HPC & SUSTAINABILITY: A NEW PARADIGM

Transforming Computational Power into an Environmental Asset

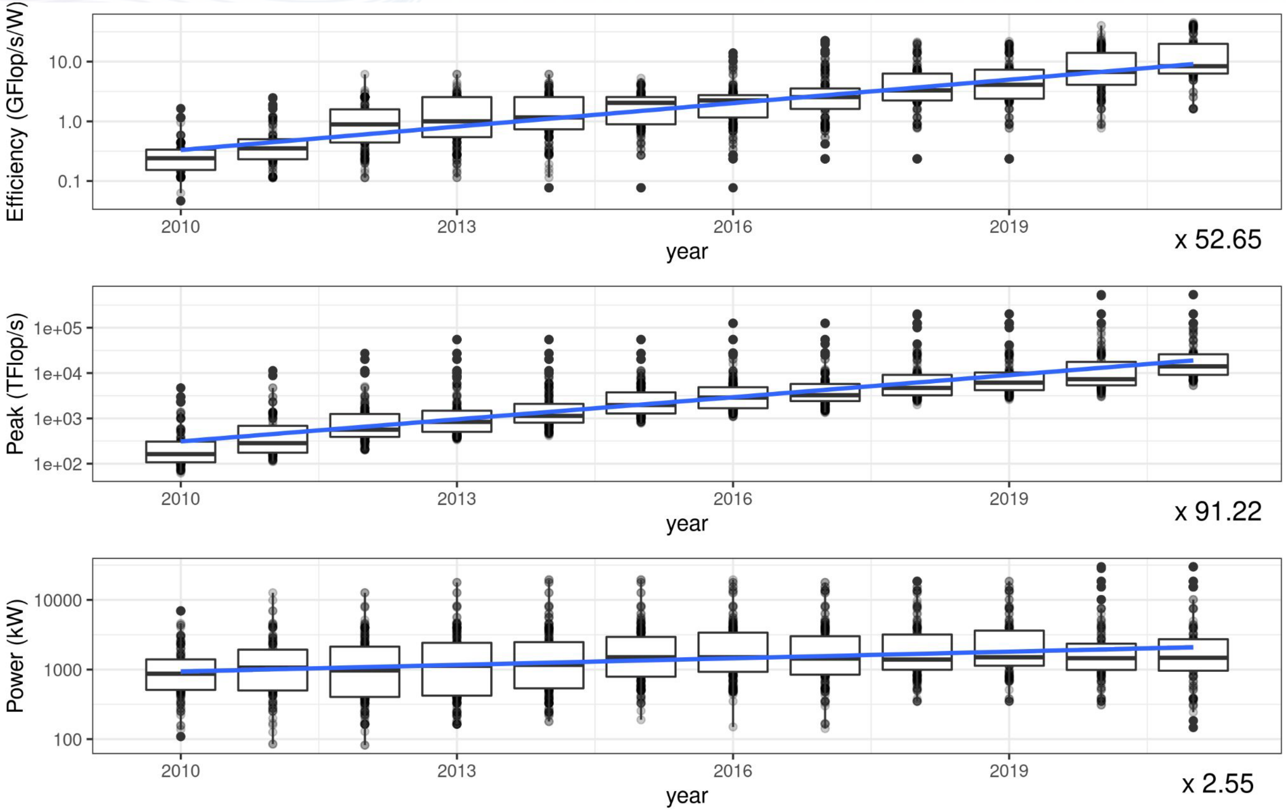
Guillaume Trainar

May 2025

Today's Agenda

- Metrics
- The Hidden Costs of HPC
- Actionable Levers by stakeholders
- Future Directions & Key Challenges

