Energy Consumption of AI

Thomas Fricke

December 27, 2023

Who are we?

Innovationsverbund Öffentliche Gesundheit

- Born in *WirVsVirus* Hackathon
- Funding Holistic Foundation
- Projects
  - Open Source
  - Privacy
  - Open Social Innovation
- Iris Connect
  Contact Tracking Public Health Departments Björn Steiger Stiftung

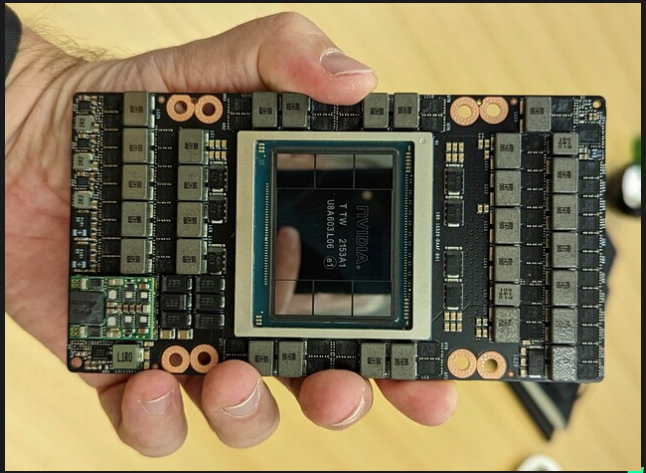
Thomas Fricke

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

Hardware NVIDIA Hopper H100

Energy Consumption

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

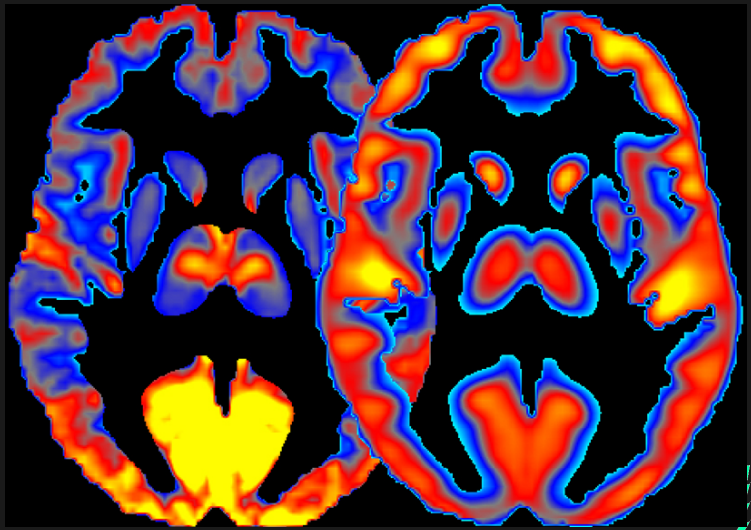


NVDIA Hopper H100 in a Hand

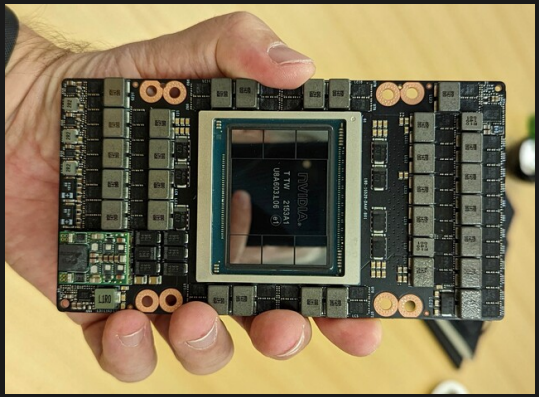
Legacy Model – Homo Sapiens Sapiens

NIH scientists present a new method for combining measures of brain activity (left) and glucose consumption (right) to study regional specialization and to better understand the effects of alcohol on the human brain.

