

# Energy Consumption of Datacenters

Thomas Fricke

October 31, 2024  
SRECON EMEA 2024, Dublin





# Innovationsverbund Öffentliche Gesundheit

Thomas Fricke

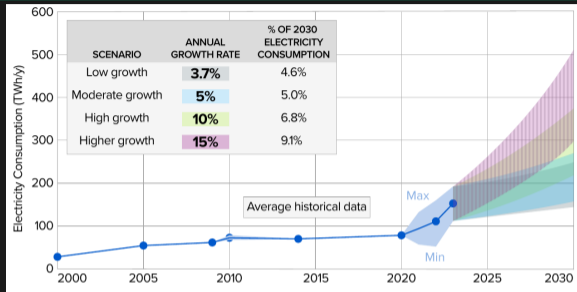
- ▶ Kubernetes Cloud Security
- ▶ Statistical Physics
- ▶ Disclaimer
  - ▶ Pro Bono: OpenCode, Beratung IT Planungsrat
  - ▶ Payed: OpenDesk, FITKO



# Electric Power Research Institute – EPRI Projections, May 2024

From Bloomberg: **Sam Altman's Energy 'New Deal' Is Good for AI. What About Americans?**

- ▶ Demand per Hyperscaler 5GW (roughly 8-10 power station blocks)
- ▶ Total 47GW (> 150 reactors of 300 MW)
- ▶ Small ... Reactors Have A Big Problem
- ▶ China will lead in 2030
- ▶ Retain US leadership in AI
- ▶ US Gov: AI linchpin of our economy
- ▶ AI New Deal
- ▶ Illinois: \$468 million in subsidies for only 339 jobs (\$1.4 million per job)
- ▶ Nebraska: costs for Google and Meta passed onto residents, estimated rate increase 2.5% to 3% per year
- ▶ Datacenters are extremely unequally distributed (Chicago, Texas, Virginia)



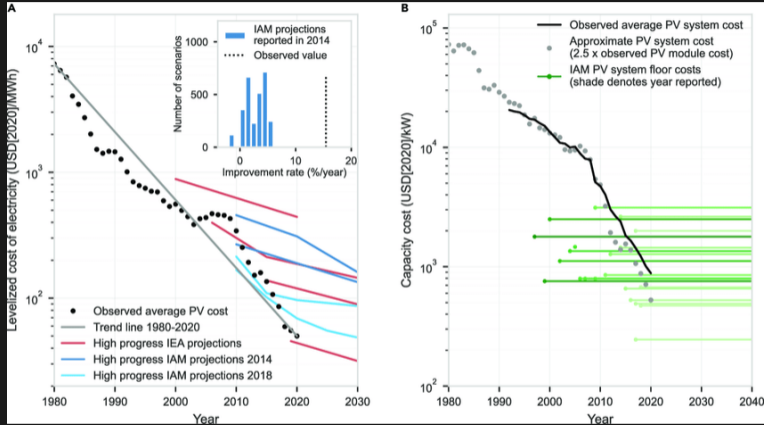
1 TWh /y = 0.114 GW  $\approx$  1/7 GW

Electric Power Research Institute –  
EPRI Projections, May 2024

- ▶ Ireland: 20% of electricity consumption
- ▶ Energy Consumption in Data Centres and Broadband Communication Networks in the EU



# Remote Nuclear Fusion

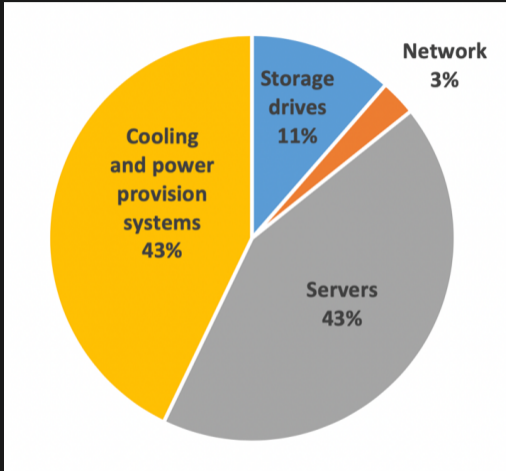


Rupert Way, Matthew C Ives, Penny Mealy University of Oxford, J. Doyne Farmer:

