

L5: HPC for AI applications & Environmental impact of computation

P. de Oliveira Castro, M. Jam
September 9, 2025

Master Calcul Haute Performance et Simulation - GLHPC | UVSQ

1. HPC for AI & Environmental impact of computation
2. Introduction to AI applications
3. Environmental impact of computation
4. Energy consumption of HPC
5. AI energy and computation costs
6. More frugal computing?

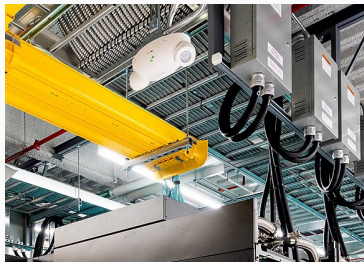
HPC for AI & Environmental impact of computation

Introduction to AI applications

Environmental impact of computation

Introduction

- Major ecological crisis: French roadmap targets carbon neutrality in 2050 (Stratégie Nationale Bas Carbone).
- Requires a **40% energy consumption reduction**.
- HPC part of the solution: modeling and improving complex systems
- HPC part of the problem: Frontier system at ORNL
 - More than 10^{18} floating point operations per second
 - Consumes 21MW: the energy of a small town (16 000 french houses)



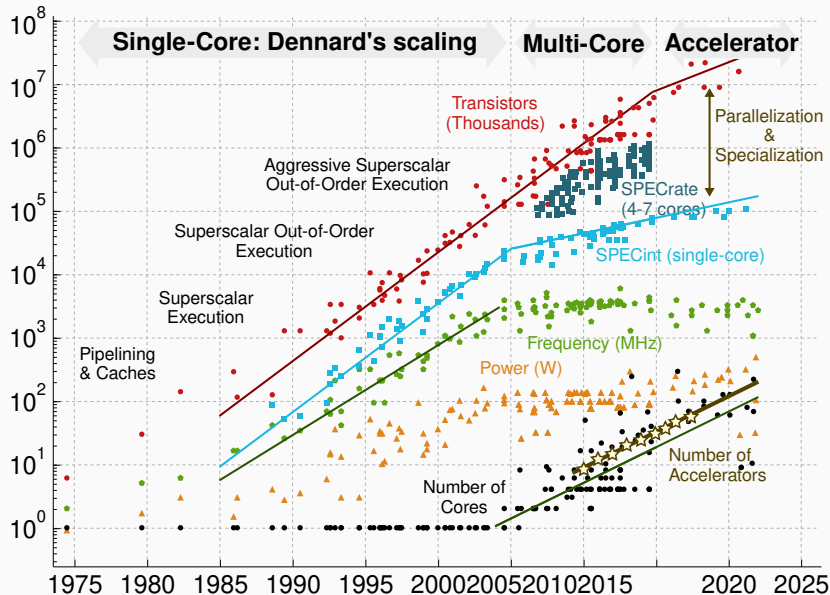
- The ICT sector consumes $\approx 5\%$ of the energy worldwide
- It accounts for **1.8% - 2.8%** of emitted GHG [Freitag, 2021]:
 - Accounts for embodied emissions.
 - Shadow energy during the whole life-cycle: mining, fabrication, transportation, recycling.
- GHG emissions are only one of the sustainability issues
 - rare-earth mining and waste disposal (eg. Agbogbloshie).
 - human-right abuses, health issues, pollution.
- This presentation focus on energy consumption of HPC

What about renewable energies?

- Low-carbon electricity is a **limited** resource
- Decarbonation → huge increase in electricity demand
 - Heating, Transportation, Industry
 - Computing will compete for low-carbon electricity.

Energy consumption of HPC

Evolution of processing units [Batten, 2023]



$$\text{CMOS Power } P = \underbrace{1/2 \cdot C \cdot V^2 \cdot f}_{P_{\text{dynamic}}} + \underbrace{V \cdot I_{\text{leak}}}_{P_{\text{static}}}$$

For each generation, transistors dimensions reduced by 30%,

- Voltage and capacitance reduced by 30%
- Frequency increases: $\times 1.4 \approx 1/0.7$
- Surface halved: $0.5 \approx 0.7 \times 0.7$
- Power halved: $\Delta P = 0.7 \times 0.7^2 \times 1/0.7 \approx 0.5$

Power per surface unit remains constant but manufacturers double number of transistors and frequency increases:

- Power efficiency doubles every 1.57 years
- Total power increases

- At current scale, leak currents start increasing (P_{static} \nearrow).
Power wall slows Dennard's scaling.
- Computing demand \rightarrow **parallelism** and **specialization**.
- Number of cores increases exponentially since 2005.
- Power efficiency still improving:
 - selectively turning-off inactive transistors;
 - architecture design optimizations;
 - software optimizations.