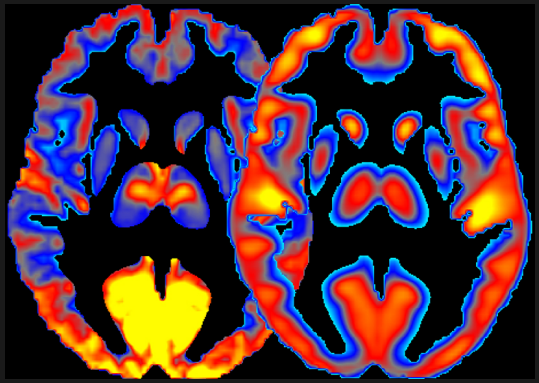
Dr. Ehsan Shokri Kojori, NIAAA



Comparison

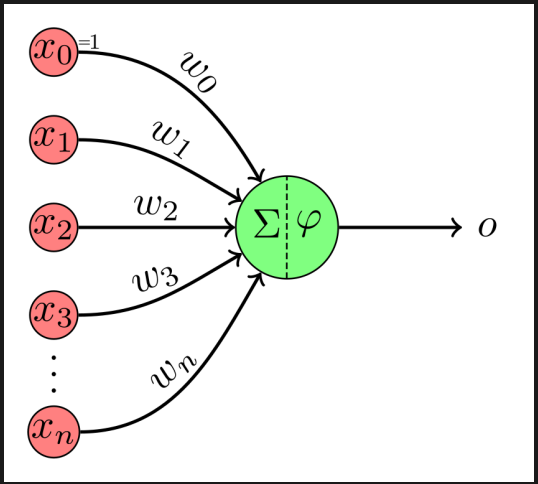


700 Watt

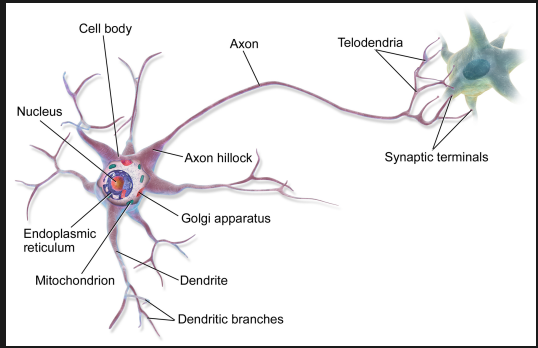


20 Watt

Some Inconvenient Truth



Perceptron



Neuron

- The AI neuron is not even a biological synapse
- The synapse computes an has the complexity of some handfull of perceptrons

Numbers

NVIDIA Tensor Core Datasheel

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

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

we will introduce the basic compositional units of the human brain, which will further illustrate the cell-level bio-structure of the brain. 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 each other via as many as 1,000 trillion synapses.

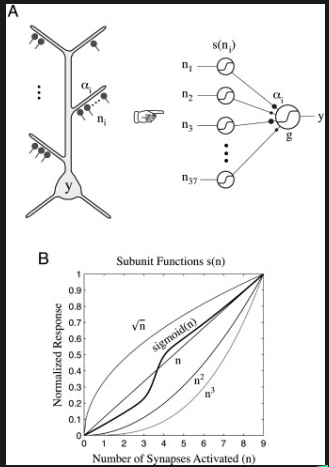
- German Milliarde: American Billion $= 10^9$
- German Billion: American Trillion $= 10^{12}$

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.

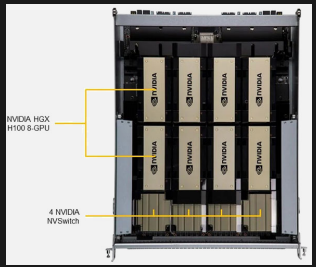
Pyramidal Neuron as Two-Layer Neural Network
by Panayiota Poirazi, Terrence Brannon, Bartlett W. Mel, 2003