Empirically grounded technology forecasts and the energy transition, September 2022



# Power Usage in a typical Data Center

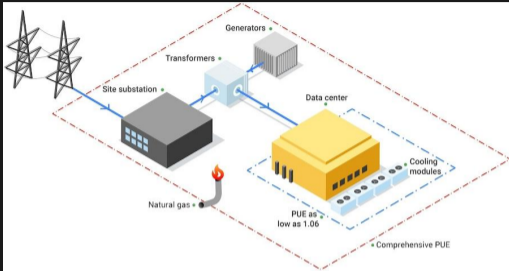


Fraction of U.S. data center electricity use in 2014, by end use. Source: Shehabi 2016

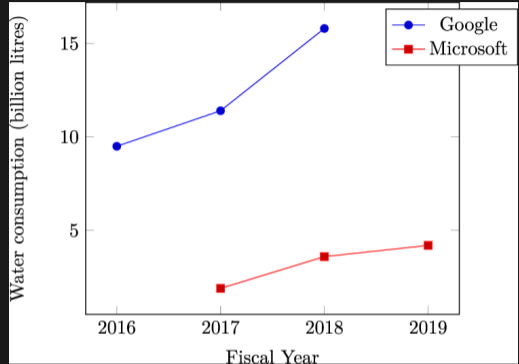
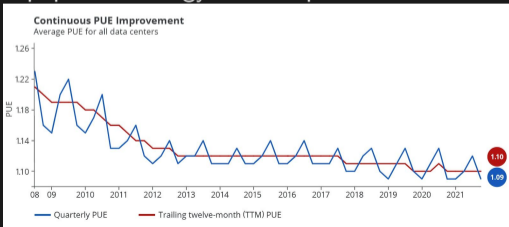
- ▶ mostly wasted
- ▶ not changed in the last 10 years
- ▶ cooling and power as much as compute
  - ▶ water evaporation
  - ▶ immersion cooling
  - ▶ adiabatic cooling



# Google Power Usage Effectiveness – PUE Greenwashing



Centre Total Energy Consumption PUE= ICT Equipment Energy Consumption

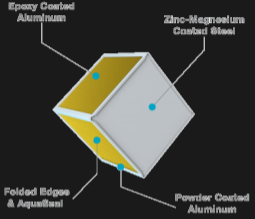
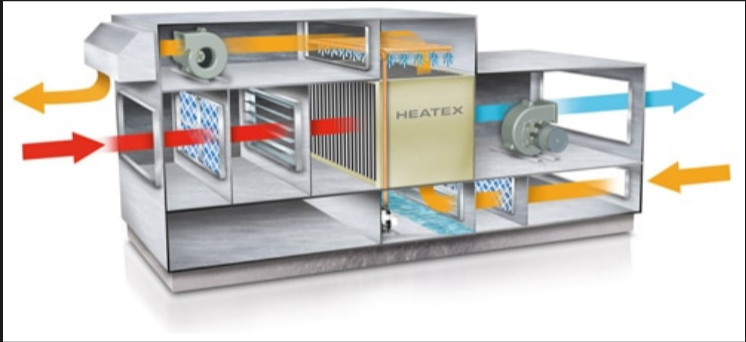


$$PUE = \frac{\text{Data Centre Total Energy Consumption}}{\text{ICT Equipment Energy Consumption}}$$

Source: Google(left), Nature (right)



# Cooling



Source: Heatex



# Water cooling for Machine learning

$$p_t = 1.58 \frac{t (p_c + p_r + gp_g)}{1000} \text{ kWh}$$

The diagram illustrates the equation for training energy consumption  $p_t$  in kWh. The equation is  $p_t = 1.58 \frac{t (p_c + p_r + gp_g)}{1000}$ . The value 1.58 is highlighted in blue and labeled 'PUE'. The variable  $t$  is highlighted in green and labeled 'training time'. The terms  $p_c$ ,  $p_r$ , and  $gp_g$  are highlighted in red and labeled 'CPU', 'DRAM', and 'GPU' respectively, with a bracketed group labeled 'power draw'.





