



Most abundant artificial object fabricated by humans
Orders of magnitude greater than 400 Billions of stars in the Milky Way.
13’000’000’000’000’000’000’000
Thirteen Sextillion Transistors
13 x 10^21

Digital Era and the Moore's law

>100million /mm^2

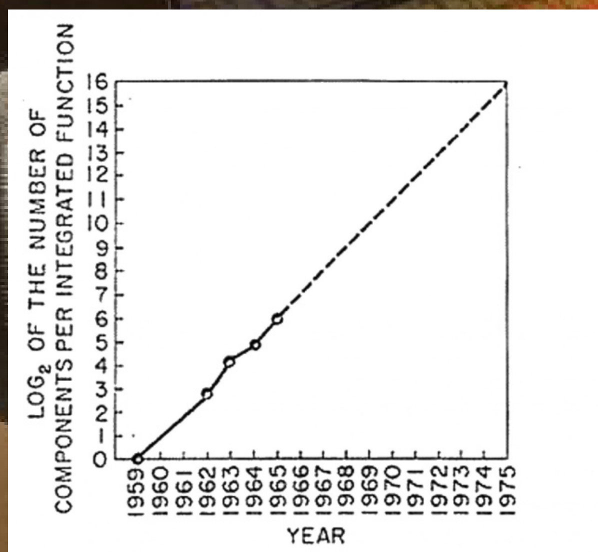
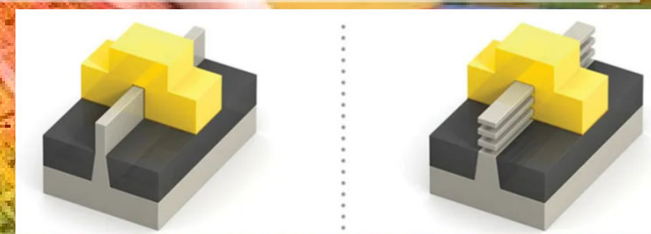
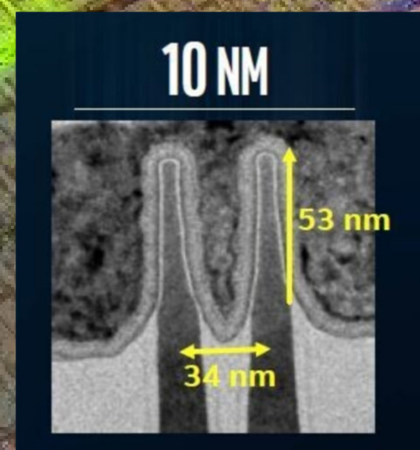
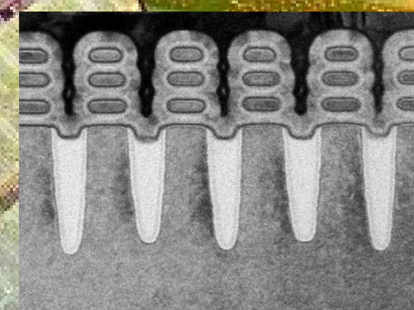


Fig. 2 Number of components per integrated function for minimum cost per component extrapolated vs time.

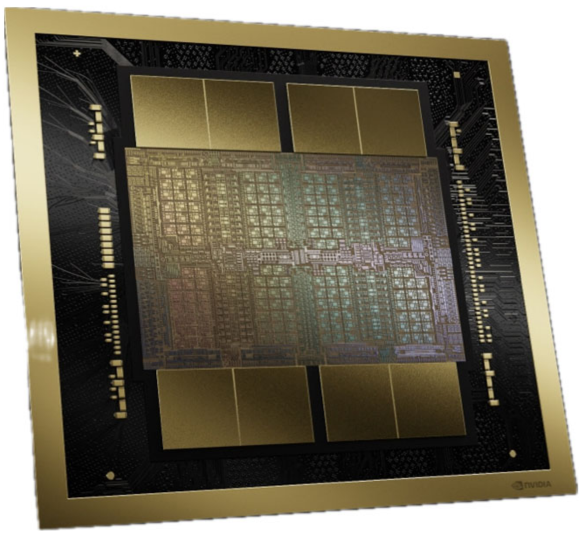


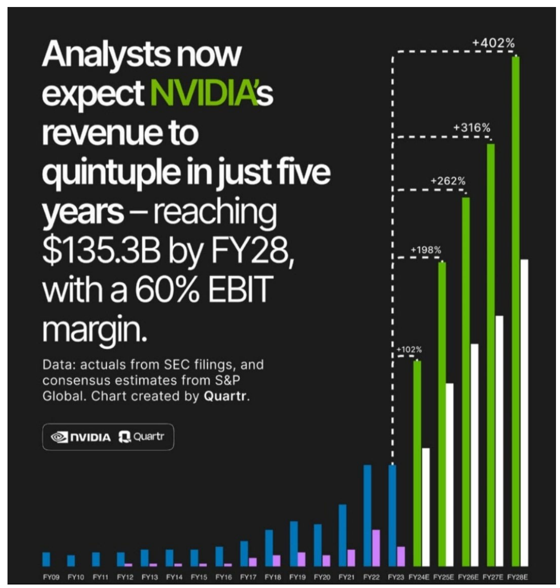


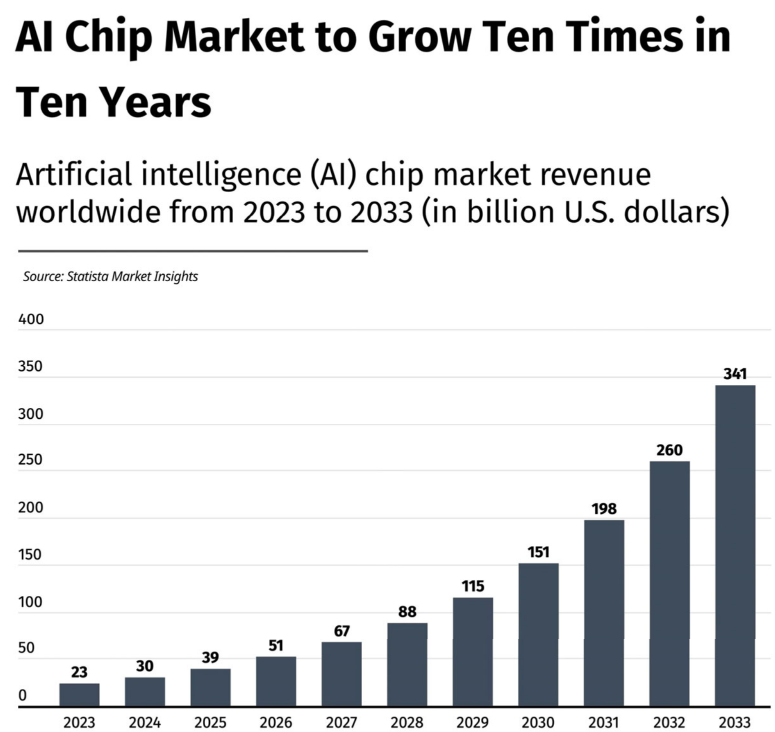
1 NM

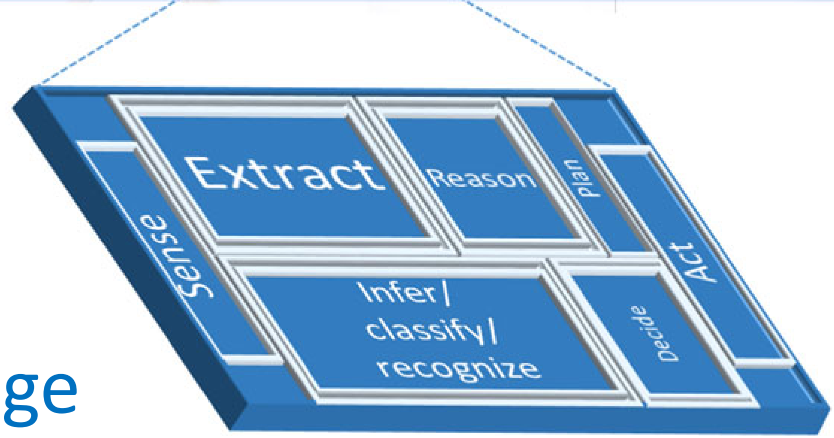


- 208B transistors in NVIDIA Blackwell, 576 GPUs, 10 TB/second chip-to-chip link









AI
@ the Edge

Edge to Cloud information processing in Digital Era

EPFL

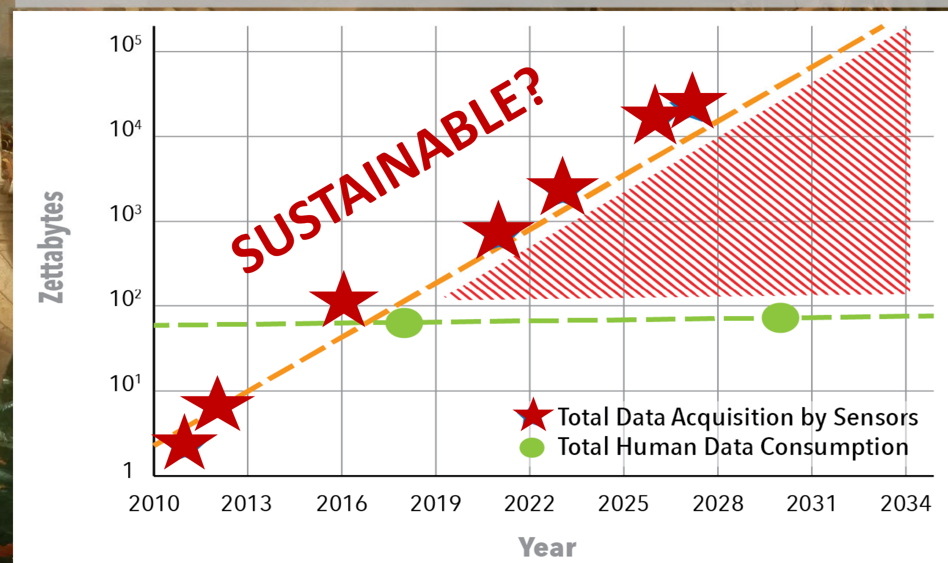
Energy efficiency and data proliferation