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



Neuronal Network Tree

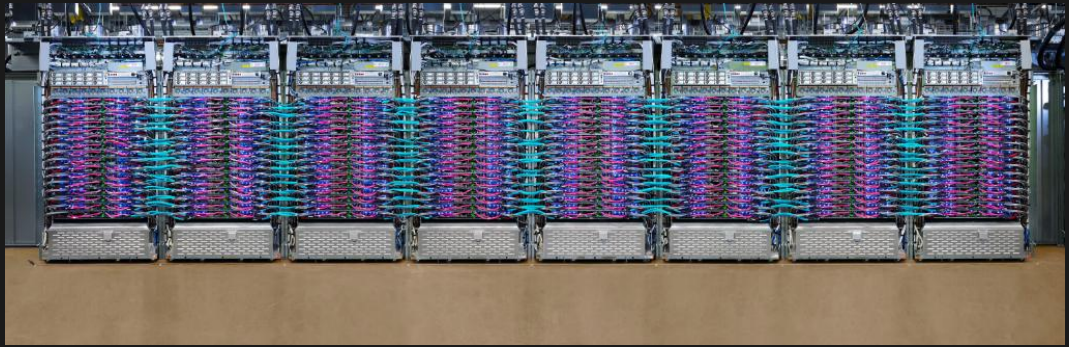
Racks



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

Datacenters

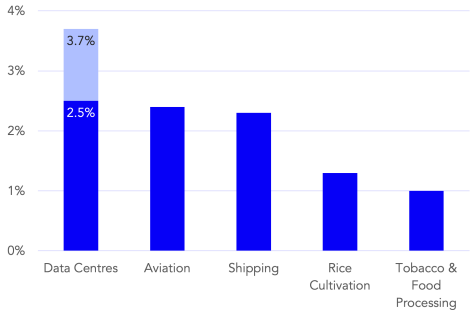


Liquid Cooling Google

Compared to other businesses

Global cloud computing emissions exceed those from commercial aviation

Share of global CO_2 emission generated by sector/category



Source: Climatiq Analysis, The Shift Project, OurWorldinData

climatiq

Hyperscale data centres are significantly more efficient than internal data centres

Category	Energy use (Million MWh)	Computing workloads (million)	Water intensity (M_3 MWh_{-1})	Carbon intensity (ton CO_2-eq MWh_{-1})	Water intensity (m_3/workload)	Carbon intensity (Ton CO_2-eq /workload)
Internal	26.90	16	7.20	0.45	12.15	0.75
Colocation	22.4	41	7.00	0.42	3.85	0.25
Hyperscale	22.85	76	7.00	0.44	2.10	0.15

Source: Siddik & Sehab 2021

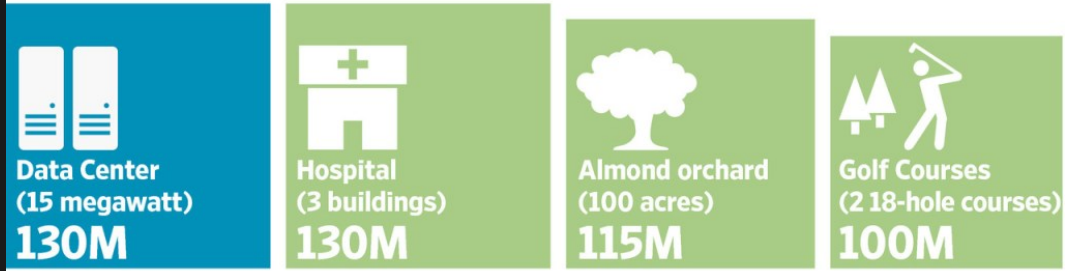
climatiq

Compared to other usages

Waterlogged

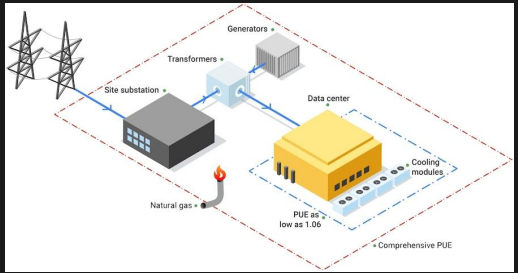
A midsize data center uses roughly as much water as about 100 acres of almond trees or three average hospitals, and more than two 18-hole golf courses.