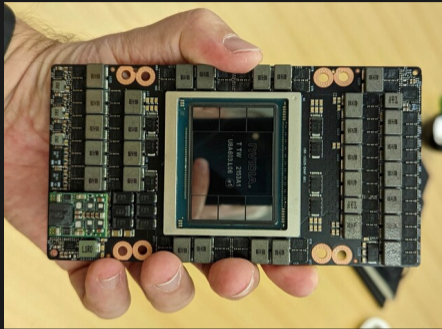
## Rough thumbrules, lies damn lies and PuEs

PuE	Technology	?
1.03	evaporation	transformers ignored
1.34	adiabatic cooling	transformers included
1.58	water cooling	for high density AI workloads
1.057	pPuE	transformers only

Benjamin Petschke: My understanding of PUE and pPUE  
(2015)



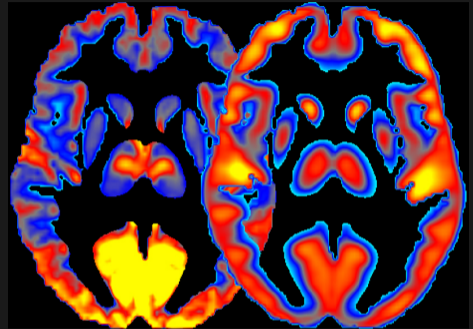
# Comparison NVIDIA Hopper H100 vs Homo Sapiens<sup>2</sup>



700 Watts

## Energy Consumption

- ▶ Single Graphics Card
- ▶ 700 Watts = 0.7kW
- ▶ ~ 30 100 kW / rack
- ▶ instead of 3 to 6 KW / rack



20 Watts

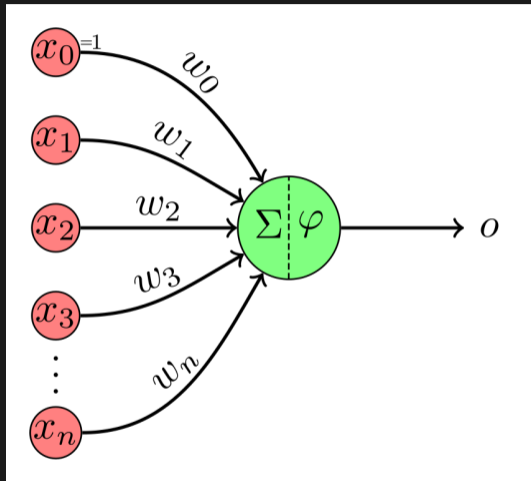
New method for combining measures of brain activity (left) and glucose consumption (right)

...

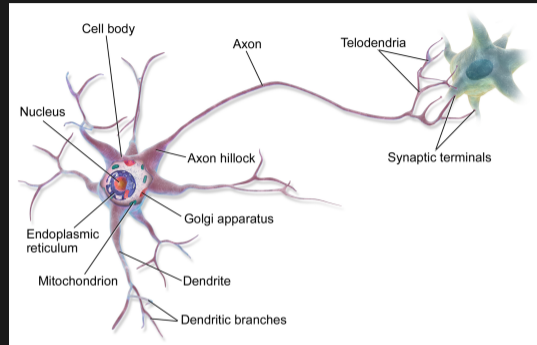
Dr. Ehsan Shokri Kojori, NIAAA



# Some Inconvenient Truth



Perceptron



Neuron

- ▶ The AI neuron is not even a biological synapse
- ▶ The synapse computes and has the complexity of some handful of perceptrons