Data Centers = Big Brains

- global scale = 416 terawatts, or **3% of all electricity on Earth**
- **Energy inefficient**

Internet of Things Nodes = Tiny Brains

+ 1 trillion IoT devices by 2035 with annual growth >20% (ARM)



The rise and the fall of the Roman Empire

Part I: Cloud Computing & AI Data Centers

- Data centers = the Big Brains of Internet
- About energy (in)efficiency
- Power Usage Effectiveness (PUE) and its limits
- Cooling technologies: air, liquid, immersion
- Renewable energy integration challenges

Data Centers

Big Brains of Internet

- global scale = 416 terawatts, or **3% of all electricity on Earth**
- 4% of Swiss electricity usage, will double in next 5 years
- Ireland: 14% of national usage, up to 27% by 2029
- **Very energy inefficient**

Data Center Components

1. Servers: The Workhorses of Data Processing

- powerful computers are the heartbeat of the operation, handling applications, computations, and storage tasks

Processing Power

2. Networking Equipment

- connects servers, devices, and users within and beyond the data center. Routers, switches, and other devices facilitate the transmission of data, ensuring that information flows seamlessly and securely between different components. This connectivity is the lifeline that enables real-time processing.

Data Transfer, Communication and Security.

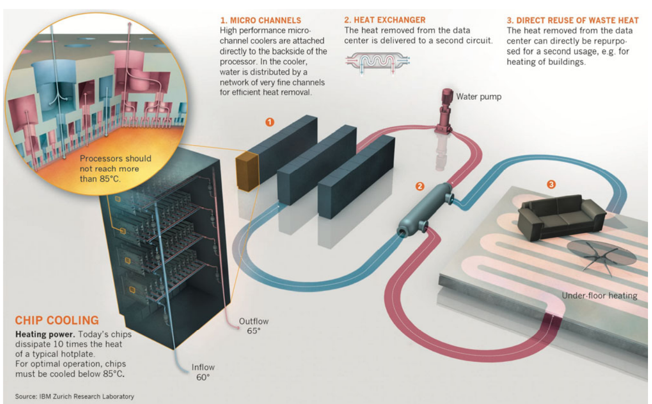
3. Storage Systems

- serve as the repositories for the immense volumes of data generated and processed daily. These include hard disk drives (HDDs), solid-state drives (SSDs), and other storage solutions. Storage systems provide the necessary space for

Data Retention, Redundancy, Backup and Data Accessibility.

Energy efficiency challenge in the cloud

Data centers:
The average data center uses the same amount of electricity needed to power a small city.



IBM's Aquasar data center with innovative water-cooling system:
6 kilowatts of thermal power to heat ETH Zurich.

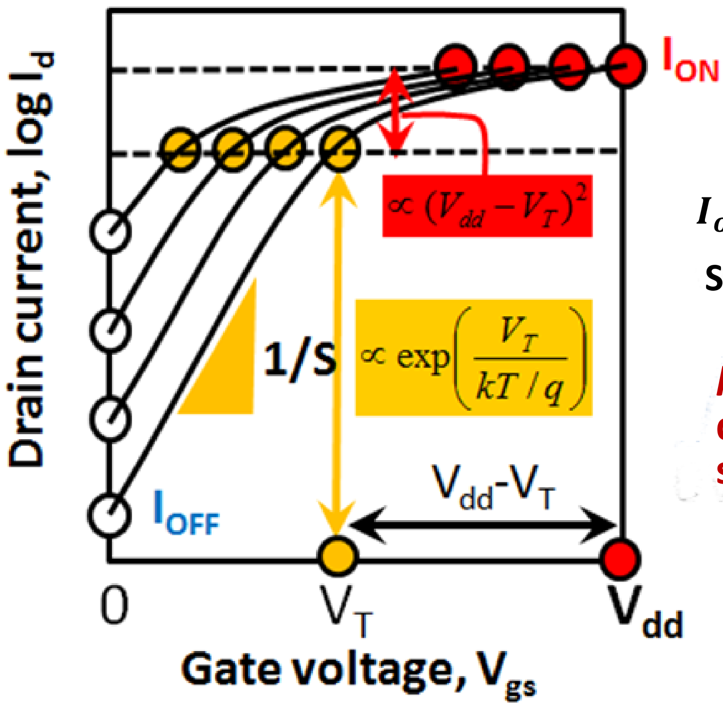
- Energy expenditures are becoming more significant than the cost of machines.
- Energy efficient strategies for data centers!

- Advanced CMOS processors in servers: leakage power dominant over dynamic.

$$P = \alpha L_D C V_{dd}^2 f + I_{off} V_{dd}$$

> 8 millions data centers in 2024.

The MOSFET switch: key benchmarks

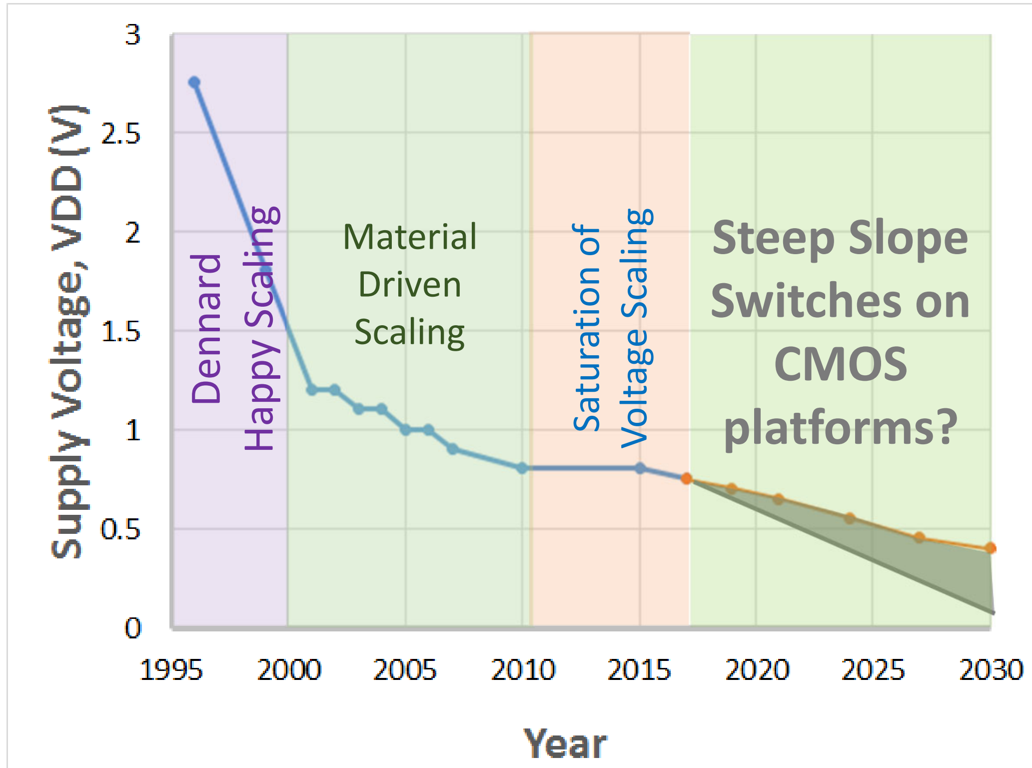


Ionescu & Riel, Nature, 2011.