Approximate annual water usage, in gallons*

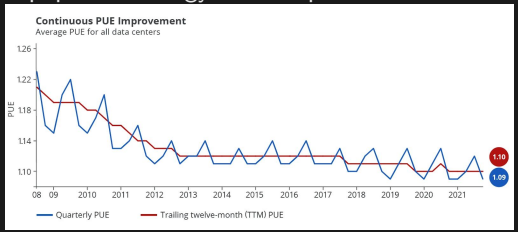


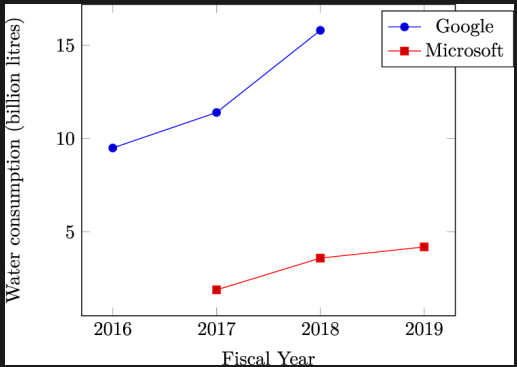
*Use varies depending on climate and other factors
Sources: California Department of Water Resources (orchards); James Hamilton (data centers); U.S. Department of Energy (hospitals); Golf Course

Google Power Usage Effectiveness – PUE Greenwashing



Centre Total Energy Consumption PUE= ICT Equipment Energy Consumption





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

Source: Google(left), Nature (right)

Stop Building Data Centers

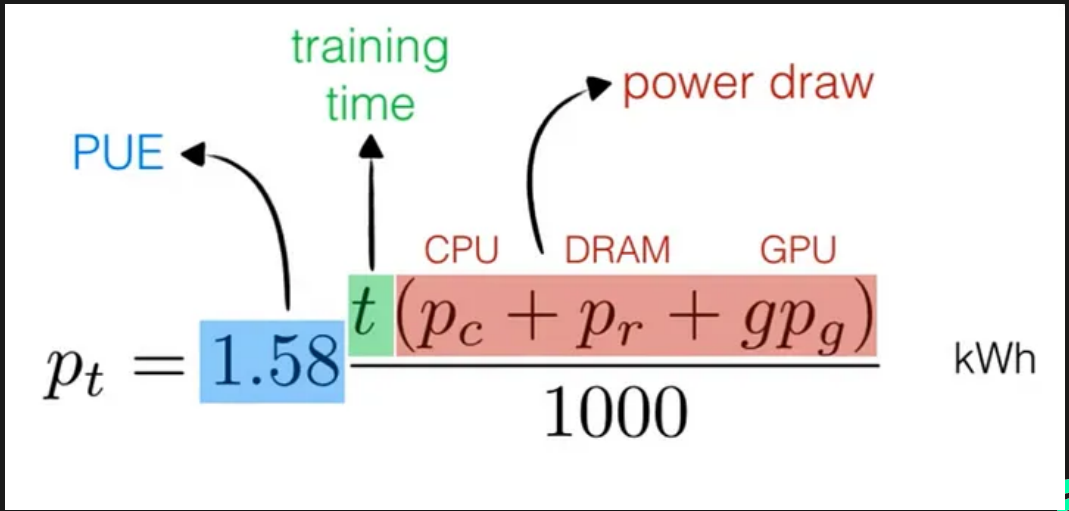
- Ireland: Microsoft and Amazon reportedly halt plans to build data centers ...
- Netherlands: Inside the data centre moratorium movement
- Germany, Brandenburg, Neuenhagen: Alphabet darf kein Rechenzentrum bei Berlin bauen now in Mittenwalde

This was all before the AI Boom took of

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.