TOP500 Supercomputers - Efficiency and Power Consumption



source: https://theses.hal.science/tel-03831483/ p135

The Hidden Costs of HPC – Beyond Energy Consumption

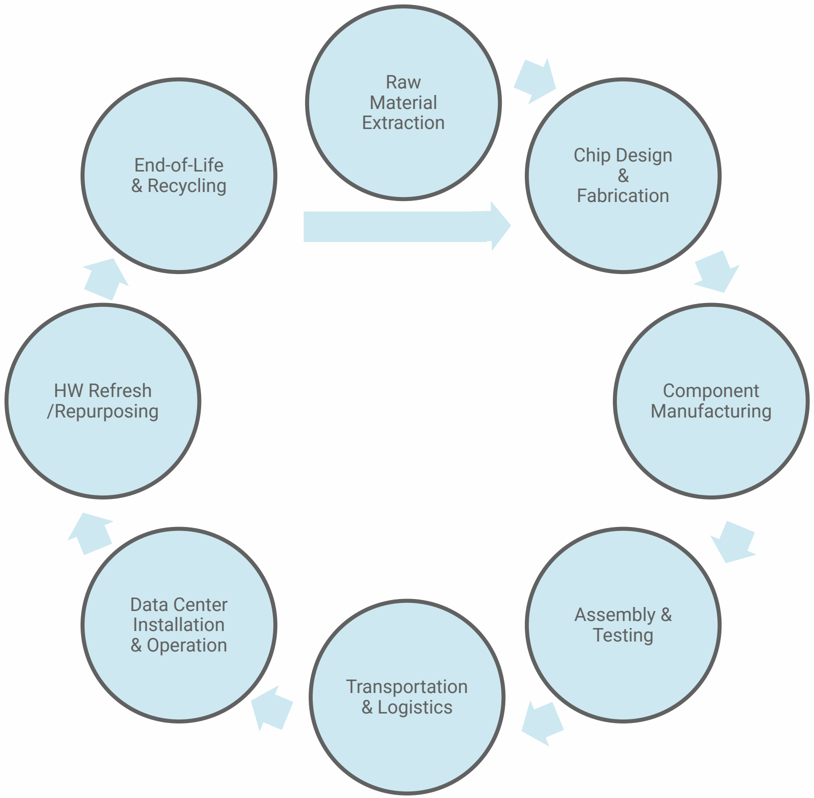
It's More Than Just Compute Watts.

TSMC now produces more than 30 million 8-inch (equivalent) wafers a year, and consumes more than 25 GWh per year.

Shipping one server via air freight from Asia to North America can generate over 1 ton of CO2e. Sea freight for the same distance is roughly 0.2-0.5 tons CO2e.

Refurbishing a server can require 10-30 kWh.

Recycling one server can require approximately 50-100 kWh.



The E-Waste Challenge – A Growing Problem

What do Hyperscalers do to mitigate the impact?

Rapid hardware obsolescence contributes to growing hazardous e-waste, posing risks if not properly recycled.

- Obsolete equipment contains toxic substances like lead, mercury, and cadmium that contaminate the environment.
- Low recycling rates mean a large portion of e-waste ends up in landfills or is exported to developing countries with lax environmental standards.

Effective e-waste management systems are crucial for recovering valuable materials and minimizing pollution.

Cloud HPC: An Opportunity for Optimization

The Hyperscalers have bespoke programs to:

- Design Circular data centers to repurpose and recycle servers and hardware from their data centers
- Partner with other tech companies to reclaim valuable materials from unused electronics

Resource Pooling:

- Share computing resources among multiple users or applications, resource pooling reduces the need for excess hardware.
- Optimize the utilization of existing resources, ensuring that no server operates underutilized

Virtual Machines:

- Allows multiple virtual servers to run on a single physical machine. This reduces the number of physical servers required
- Enable dynamic allocation of resources, ensuring that computing power is used efficiently and only when needed

Cloud HPC – Tradeoffs and Considerations

Data Transfer Energy Costs

Data transfer energy costs are optimized by cloud providers pooling resources and using efficient data centers. However, transferring data between regions or to external networks increases energy consumption due to network infrastructure demands.

Location Considerations

Location affects latency, energy efficiency, and costs in cloud data centers. Providers often locate data centers in regions with renewable energy and cooler climates to reduce environmental impact.