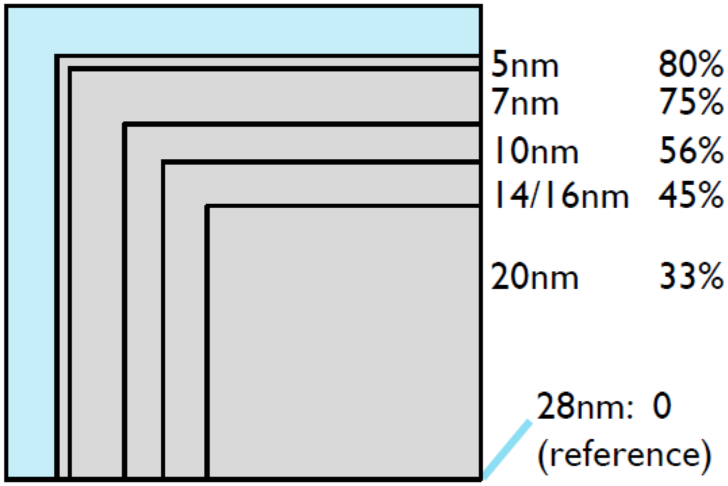
$$I_{off} = I_0 10^{-V_{th}/SS}$$

$$SS = \frac{dV_g}{dlog(I_d)} \text{ @ RT}$$

***kT/q* does not scale!**



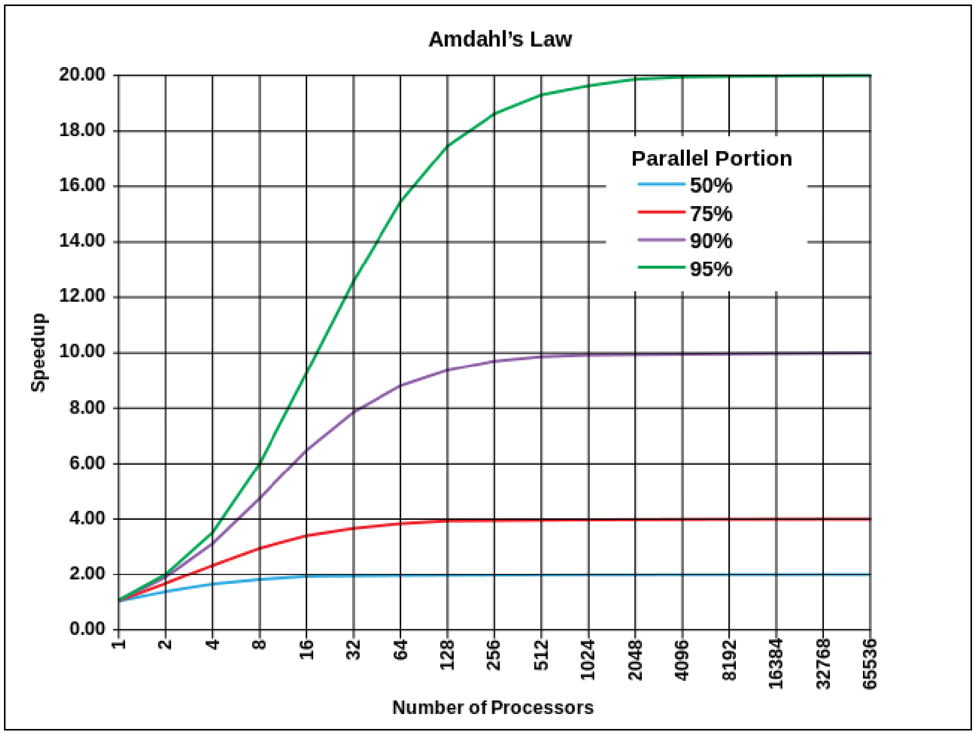
Power density and dark silicon



We get more transistors, we just can't afford to turn them all!

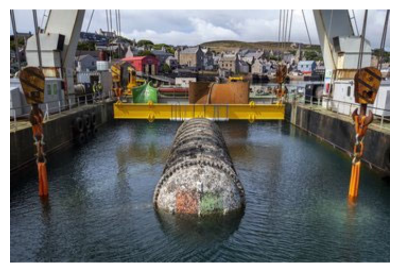
Greg Yeric, ARM @ IEDM 2015

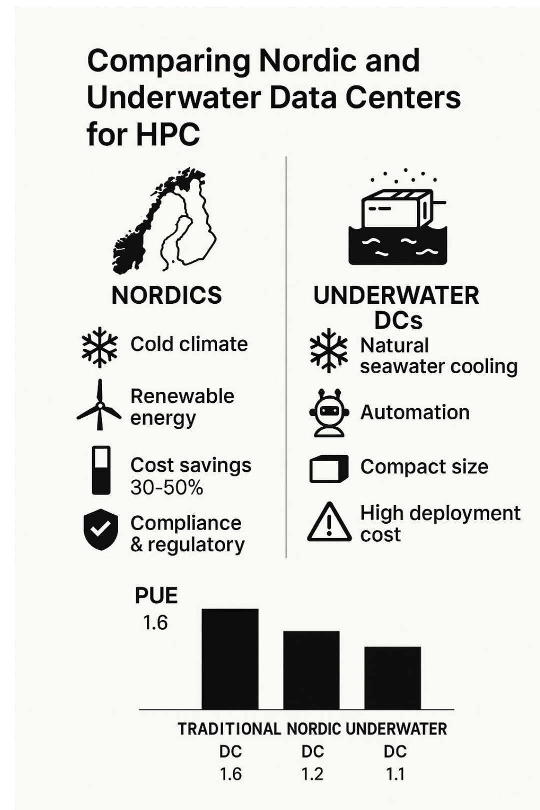
One or two walls?



Data Centers for HPC: Strategic Placement for Sustainability & Performance

- **Nordic Regions (Norway, Sweden, Finland, Iceland, etc.):**
  - 🌡️ Naturally cold climate reduces cooling costs significantly.
  - ⚡ Abundant renewable energy (hydro, wind, geothermal).
  - EU Strong data privacy laws and political stability.
  - 🏗️ Government incentives for green infrastructure projects.
- **Underwater Data Centers (e.g., Microsoft's Project Natick):**
  - 🌊 Seawater cooling provides efficient thermal management.
  - 🚫 Reduced real estate usage and land footprint.
  - 🤖 Fully automated systems reduce need for on-site staff.
  - 🔒 Physically secure and isolated from terrestrial threats.

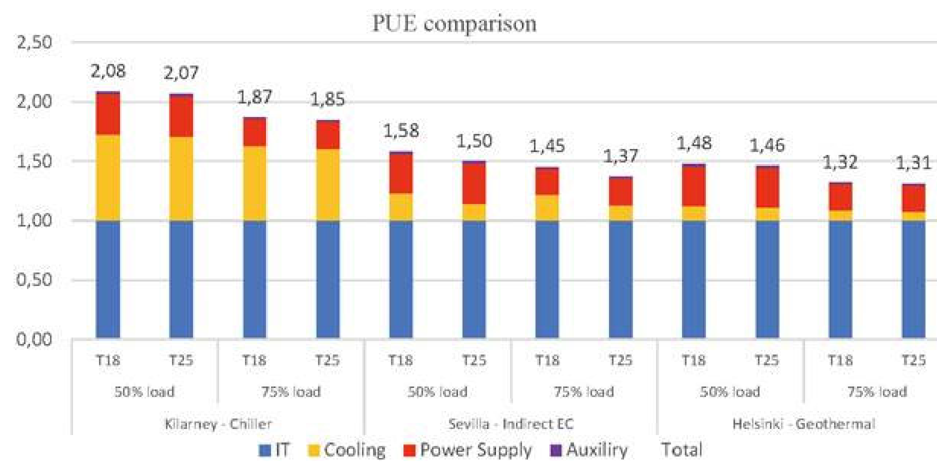




Metrics for Data Center Efficiency

- The main indicator used to assess overall data center energy efficiency is **PUE (=Power Usage Efficiency)**, which shows the ratio between total facility power use and IT equipment power use (Avelar et al., 2012):
- The optimal value for PUE is 1.0, the max value is infinity.

Metric Description	Metric Formulation
Power Usage Efficiency	$PUE = \dfrac{Total\ facility\ power}{Total\ IT\ power}$
Data Center Infrastructure Efficiency	$DCiE = \dfrac{Total\ IT\ power}{Total\ facility\ power}$
Carbon Usage Effectiveness	$CUE = \dfrac{Total\ CO2\ emissions\ from\ DC\ energy}{Total\ IT\ Equipment\ energy}$
IT Equipment Utilization	$ITEU = \dfrac{Total\ measured\ energy\ of\ IT}{Total\ specification\ energy\ of\ IT}$



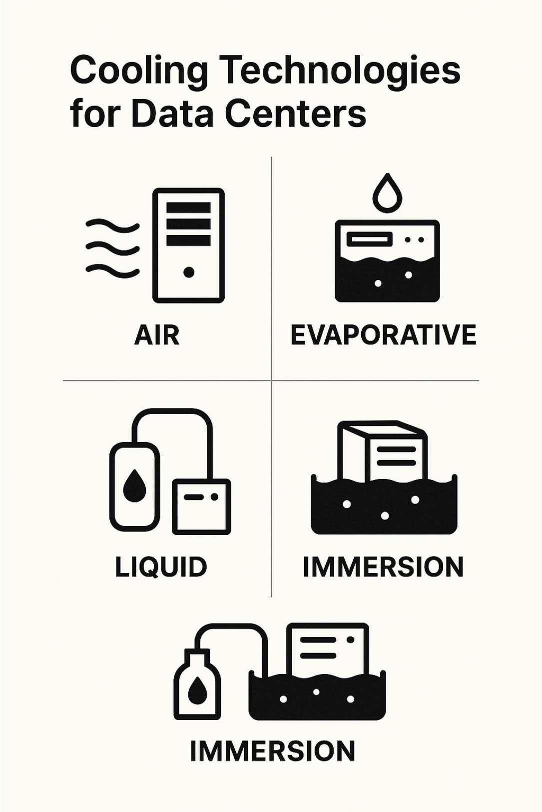
Cooling technologies for Data Centers

Air Cooling
- Uses fans and airflow to dissipate heat.
- Most common and cost-effective method.
- Limited efficiency in high-density HPC setups.