Machine learning



Typical numbers

Model	Hardware	Hours	CO_2e (lbs)	Cloud compute (USD)
$\text{Transformer}_{\text{base}}$	P100x8	12	26	$41-$140
$\text{Transformer}_{\text{big}}$	P100x8	84	192	$289-$981
ELMo	P100x3	336	262	$433-$1472
$\text{BERT}_{\text{base}}$	V100x64	79	1438	$3751-$12,571
$\text{BERT}_{\text{base}}$	TPUv2x16	96	---	$2074-$6912
NAS	P100x8	274,120	626,155	$942,973-$3,201,722
NAS	TPUv2x1	32,623	---	$44,055-$146,848
GPT-2	TPUv3x32	168	---	$12,902-$43,008

TPU
GPU

Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

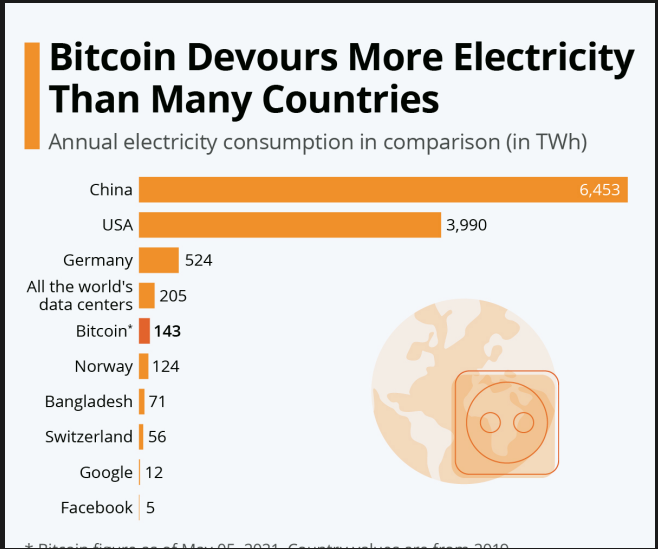
Moore's Law for Training Neural Networks
How AI will really kill us

Moore's Law

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

some 67 MW hours

Bitcoin is the worst



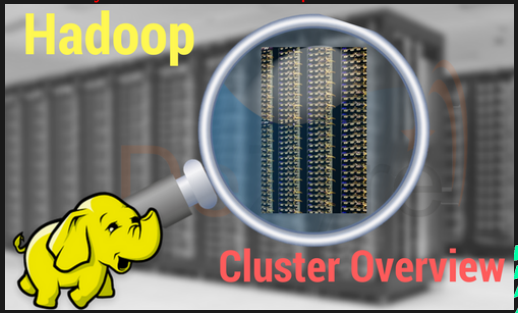
Applications

From the trenches

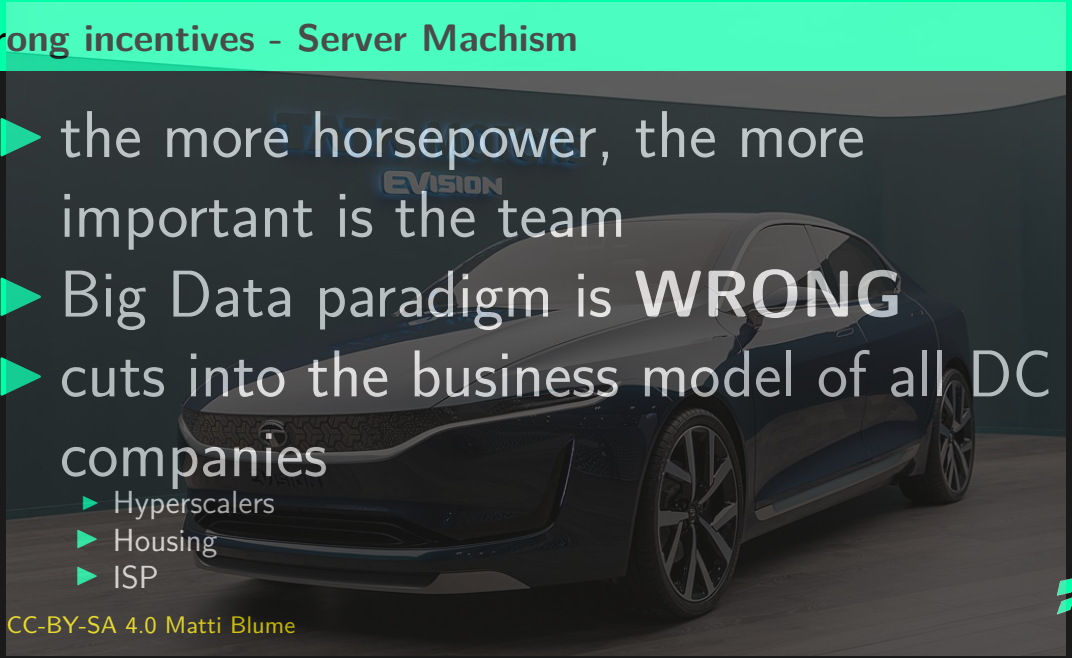
- going life in critical infrastructure
- trading software
- Open Shift
- vendor came up with 10 nodes of HBase
- tried to measure the load neglectable
- calculated the needs
  - MySQL would be oversized
  - SQLite on a Raspi would have done it