- For domain specific applications, such as AI, specialized accelerators are used
 - Memory and compute units tuned for a specific problem (matrix multiplication) ;
 - Faster and better power efficiency: GPU, TPU, FPGA, ASIC.

Analysis of TOP-100 HPC systems

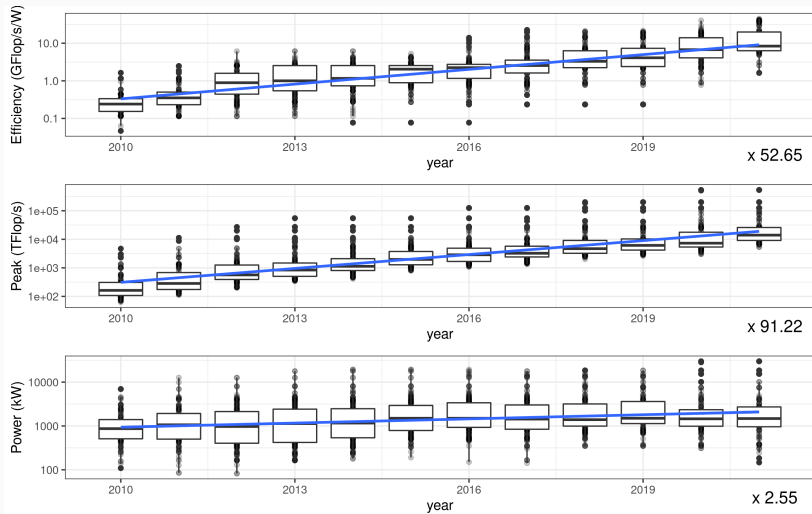


Figure 3: image

- In 1865, Jevons shows that steam engine improvements translate into increased coal consumption.
- In HPC, efficiency gains contribute to the rising computation demand.
 1. net increase of the total power consumption.
- Rebound effects for data-centers [Masanet, 2020]
 1. 6% increase in energy consumption from 2010 to 2018 (255 % increase in nodes).
- **Indirect rebound effects:** computation advances can contribute to the acceleration of other fields.

AI energy and computation costs

- 2012: **AI renaissance** brought by increased data availability and computation resources
 - breakthroughs in multiple domains
 - many innovations: algorithms, specialized processors, optimizations
- Most systems use **neural networks**:
 - Training (stochastic gradient descent + backpropagation)
 - Inference (forward pass)
- For both, the **bottleneck is matrix multiplication**

Training cost doubles every 3.4 months [OpenAI, 2020]

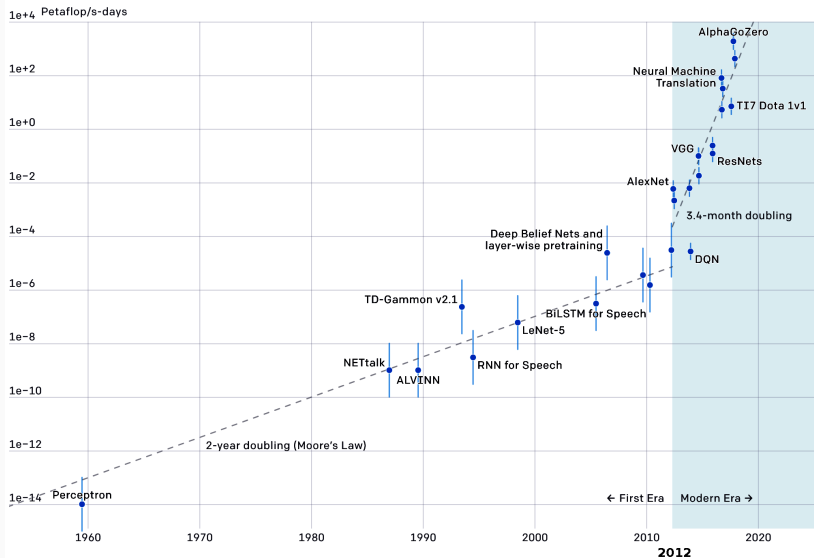
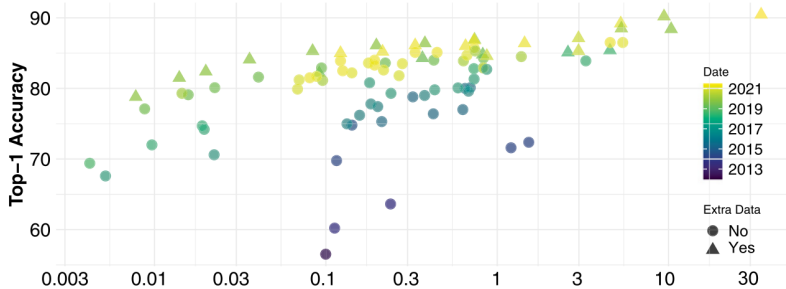
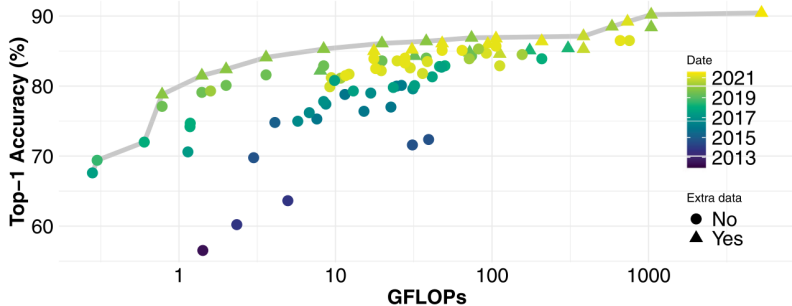