Evaporative Cooling
- Cools air through water evaporation.
- More energy-efficient than traditional air cooling.
- Requires consistent water supply and humidity control.

Liquid Cooling
- Circulates coolants (e.g., water, glycol) through pipes near heat sources.
- Higher thermal efficiency than air cooling.
- Suitable for high-performance or densely packed servers.

Immersion Cooling
- Servers are submerged in thermally conductive dielectric fluid.
- Enables extreme heat removal and compact design.
- Reduces energy usage for cooling dramatically.
- Ideal for edge computing and extreme HPC environments.



Renewable energy integration challenges

Power Supply Intermittency

- Solar and wind are **weather-dependent**.
- Leads to **load balancing issues** and **reliability risks**.

Energy Storage Limitations

- High-performance batteries are **expensive** and **space-intensive**.
- Current storage tech **lacks scalability** for 24/7 uptime needs.

Grid Infrastructure Constraints

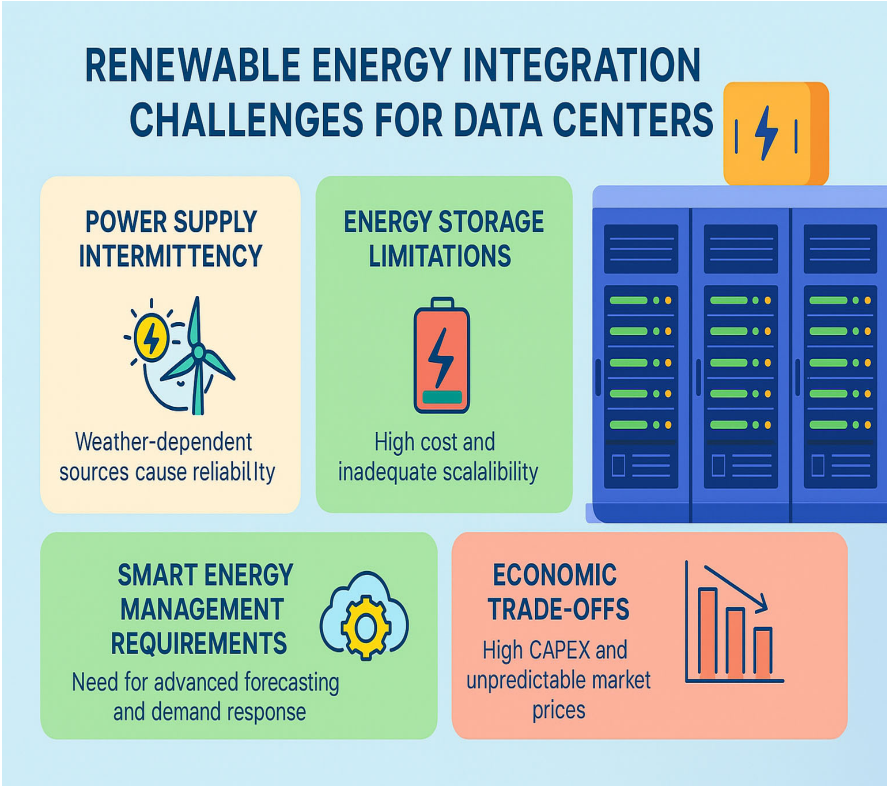
- Legacy grids struggle with **bidirectional flow** and **volatility**.
- **Transmission bottlenecks** in remote renewable-rich regions.

Smart Energy Management Requirements

- Need for **AI-driven load forecasting**, demand response.
- Requires **real-time integration** with cloud, edge systems.

Economic Trade-Offs

- Higher **CAPEX for renewable installations**.
- Unpredictable **energy market prices** can hurt OPEX.



Growth of AI model size and computation demands

Explosion in AI Model Sizes

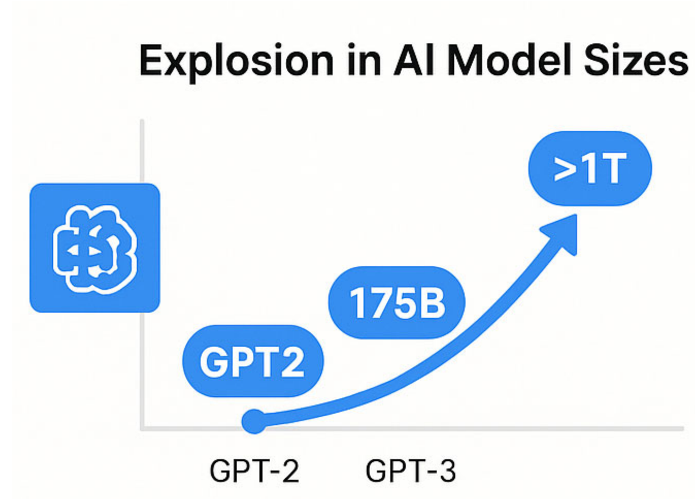
- GPT-2 (2019): 1.5B parameters
- GPT-3 (2020): 175B parameters
- GPT-4 (2023+): >1T parameters
- DALL·E 2, Stable Diffusion, Gemini, Claude — growing multimodal capabilities
- Implication: exponential compute, memory, and storage requirements

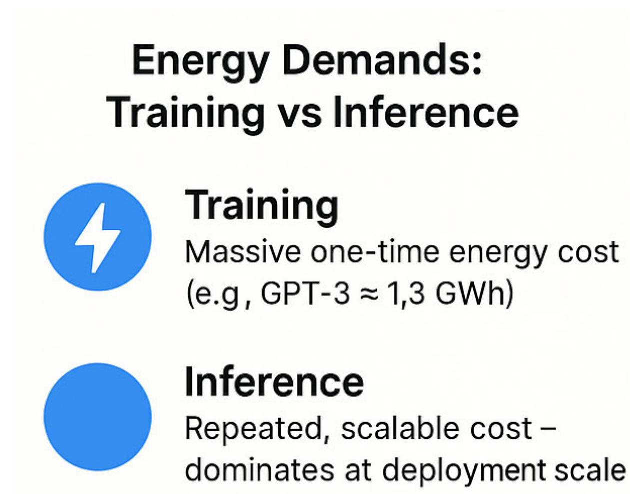
Rise of Cloud Hyperscalers and AI Workloads

- Major players: AWS, Microsoft Azure, Google Cloud Platform (GCP)
- Massive investment in GPU/TPU clusters & AI-specific infrastructure
- AI workloads dominate cloud revenue growth (training & inference)