Shrink your app

- Most of the apps are hopelessly overdimensioned
- Factor of 10 is typical
- dissolved the Oracle team at a customer 10 years ago
- they moved to Hadoop

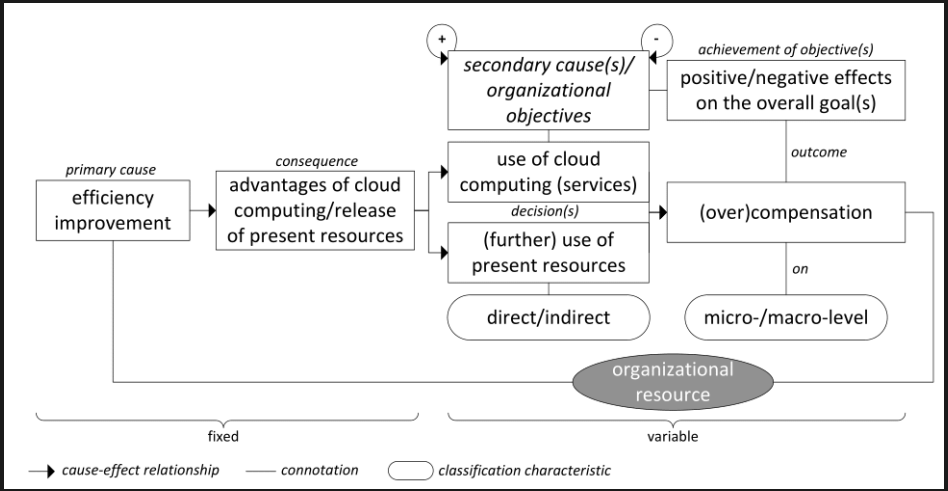


Wrong incentives - Server Machism



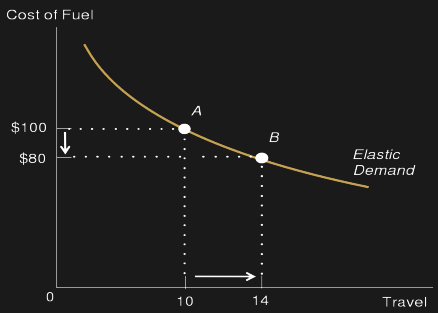
- the more horsepower, the more important is the team
- Big Data paradigm is **WRONG**
- cuts into the business model of all DC companies
  - Hyperscalers
  - Housing
  - ISP

Rebound Effect



Rebound Effects in Cloud Computing

Jevons Paradox



Jevons Paradox

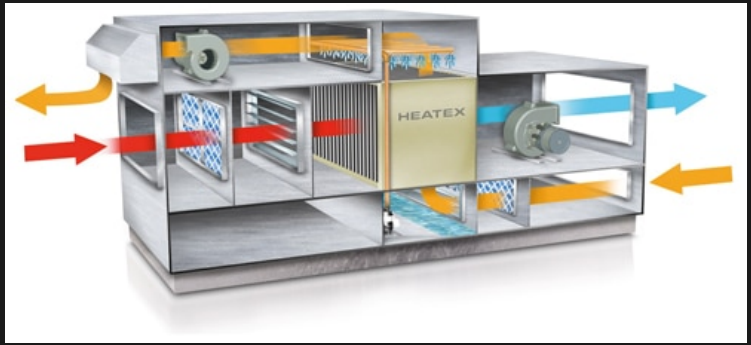
Jevons Paradox
- first described for steam engines
- example is for travelling costs
- Rebound Effects in Cloud Computing: Towards a Conceptual Framework