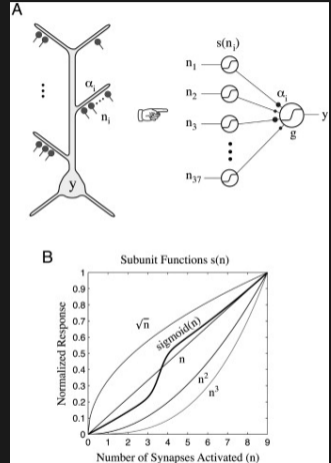


# Pyramidal Neuron as Two-Layer Neural Perceptron Network

*We found the cell's firing rate could be predicted by a simple formula that maps the physical components of the cell onto those of an abstract two-layer "neural network." In the first layer, synaptic inputs drive independent sigmoidal subunits corresponding to the cell's several dozen long, thin terminal dendrites.*

Pyramidal Neuron as Two-Layer Neural Network  
by Panayiota Poirazi, Terrence Brannon, Bartlett W. Mel, 2003

- ▶ article is old
- ▶ simulation of real firing synapses
- ▶ consistent result
- ▶ hundreds of different types of synapses
  - ▶ chemical
  - ▶ electrical



Neuronal Network Tree



## NVIDIA Tensor Core Datasheet

- ▶ *Built with 80 billion transistors using a cutting-edge TSMC 4N process custom tailored for NVIDIA's accelerated compute needs, H100 is the world's most advanced chip ever built*
- ▶ **Blackwell B-200:**
  - ▶ roughly *H100 x2.5 208 Billion Transistors*
  - ▶ *LLM Inference x30*
  - ▶ *LLM Training x4*
  - ▶ *Energy Efficiency ? x25x*
  - ▶ *Data Processing ? x18 vs. CPU !*

## Basic Neural Units of the Brain: Neurons, Synapses and Action Potential by Jiawei Zhang

*On average, the human brain contains about 100 billion neurons and many more neuroglia which serve to support and protect the neurons. Each neuron may be connected to up to 10,000 other neurons, passing signals to ... as many as 1,000 trillion synapses.*

- ▶ German Milliarde: American Billion =  $10^9$
- ▶ German Billion: American Trillion =  $10^{12}$



# Comparison

Transistor  $\approx$  Synapse

4.000 Blackwell  $\approx$  10.000 H100  $\approx$  Brain

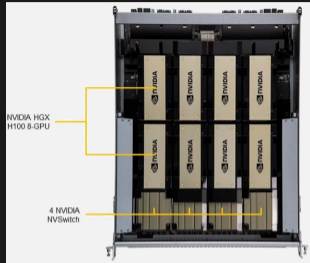
7 MW  $\equiv$  20 W

x10 Transistors and x2 for cooling and network  $\approx$  140MW (50MW Blackwell)  
that is the true reason why the Matrix AI is using humans to live in

Average usage of a GPU in Kubernetes is 20%



# Racks



Nvidia Rack

- ▶ Key Applications : High Performance Computing, AI, Deep Learning and Industrial - Automation.
- ▶ Dual AMD EPYC 9004 Series Processors (Socket SP5)
- ▶ 8x NIC for GPU direct RDMA (1:1 GPU Ratio)
- ▶ High density 8U system with NVIDIA® HGX™ H100 8-GPU
- ▶ Highest GPU communication using NVIDIA® NVLINK™ + NVIDIA® NVSwitch™
- ▶ 24x DIMM Slots, Up to 6TB DRAM, 4800 ECC DDR5 LRDIMM;RDIMM;
- ▶ 8x PCIe Gen 5.0 X16 LP, and up to 4 PCIe Gen 5.0 X16 FHFL Slots
- ▶ Flexible networking options
- ▶ 1x M.2 NVMe for boot drive only
- ▶ 2x 2.5" hot-swap NVMe/SATA drive bays (12x 2.5" NVMe dedicated)
- ▶ 2x 2.5" Hot-swap SATA drive bays
- ▶ 10x heavy duty fans with optimal fan speed control
- ▶ 6x 3000W redundant Titanium level power supplies