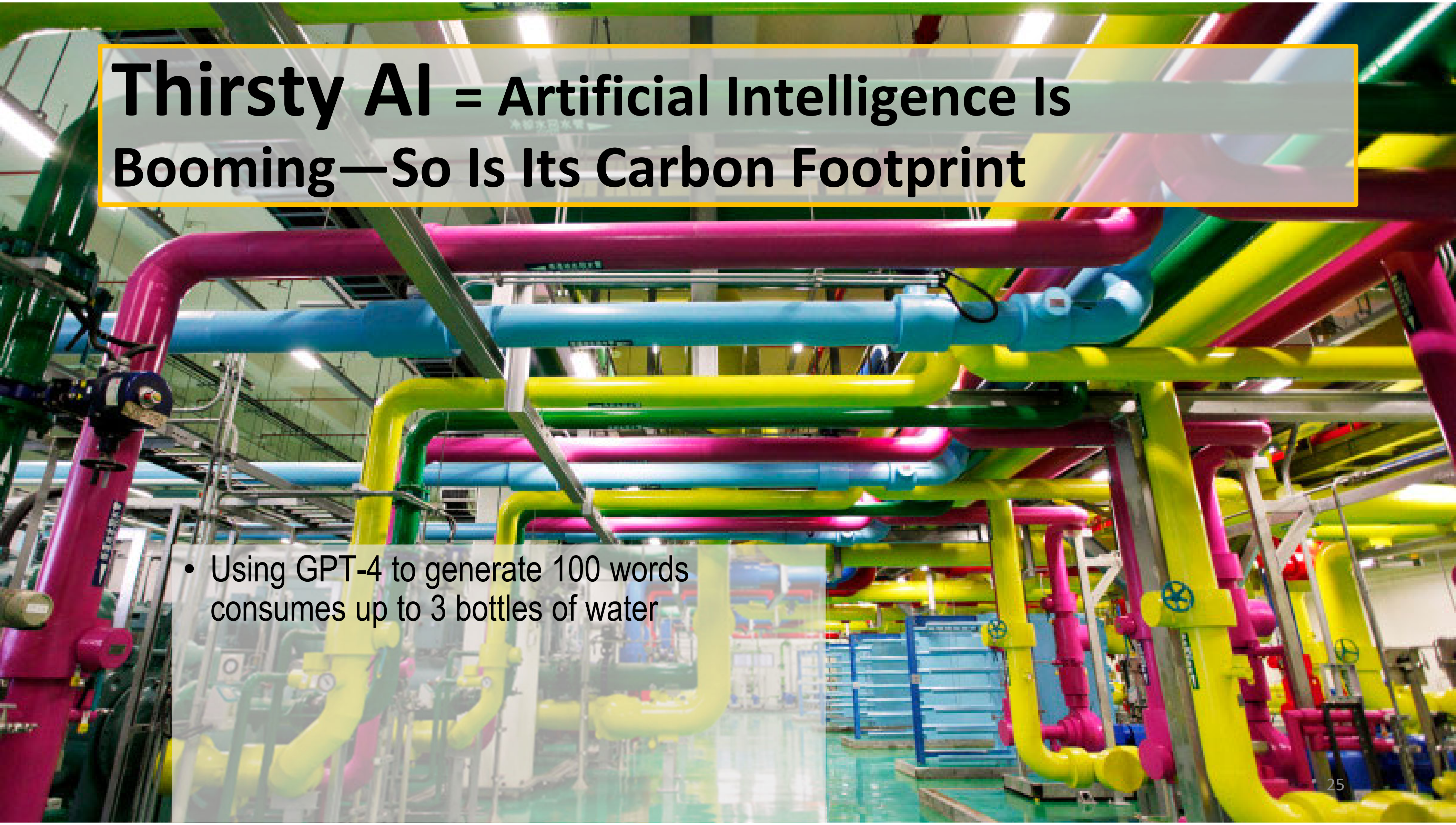
Energy Demands: Training vs Inference

- **Training**: Massive one-time energy cost (e.g., GPT-3 ≈ 1.3 GWh)
- **Inference**: Repeated, scalable cost — dominates at deployment scale
- Urgency to improve energy efficiency of both phases (hardware & algorithmic optimization)





Thirsty AI = Artificial Intelligence Is Booming—So Is Its Carbon Footprint



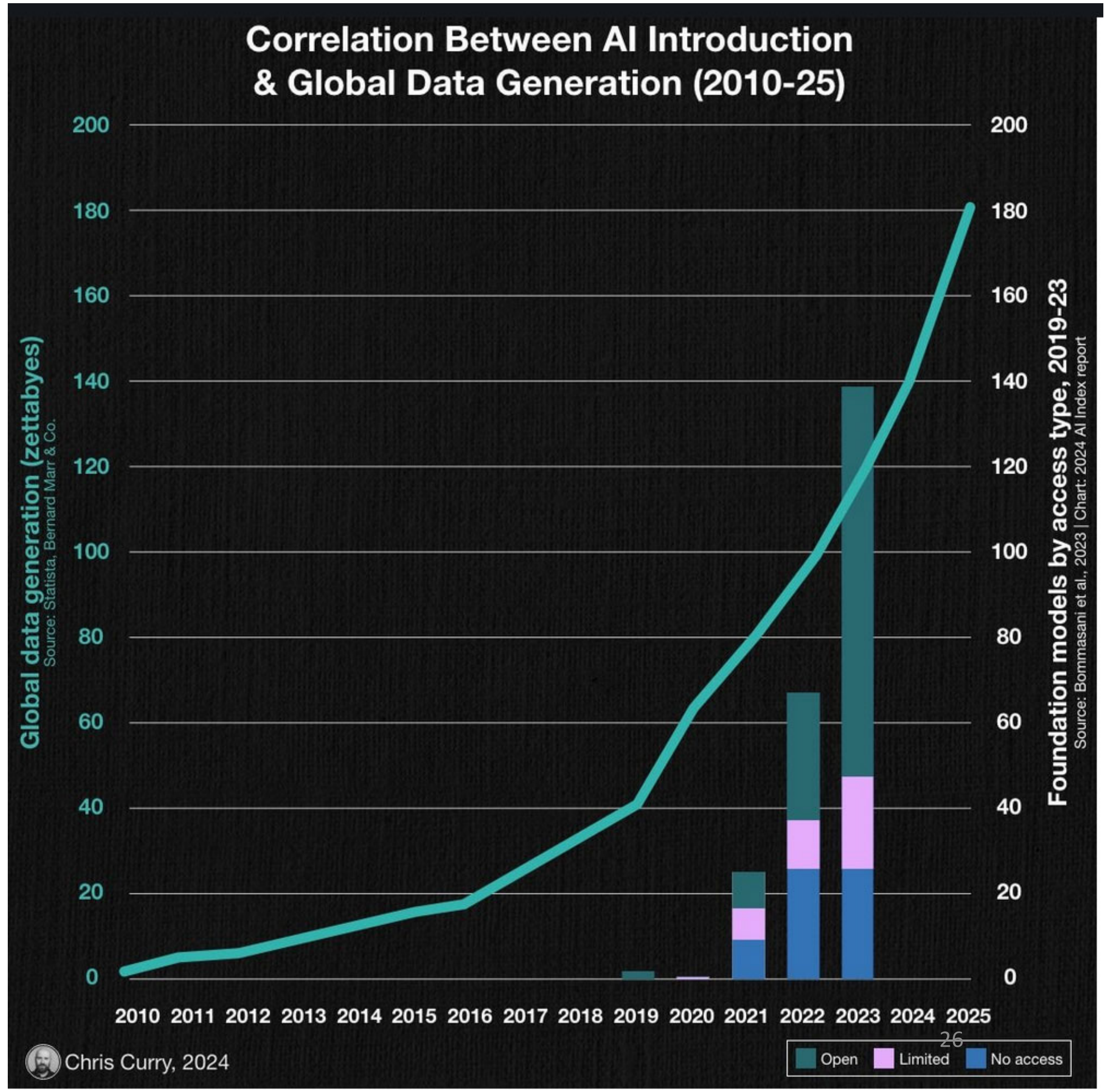
- Using GPT-4 to generate 100 words consumes up to 3 bottles of water

Data generation vs AI Introduction

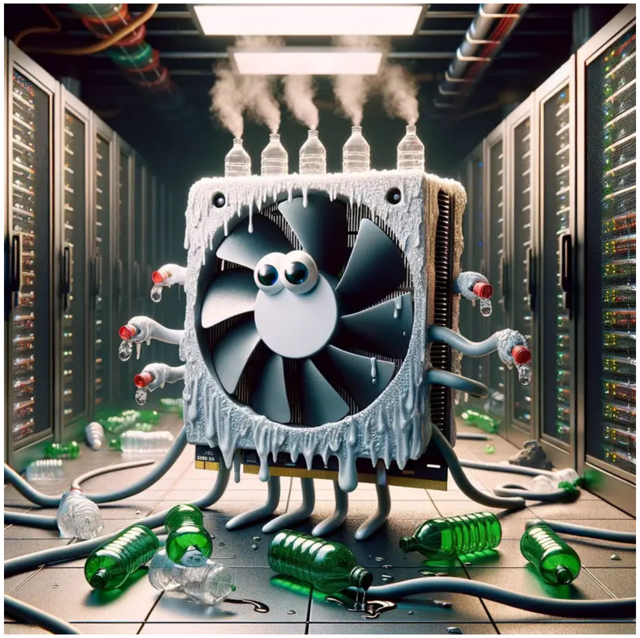
Shift from Passive to Purposeful Data Collection
→ AI requires structured, high-quality, and labeled data, so data generation has become more strategic and goal-driven.

Automated Data Labeling & Augmentation
→ AI tools (like computer vision) are now used to annotate and augment datasets, speeding up and scaling the generation process.

Synthetic Data Creation
→ AI models can generate synthetic datasets when real data is limited, especially valuable in medical or rare-event contexts.



Energy Footprint of AI Training



Training Large Models

- GPT-3 training required: ~3.14×10^{23} FLOPs
- ≈ 552 metric tons CO_2e (equivalent to 125 round-trip flights between NYC and London)

Carbon Emissions & Water Usage

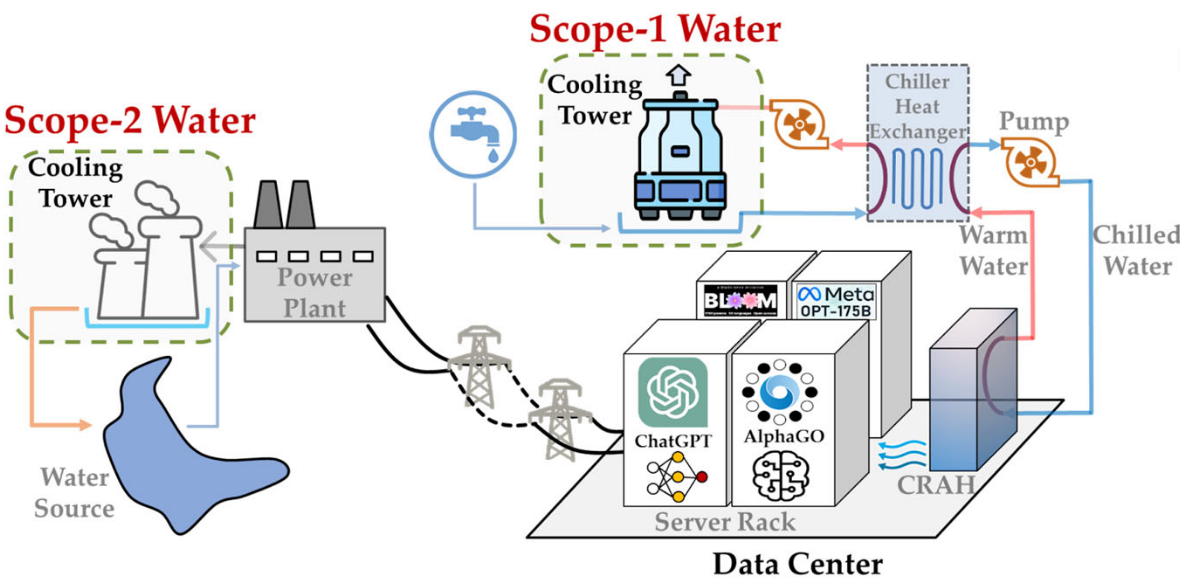
- Average model training (1 GPU over 1 week): ~0.5 tons CO_2e
- Data center cooling (per training run): ~700,000 liters of water used for cooling per 1 MWh of compute energy (equivalent to 4,600 bathtubs)