Figure 4: image

Should we study training or inference?

- **Training:** huge cost but done once
 - GPT3, 175 billion parameters, ≈ 314 ZettaFLOP
 - GPT4, 1.7 trillion parameters
- **Inference:** millions of users and requests
 - 80-90% cost of a deployed AI system is spend on inference [NVIDIA, 2019]

Inference cost - Diminishing returns for computer vision



More frugal computing?

Smaller precision / Smaller models for AI

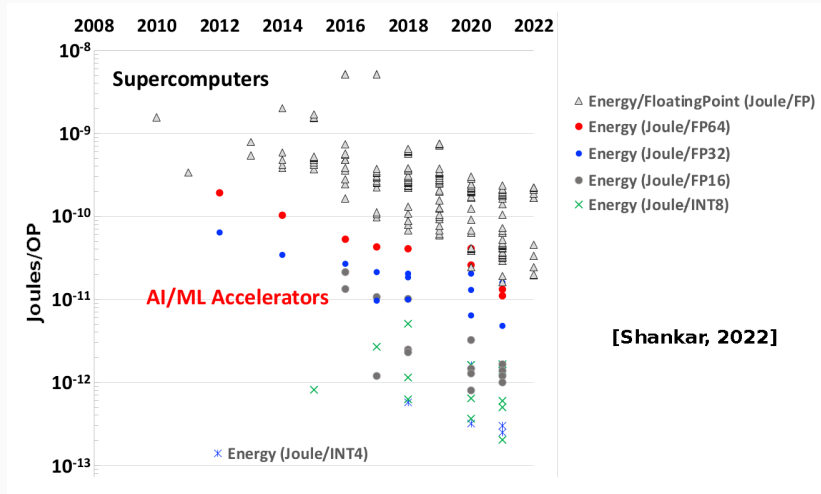
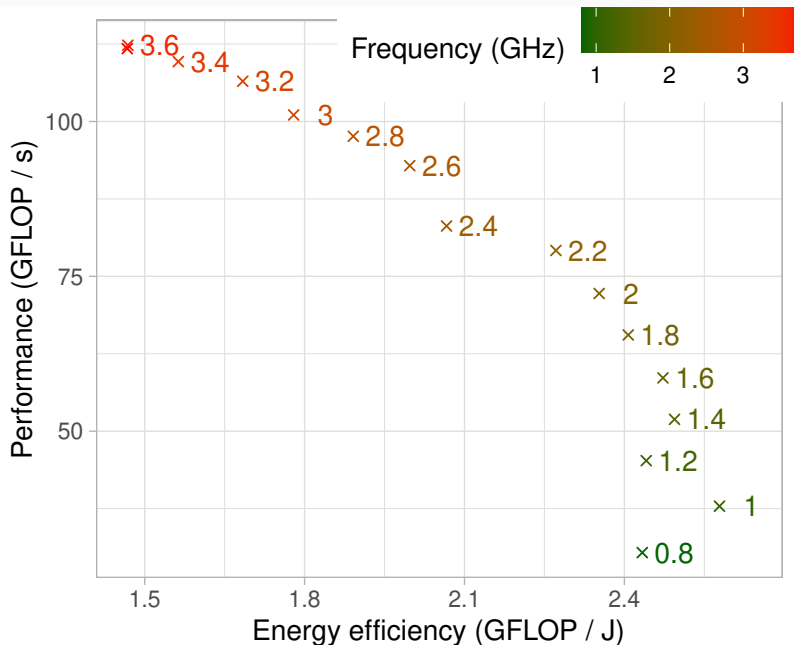