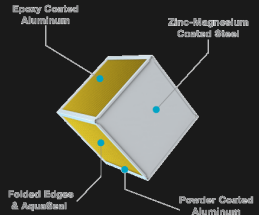
Personal observations
- provisioning times are hidden costs
- self provisioning
- cloud enabling
- virtualisation
- containers
- Kubernetes
- CI/CD pipelines

New Datatacenter Berlin Lichtenberg

- 100MW power usage
  - transformer substation
  - integration of energy management
  - cooperation with local Berlin power network
- integration into district heating
  - 50-60% use of heating energy
  - only 6% of datacenters are integrated
- $\mathrm{PUE} \ll 1.3$
- adiabatic cooling

Cooling





Source: Heatex

Local Clouds

- Sovereign Cloud Stack
  - Summit
  - Hochschule Osnabrück
    - Campus Netz
    - Datacenter in a Real Container
      Disaster management
      Distributed Usage
- Energy Transition

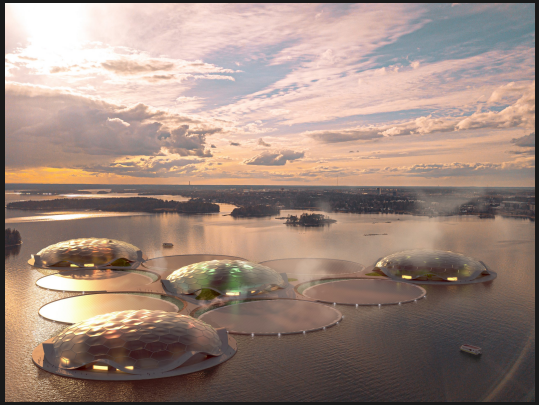
  Datacenter without losses
  - Renewable Energy
  - Cloud Computing
  - local remote heating

  by JH-Computers and OSISM

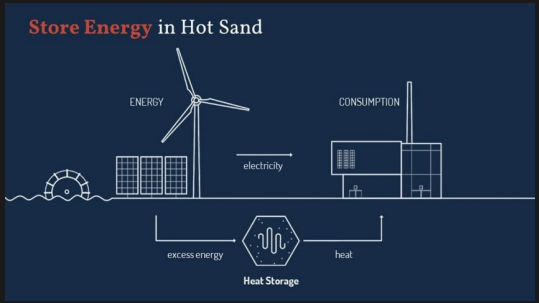
JH Computers

- upcycling heat to 85C
- scalable between 50kW and 50MW
- makes economic sense $\geq$ 250KW
  - 2GWh/year heat production
  - 400 houses
- customers want at least 5MW datacenters
  - 40 GWh/year
  - 8000 houses
  - industrial usage
- real estate companies are interested