Rittal Megawatt Cooling



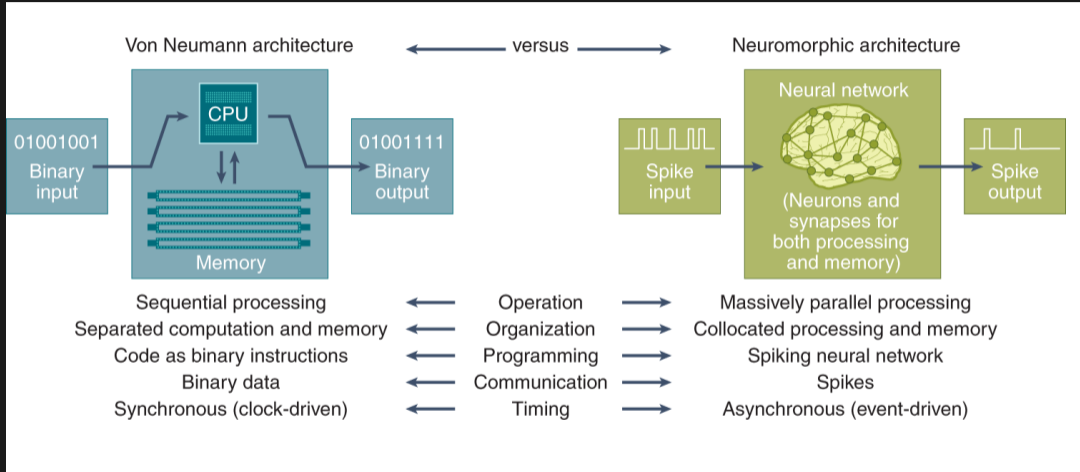
# Price of Trainings

- ▶ Everybody is complaining
- ▶ Access to Compute Power is **the gatekeeper**
- ▶ 1.2 M€ for an academic research
- ▶ Building a datacenter
  - ▶ starts at 300M€
  - ▶ planning several years
  - ▶ lack of H100
- ▶ training in the US clouds
  - ▶ coal and gas power plants
  - ▶ good bye sovereignty
  - ▶ dependency





# Neuromorphic Computing – Nature



# Neuromorphic Computing – Intel

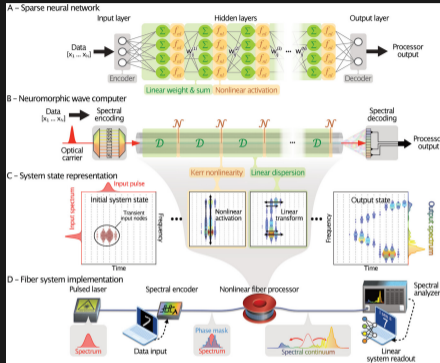
*Intel Labs' second-generation neuromorphic research chip, codenamed Loihi 2, and Lava, an open-source software framework, will drive innovation and adoption of neuromorphic computing solutions.*

Enhancements include:

- ▶ Up to 10x faster processing capability<sup>1</sup>
- ▶ Up to 60x more inter-chip bandwidth<sup>2</sup>
- ▶ Up to 1 million neurons with 15x greater resource density
- ▶ 3D Scalable with native Ethernet support
- ▶ A new, open-source software framework called Lava
- ▶ Fully programmable neuron models with graded spikes
- ▶ Enhanced learning and adaptation capabilities