Figure 5: image

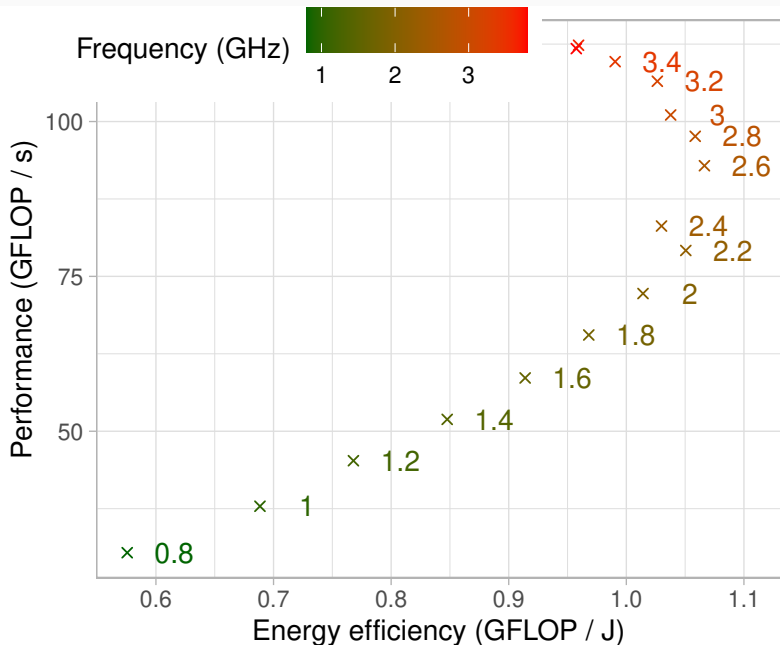
LLM success of smaller models (Llama, Chinchilla) fine-tuned for specific tasks with LoRA.

- Inference cost grows with model complexity
- Simpler models are often more interpretable
 - Traditional science also prefers simpler models
- DNN not necessary for all tasks

DVFS study of LU decomposition



When accounting for the whole system



Need for an interdisciplinary discussion

- AI / HPC can contribute towards sustainability (eg. acceleration of weather forecast models) ... **but its energy cost must be reduced**
- **Efficiency:**
 - Improve hardware and software
 - Use smaller models / smaller precision
- ... subject to rebound effects
- **Frugality in computing:**
 - Balance computation cost vs. outcomes for each task
 - Choose the right sized model
 - Assess the environmental impact

Treatment of febrile children illnesses in dispensaries.

- IMCI: Paper-based decision tree WHO
- e-POCT CART tree tailored to real data on a standalone tablet
 - Final CART tree easy to interpret and manually checked
 - Randomized-trial → better clinical outcomes and antibiotic prescription reduction
- Sophisticated AI that continuously collects patient data and adapts the algorithm ?
 - Increase in hardware and computation costs.
 - Loss in explainability and verification of the algorithm.

References - Environmental impact of computation

- Jones, Nicola (2018) 'How to stop data centres from gobbling up the world's electricity'. *Nature*, 561(7722), pp. 163–167.
- Freitag, Charlotte, Berners-Lee, Mike, Widdicks, Kelly, Knowles, Bran, et al. (2021) 'The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations'. *Patterns*, 2(9), p. 100340.
<https://www.sciencedirect.com/science/article/pii/S2666389921001884>
- Masanet, Eric, Shehabi, Arman, Lei, Nuoa, Smith, Sarah and Koomey, Jonathan (2020) 'Recalibrating global data center energy-use estimates'. *Science*, 367(6481), pp. 984–986.
- Schwartz, Roy, Dodge, Jesse, Smith, Noah A. and Etzioni, Oren (2019) 'Green AI'. arXiv:1907.10597
- Amodei, Dario, Hernandez, Danny, Sastry, Girish, Clark, Jack, et al. (2018) 'AI and compute. OpenAI'.
<https://openai.com/blog/ai-and-compute/>
- D'Acremont presentation: https://youtu.be/oKcy_cY0QOw