Energy or Heat Storage in Finnland

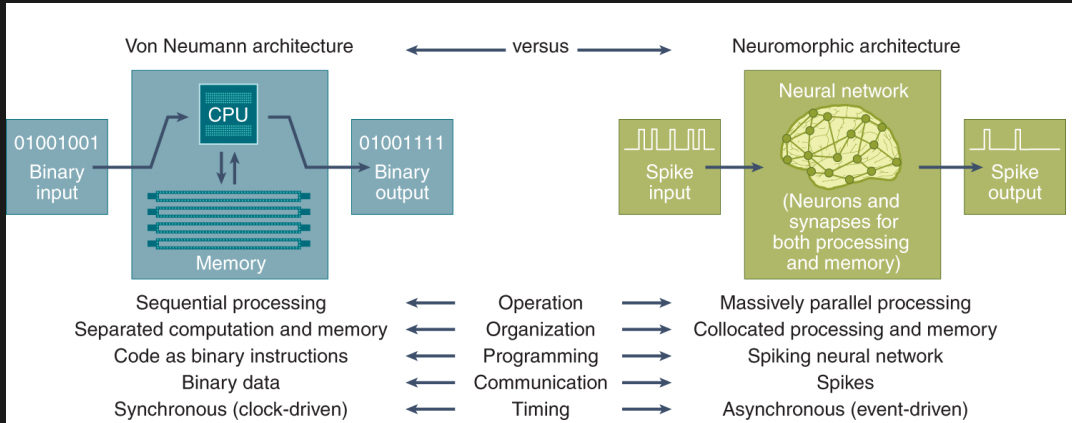


Hot Heart Project



Finnish “sand battery” offers solution for renewable energy storage

Neuromorphic Computing – Nature



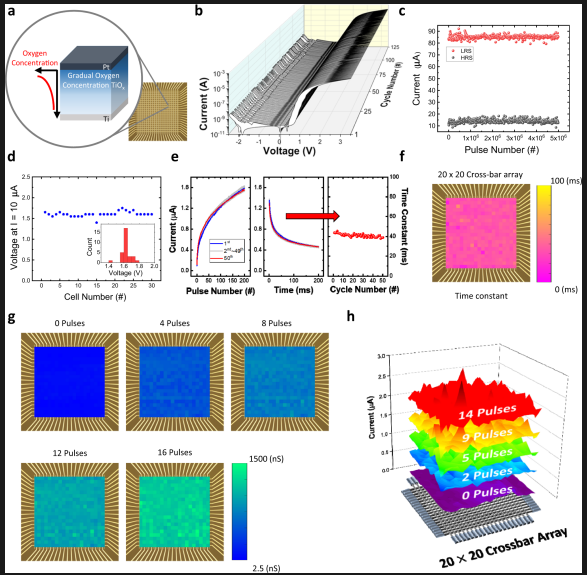
Neuromorphic Computing – Intel

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 capability1
- Up to 60x more inter-chip bandwidth2
- Up to 1 million neurons with 15x greater resource density3
- 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 – Nature about Memristor

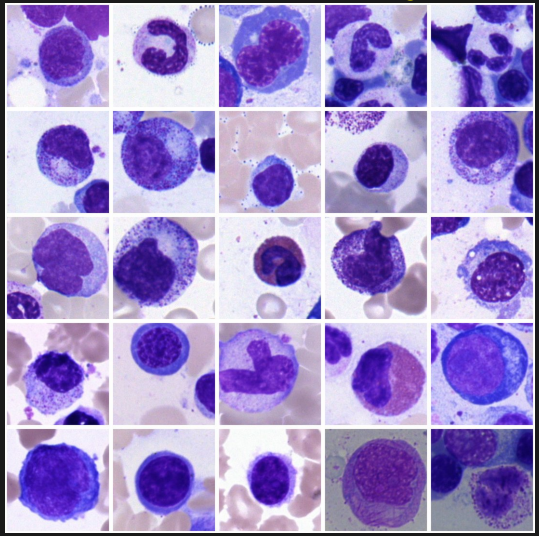


DeepSouth

Western Sydney University to get DeepSouth neuromorphic supercomputer in 2024

Where is the positive? We want AI!

Helmholtz München Leucemia Recognition



Alpha Fold Protein Folding

AI to get nuclear reactors approved

- Microsoft is training an AI to help get nuclear reactors approved
- Getting new nuclear reactors approved by regulators is an expensive, complex process.
- We're really excited about the game-changing potential for AI in this space. MICHELLE PATRON
- Is advanced nuclear in trouble? What's next after NuScale cancellation

Conclusion

- AI will use a lot of resources
  - without care we will see a factor of 5-10 increase in energy consumption
  - this is a danger to the local climate
- energy hunger cannot be satisfied
  - competion with other industries
  - nuclear power plants
    - really?
    - we are running out of Uranium btw.
    - not under democratic control
- GPUS are a naive way of implementing perceptrons
- research on alternatives?
- there is a problem with bias
  - careful selection of sources
  - expensive human control
- charlatanry
- massive financial interest

Question? Remarks?

Some Answers

Mail: ai@thomasfricke.de

LinkedIn

Mastodon: @thomasfricke@chaos.social