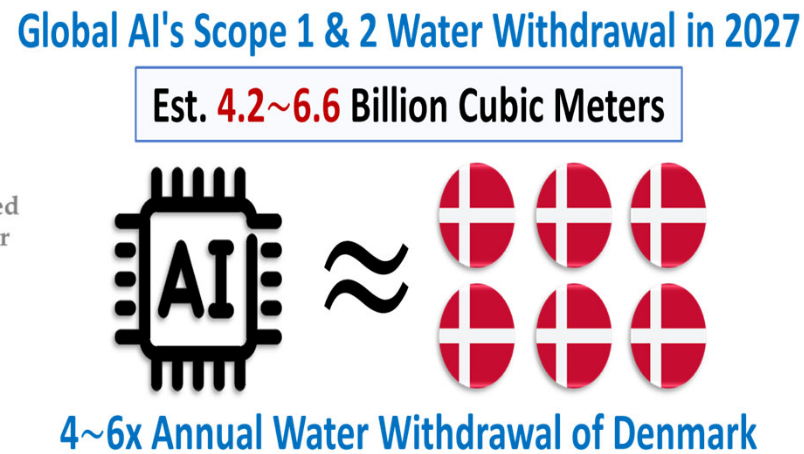
> - Google's AI subsidiary, DeepMind, applied machine learning to enhance the efficiency of Google's data centers, achieving a 40% reduction in energy use for cooling.
> - This advance translates to a 15% reduction in overall Power Usage Effectiveness (PUE) overhead

Thirsty AI data Centers

AI data centre's operational water usage:

- on-site scope-1 water for server cooling (via cooling towers in the example)
- off-site scope-2 water usage for electricity generation.





https://oecd.ai/en/wonk/how-much-water-does-ai-consume

Research & Innovation for a Greener AI Cloud

Efficient AI Models

- **Model distillation**: Compress large models into smaller ones with minimal accuracy loss.
- **Quantization & pruning**: Reduce precision and unnecessary weights to lower compute and memory use.

Green Software Engineering

- **Energy-aware coding**, efficient algorithms, and adaptive compute scheduling.
- Promote **carbon-aware deployments** & **open-source energy profiling tools**.

Future Outlook

- **Neuromorphic computing**: Brain-inspired chips (e.g., spiking neural nets) offering ultra-low power AI.
- **Photonic computing**: Light-based computation for faster, energy-efficient data processing.

Part II: Edge AI, IoT & Data Proliferation

Edge Computing and AI Shift

- Why move AI to the edge?

  Latency, privacy, bandwidth, data reduction

- Explosion of IoT devices and embedded AI. Examples: smart homes, health wearables, autonomous vehicles, environmental monitoring, etc.

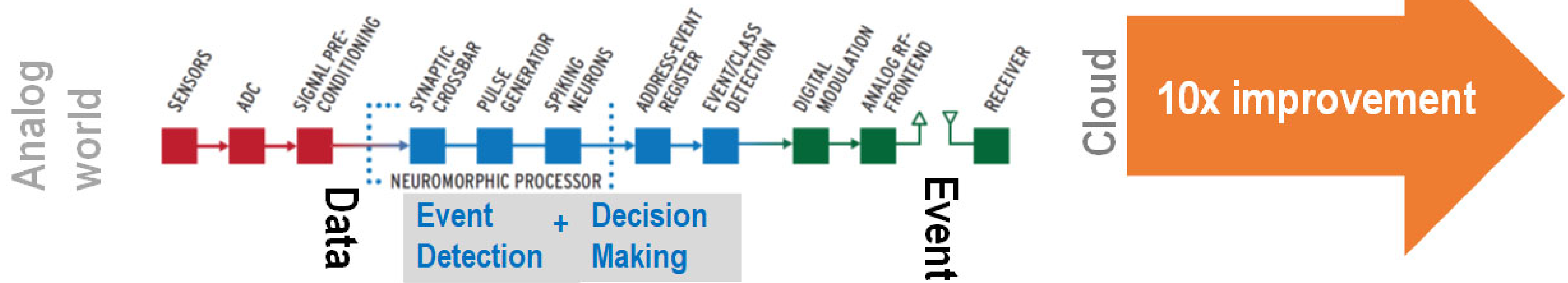
  1 trillion sensor planet, battery operated, electronic waste


How Much of Our Brain Do We Use?
The 10% myth.