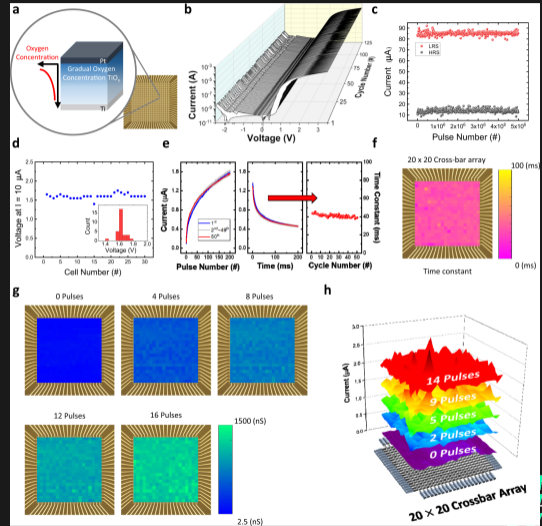


# Neuromorphic Computing – Optical and Memristor

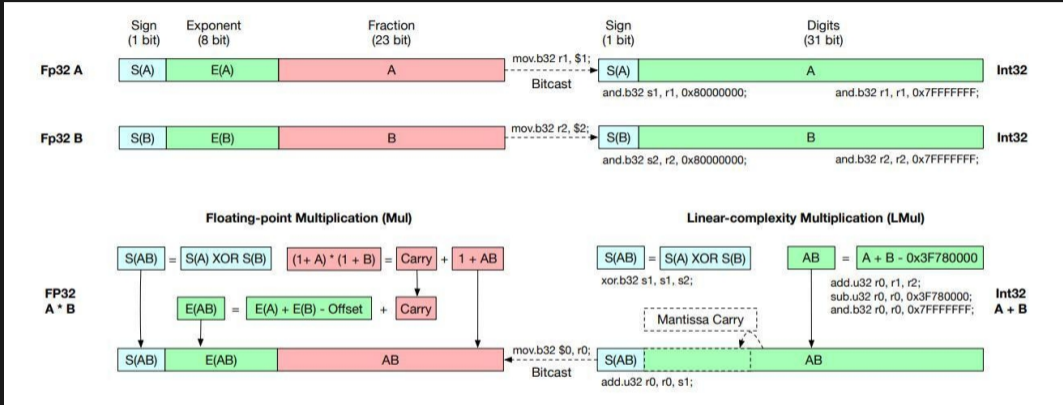


Neural Networks Made of Light: Jena Research Team Develops AI System in Optical Fibers

Experimental demonstration of highly reliable dynamic memristor for artificial neuron and neuromorphic computing



# Use more Integers! The L-Mul Algorithm



Addition is All You Need for Energy-efficient Language Models by Hongyin Luo, Wei Sun


Could save 95% of the energy needed



# Where is the positive? We want AI!

Illustrations: Niklas Elmehed

THE NOBEL PRIZE  
IN CHEMISTRY 2024



David Baker  
"for computational protein design"

Demis Hassabis  
"for protein structure prediction"

John M. Jumper

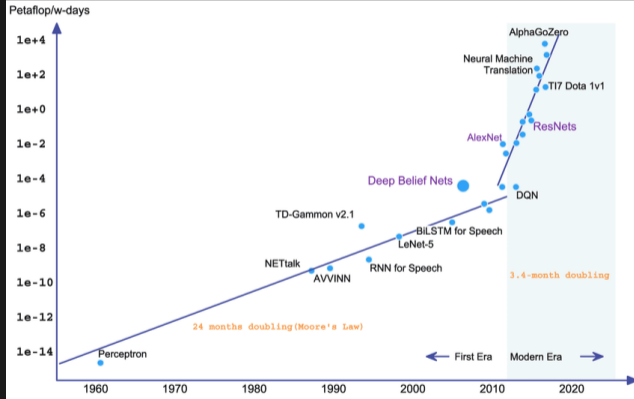
THE ROYAL SWEDISH ACADEMY OF SCIENCES

## Alpha Fold Protein Folding



# Moore's Law for Training Neural Networks

## How AI will really kill us



Moore's Law by Open AI AI and Compute

- ▶ H100
  - ▶ 10.6 TFlops single precision
  - ▶ 5.3 TFlops double precision
- ▶ 10000 TFlops
  - ▶ 1000 H100 single precision
    - ▶ 700 kW
  - ▶ 2000 H100 double precision
    - ▶ 1400 kW
  - ▶ cooling
    - ▶ PUE=1.6

some 67 MW hours

# Explosion



## Exponential Growth

- ▶ explosives
  - ▶ nuclear chain reactions
  - ▶ population growth
  - ▶ infections at the beginning of an epidemic
- SIR Model**
- ▶ limited by resources