HPC Driving Environmental Solutions - Weather Modeling

Weather Research and Forecasting (WRF) Model

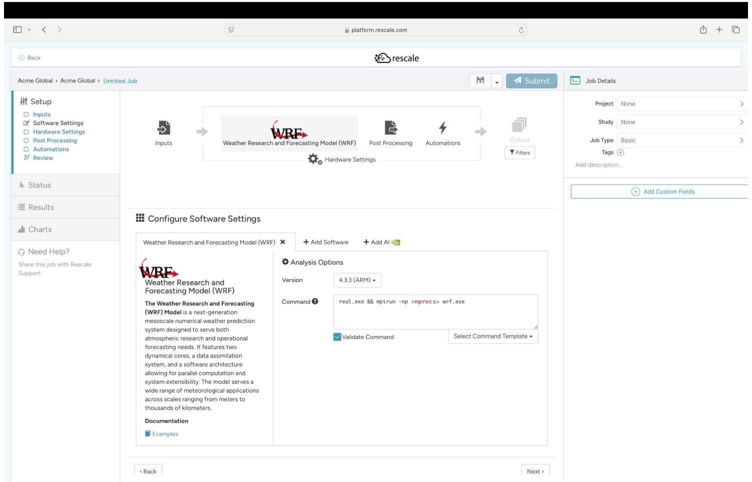
The Weather Research and Forecasting (WRF) Model is a next-generation, GPU-accelerated, mesoscale numerical weather prediction system designed to serve both atmospheric research and operational forecasting needs.

Accelerated WRF with GPUs = Faster, Smarter Forecasts

- Simulate high-fidelity weather to capture tornadoes, coastal wind shifts, and urban microclimates without downsampling
- Reduce forecast time from hours to minutes with GPU-accelerated performance

Why It Matters

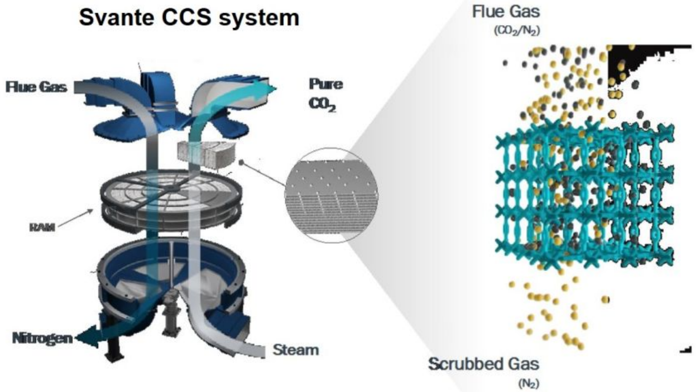
- Emergency response: Earlier, more accurate alerts
- Insurance and risk: Better climate exposure modeling
- Operations: Site-specific, real-time weather insights



The Innovation Lens

Simulation and HPC's Role in Materials Science & Carbon Capture

- **Accelerated Materials Discovery:** HPC enables rapid simulation of atomic and molecular interactions—crucial for designing next-gen materials for batteries and solar cells.
- **Battery Innovation:** Simulation is used for lithium-ion diffusion, electrolyte behavior, and solid-electrolyte interphase (SEI) formation to enhance performance and safety.
- **Solar Energy Advancements:** Models photovoltaic materials like perovskites, helping optimize efficiency, durability, and fabrication methods.
- **Carbon Capture Simulation:** Facilitates detailed modeling of sorbent materials, CO_2 adsorption mechanisms, and process optimization under varied conditions.



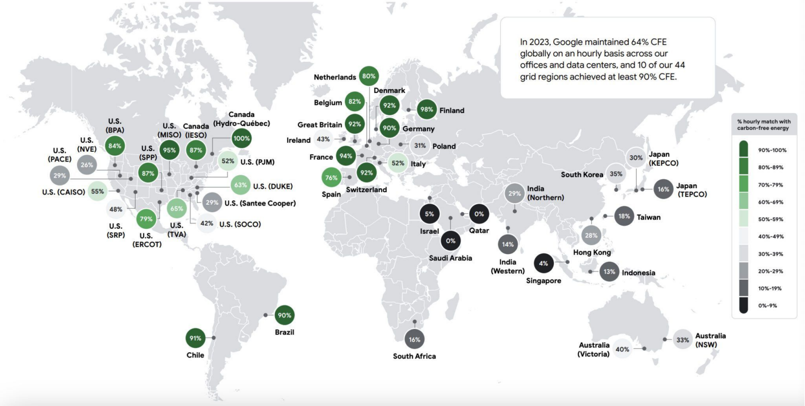
Savante is disrupting the world's CO2 marketplace

Svante

Powered By Clean Energy

The Need for Data Management for Collaboration on Clean Energy

- **Global collaboration:** A shared source of truth helps to connect engineering, R&D, and supply chain teams for faster, aligned decision-making internally and externally.
- **Traceability & compliance:** Captures the full simulation lifecycle to support regulatory needs in different regions and ensure design intent is preserved.
- **Faster time-to-market:** Streamlines workflows and eliminates duplicate effort to meet aggressive clean energy innovation timelines.
- **Sustainability alignment:** Ensures every simulation supports low-carbon goals (core-types, power consumption, usage).