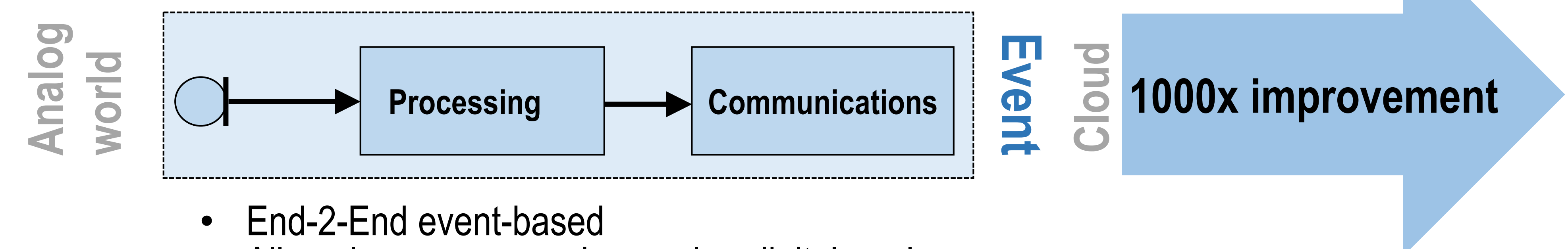
EPFL

Future Solution TINY BRAINS @ the Edge

Research sensing + processing + communications



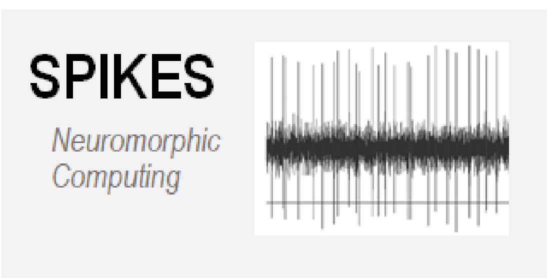
Radically new approach by SWIMS©

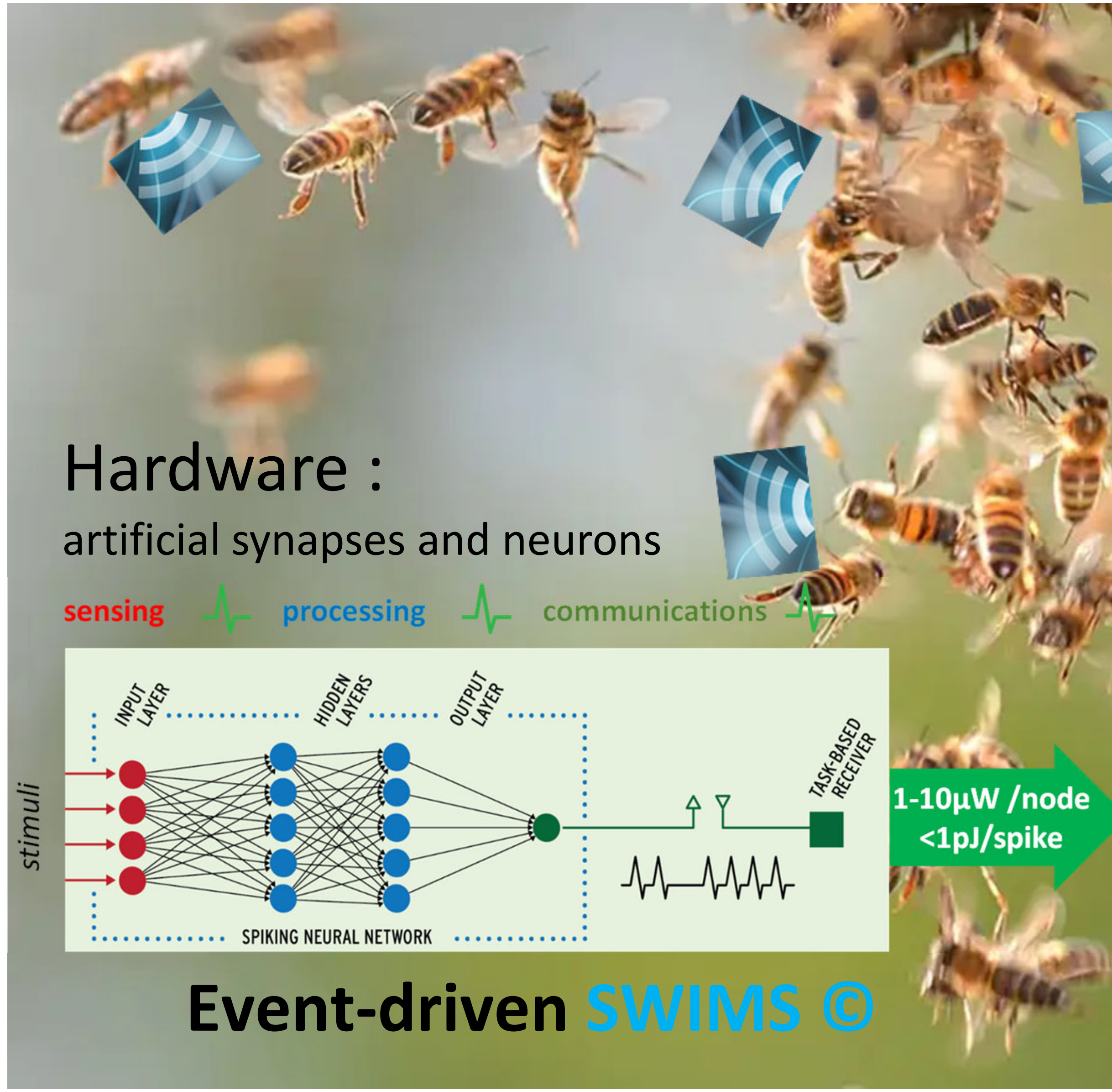


- End-2-End event-based
- All analog, no conversion analog-digital-analog
- No data stored & communicated (privacy preserved)

Neuromorphic Edge = Tiny Brains

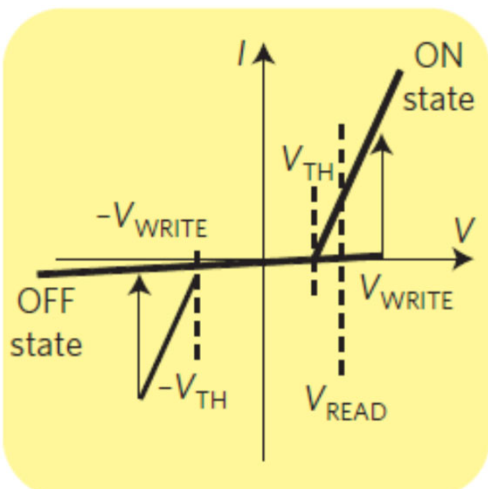
- Autonomous systems
- Decision making on the Edge
- **Real-time, energy efficiency**
- Adaptable, bio-inspired
- **Spiking Neural Networks: continuous, time-domain.**



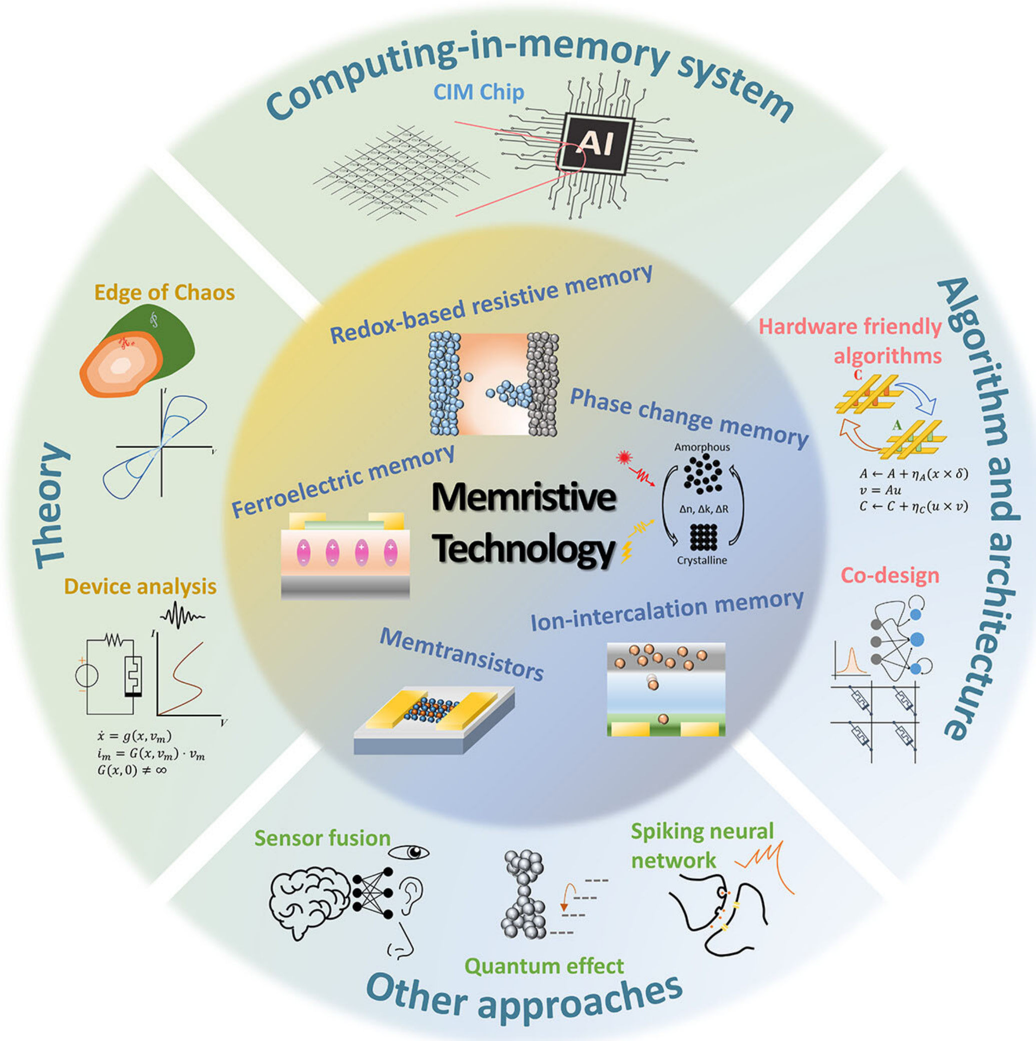


Memristive technologies for the Edge

- **can retain a state of internal resistance based on the history of applied voltage and current**



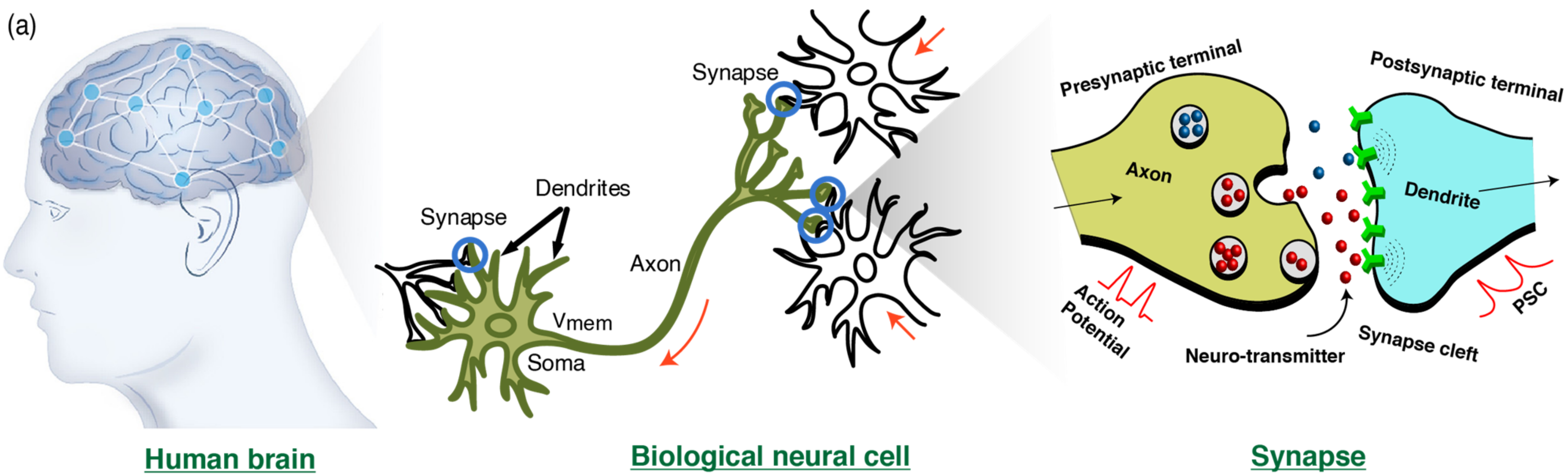
Yang, J., Strukov, D. & Stewart, D. Nature Nanotech (2013).