# Touching Limits: Energy, Water, Metal CO<sub>2</sub>

- ▶ Ireland: Microsoft and Amazon reportedly halt plans to build data centers . . .
- ▶ Netherlands: Inside the data centre moratorium movement
- ▶ Tech HQ: Heating up: how much energy does AI use? *What we do know is that training ChatGPT used 1.287 gigawatt hours, roughly equivalent to the consumption of 120 US homes for a year.*
- ▶ Moomoo: Chicago data center electricity demand increased by 900%! AI continues to detonate global energy challenges
- ▶ Cleanroom Technology: data centers run out of power
- ▶ Business Today: OpenAI might go bankrupt by end of 2024
- ▶ Business Insider: The AI boom will push America's shaky power grid to its limit
- ▶ Wired: AI's Energy Demands Are Out of Control. Welcome to the Internet's Hyper-Consumption Era
- ▶ OECD: How much water does AI consume? The public deserves to know
- ▶ Substack: The Great Salt Lake is Disappearing. So, Utah Banned the Rights of Nature.
- ▶ Straight Arrow News: AI tools consume up to 4 times more water than estimated
- ▶ Substack: Material Sacrifices To tackle climate chaos, decolonize the labor movement
- ▶ The Driller: Growing Demand for Copper Drives Need for Increased Domestic Mining, Experts Suggest
- ▶ Generative AI is reportedly tripling carbon dioxide emissions from data centers
- ▶ Odessa American Online: AI to boom natural gas market
- ▶ Arabian Gulf Business Insight: Aramco partners with US startup Groq for AI data centre





# AI goes nuclear? Not really

- ▶ Microsoft is training an AI to help get nuclear reactors approved

*We're really excited about the game-changing potential for AI in this space*

MICHELLE PATRON (MS director of sustainability)

- ▶ Is advanced nuclear in trouble? What's next after NuScale cancellation



- ▶ Guardian:
  - ▶ Google to buy nuclear power for AI datacentres in 'world first' deal
  - ▶ Three Mile Island nuclear reactor to restart to power Microsoft AI operations

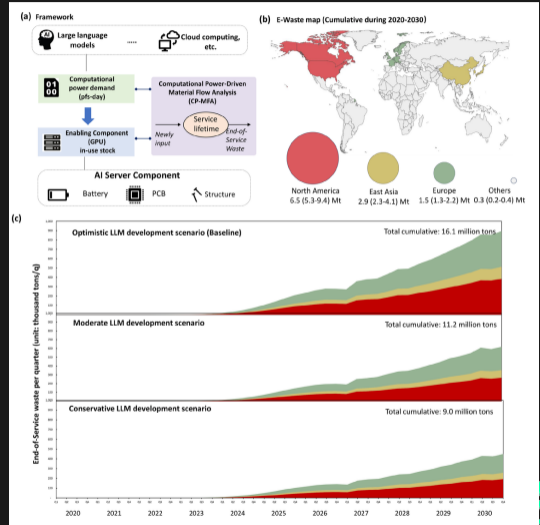
## Scam the scammers

- ▶ Getting new nuclear reactors approved by regulators is an expensive, complex process.
  - ▶ Planning 5 years
  - ▶ construction 5 years
  - ▶ construction delays are typical for common reactors
  - ▶ no experience with the new design
  - ▶ Researchgate IAEA: Typical timeline of a nuclear plant construction and start-up project
- ▶ Big companies fall in the same FOMO promotional traps they build for politics
- ▶ same as for gas and oil powered plants