[Google leverages Kairos Power to target carbon-free data centers]

Kairos Power

Future Directions & Key Challenges

energy-efficient hardware

- ARM: Assuming a 40% energy efficiency advantage for ARM (conservative estimate), a 25% shift in workload could yield approximately 10% overall energy savings.
- Quantum: 1,000x more energy efficient - challenge very limited uses cases to date

algorithmic optimization/AI-Physics,

- Studies have shown that rewriting computationally intensive Python code in C can reduce energy consumption by 20-50% or more
- AI/Physics-informed models can accelerate simulations and improve accuracy, reducing the need for computationally expensive iterations. Savings potential ranges from 10-40% depending on the application.

circular economy - repurpose aging HW

- Extending the Useful Life of 15% of Aging Servers will reduce the need to manufacture new servers and avoid the significant embodied energy and associated carbon footprint

AI-Driven Impeller Optimization

CFD Model: Design and performance optimization study with **7** parameters of a periodic, symmetric, compressible flow simulation (~100,000 tet mesh) using ANSYS CFX and OpenCAD.

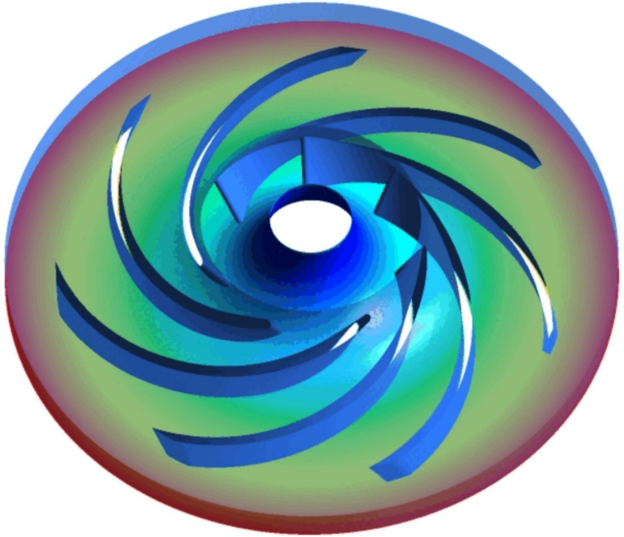
To accelerate the design and optimization process, and to scale it AI-driven surrogate models are used to approximate full-fidelity simulations.

For example, comparing two techniques the ROI is evident at scale:

- Direct Design and Optimization (Full Fidelity Simulation)
- AI-Driven Design and Optimization (NN Surrogate Models)

x5 Speed Up, 5x Cost Reduction, -80% Energy to build a 10,000 DOE suite

Approach	Compute Time	Hardware	Cost	Accuracy
Full Fidelity	~5 weeks (5m @ 10,000)	4 CPU Cores Intel Xeon Platinum P-8124 Skylake @ 3.0 GHz TDP=240W	10000*4*5/60. 3333 core hours. $.102/core-hr. **$340** for the 10k runs. **E: 44.45kW.h**	100%
AI-Driven	~7 days & 3 hrs (5m @ 2,000) build data set + ~3m AI training + 10,000 x 1s Inference speed	CPU Cores for building simulation data set V100 GPU for training TDP=300W	$4.158 / hour. $0.21 for training. Inference is negligible for 2000 runs **$68** **E: 8.9 + 0.9 = 9.7kW.h**	98.57% (Mean Absolute Percentage Error: 1.43%)



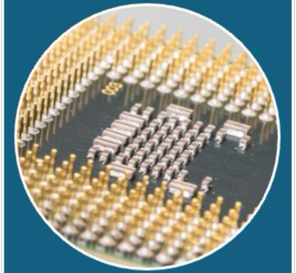
The neural network was trained using **PyTorch** with 3 fully connected layers (10 neurons each), an Adam optimizer, and MSE loss, running for 31,000 epochs in ~3 minutes on a single NVIDIA V100 GPU.

Conclusion

Moving to Cloud HPC for better sustainability

Resources Pooling

Advanced Data Center Efficiency & Renewable Energy Adoption



Optimized Hardware & Software Optimization

Husher innovation for more sustainable solutions

Superior Performance to achieve better sustainability