Song M.K. et al., ACS Nano **2023**

Biological and artificial synapses

- Synapses **transfer information between neurons** and transform this information.
- Artificial device **with programmable conductivity** = weight

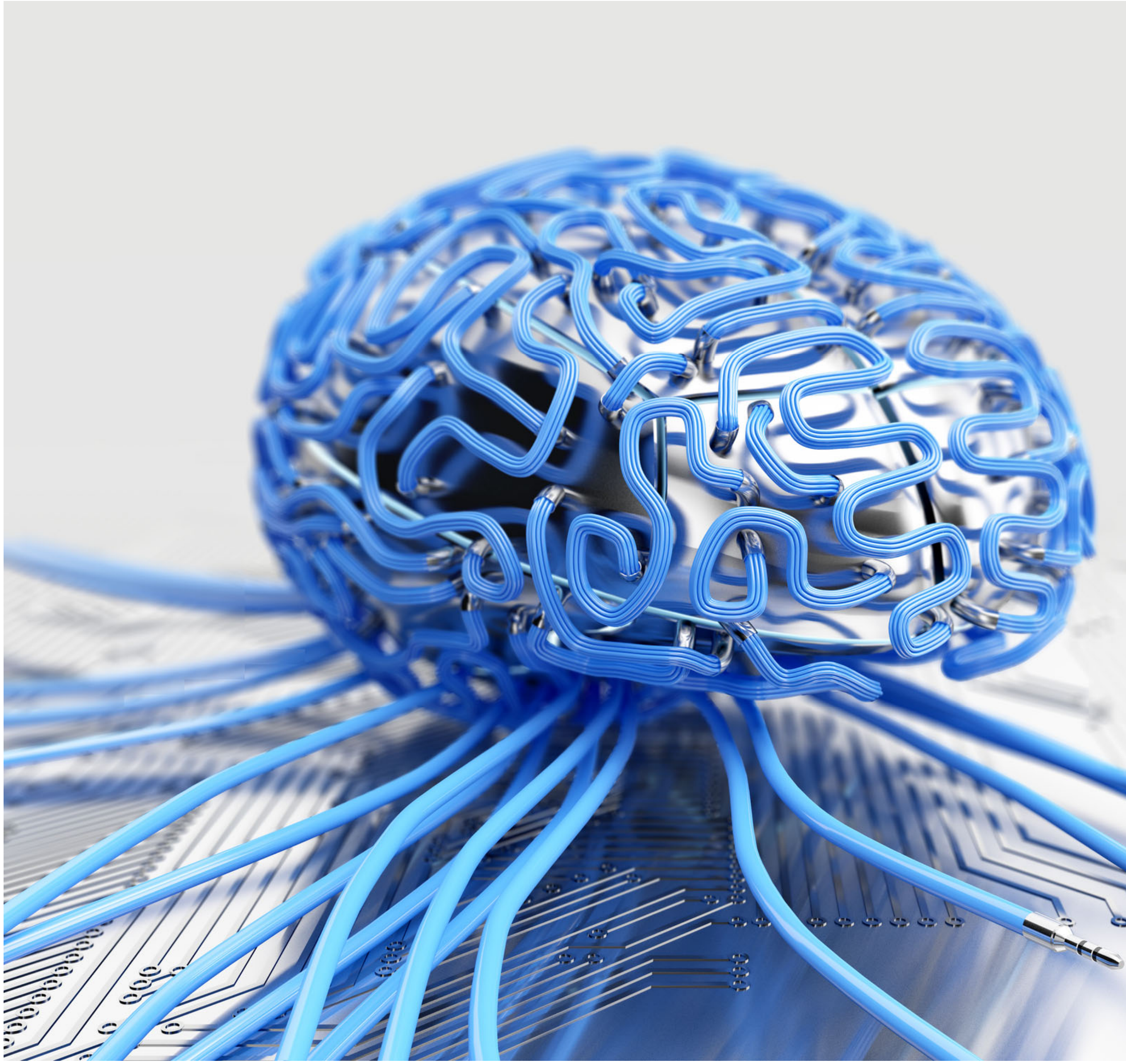


S. Kamaei, PhD thesis, EPFL, 2023.

Cognitive 3D chips

- **Imagine future chips that can sense, learn, infer and interact…**

"**Cognitive systems are probabilistic**, meaning they are designed to adapt and make sense of the complexity and unpredictability of unstructured information," John E. Kelly, senior vice president IBM Research



Smart Data Management of Edge to Cloud AI

Cost of Data Transmission

- High bandwidth use = increased operational cost & energy demand
- Data transmission can account for 30–50% of total system energy in edge-heavy deployments
- Costs scale non-linearly with latency sensitivity and geographical spread

Edge-Cloud Coordination: When to Offload?

- Trade-off between latency, accuracy, and energy usage
- Offload only when:
  - Local compute is overloaded
  - Cloud provides significant performance boost
  - Network is stable and low-latency
- Use dynamic decision models for offload policies (e.g., reinforcement learning)

Data Summarization & Compression at Source

- Apply feature extraction, compression (e.g., entropy coding, quantization) before transmission
- Reduces transmission volume by 10×–100× in many IoT/AI scenarios
- Enhances privacy and energy efficiency

Energy Efficiency Techniques for Edge AI

TinyML & Model Optimization

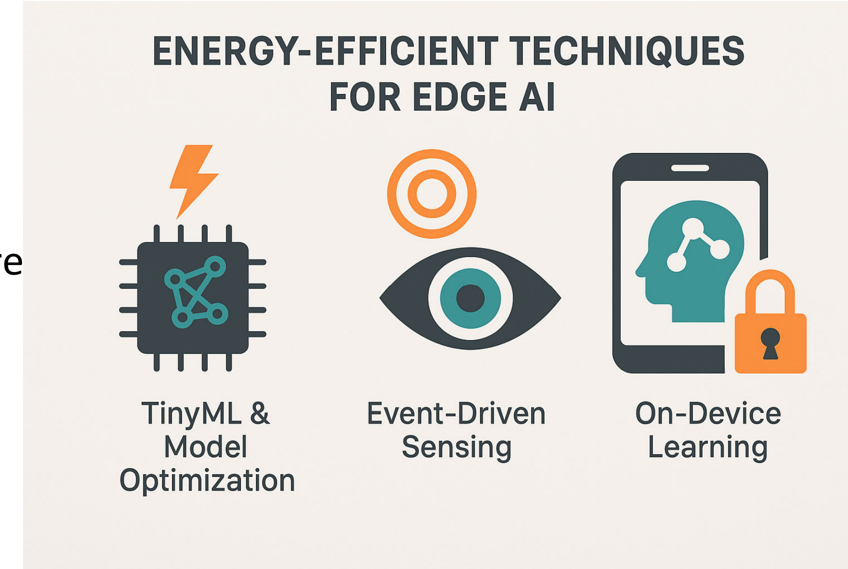
- TinyML: Running ML models on ultra-low-power microcontrollers (µW–mW)
- Model compression: reducing model techniques
- Hardware-aware training: Co-designing models for specific hardware constraints

Event-Driven Sensing

- Uses neuromorphic sensors (e.g. multi-modal sensors) to capture data only when meaningful events occur
- Significantly reduces energy and bandwidth

On-Device Learning

- Federated learning: Trains models locally and shares only updates—no raw data transmission
- Continual learning: Enables devices to adapt over time without full retraining
- Reduces cloud dependence and ensures data privacy



Electronic waste for IoT & Edge AI

⚠ Challenges

- Massive device deployment: Billions of sensors, wearables, and edge devices.
- Short product life cycles: Rapid obsolescence leads to early disposal.
- Difficult recycling: Miniaturized, composite components are hard to disassemble and recycle.
- Toxic materials: Batteries, PCBs, and rare earth metals pose environmental risks.
- Low recovery rates: Only ~17.4% of global e-waste was formally documented as collected and recycled (UN, 2020).

☑ Sustainable Measures

- Design for Circularity: Modular, repairable, and upgradeable hardware.
- Biodegradable electronics: Emerging materials for temporary or low-power edge devices.
- E-waste regulation compliance
- AI-powered asset tracking: Smarter lifecycle monitoring and recycling.
- Energy harvesting IoT: Reduces reliance on disposable batteries.

Wrap-Up & Key Takeaways

- AI sustainability must be viewed systemically/holliustically:
  - ✓**from cloud to edge**
- Innovations needed:
  - ✓ **in both hardware and algorithms**
  - ✓ **in policies**
- Transparency, regulation, and standardization will be crucial
- Research directions:
  - ✓**embodied energy of AI**
  - ✓**end-to-end LCA**

Add-ons / Class Activities

- Quick debate: "Is training large AI models ethically justifiable?"
- Case analysis #1: "Estimate CO_2 footprint of a GPT-3 or -4 query using online tools and propose counter measures to reduce it."
- Case analysis #2: "Compare carbon commitments of cloud providers"
- Case analysis #3: "Discuss advantages versus challenges when moving AI to the Edge."