# AI-Waste

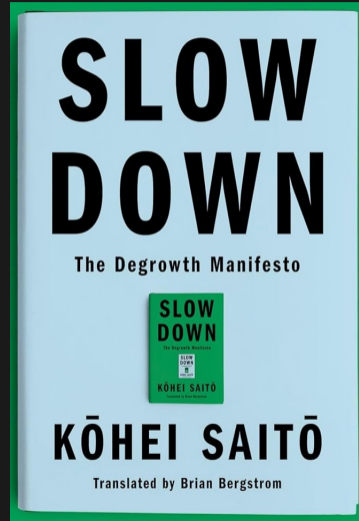
- ▶ Lifecycle of Data Center Hardware: **3 – 5 years**
- ▶ Peng Wang, Chinese Academy of Sciences, Lingyu Zhang, Institut National des Sciences Appliquées de Lyon, Asaf Tzachor, Eric Masanet, University of California, Santa Barbara:  
**E-waste Challenges of Generative Artificial Intelligence**  
also in **Nature**
- ▶ **1000** fold increase in creation of waste



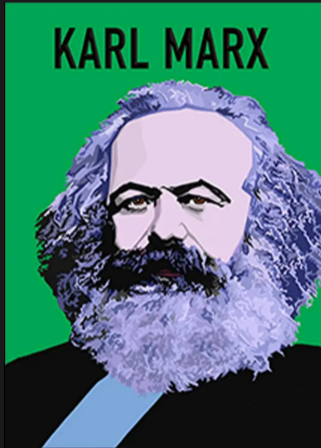
E-Waste

# Degrowth

- ▶ if you don't kill exponential growth, the explosion will kill **us**
  - ▶ **our** economy
  - ▶ **our** energy grids
  - ▶ **our** business
  - ▶ **our** environment
  - ▶ **our** entire planet
  - ▶ all limits are nearly exhausted
  - ▶ further reading
- ▶ **Degrowth**
  - ▶ will kill nearly all of your business models
    - ▶ advertising
    - ▶ surveillance
    - ▶ selling without limits
    - ▶ keeping people busy
    - ▶ anything with **Growth**
- ▶ the current economy is like a junkie looking for money to buy drugs



# A spectre is haunting Europe



Karl Marx was Green

This was a joke

- ▶ nobody wants a communist party
- ▶ we have no working class any more
- ▶ revolutions are only successful after a war
- ▶ externalisation is real
- ▶ concentration of power is real
- ▶ Open Source and Wikimedia show that collaboration is real

But we can at least start to degrow

- ▶ non GPU AI
  - ▶ memristor
  - ▶ optical fibres
  - ▶ I-mult
- ▶ saving
  - ▶ energy
  - ▶ the energy grid
  - ▶ water
  - ▶ metals
  - ▶ carbondioxide

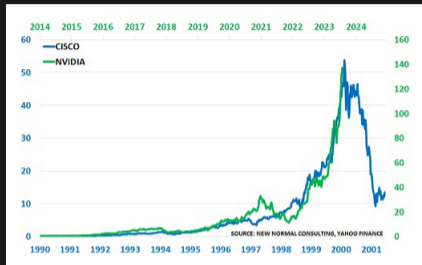


# Conclusion

*In from three to eight years we will have a machine with the general intelligence of an average human being.*

Marvin Minsky in Life magazine 1970

- ▶ Resource consumption of data centers is out of control
- ▶ Fine grained data necessary
- ▶ AI is oversold
- ▶ degrowth
  - ▶ start with different AI
  - ▶ degrow your workload
- ▶ will harm the planet on every possible scale
- ▶ charlatantry
- ▶ massive financial interest
- ▶ public protest



Stock market bubbles follow the same pattern, as Nvidia and Cisco confirm

- ▶ whatever resource is exhausted first will terminate the AI
  - ▶ money
  - ▶ energy
  - ▶ energy grid
  - ▶ metal resources



# Question? Remarks?

## Some Answers

Slides: <https://thomasfricke.de/srecon24.pdf>

Mail: [srecon24@thomasfricke.de](mailto:srecon24@thomasfricke.de)

Mastodon: [@thomasfricke@23.social](https://mastodon.social/@thomasfricke)

LinkedIn: <https://www.linkedin.com/in/thomas-fricke-9840a21/>

