Computing within environmental limits – Energy consumption of HPC

Pablo de Oliveira (pablo.oliveira@uvsq.fr)

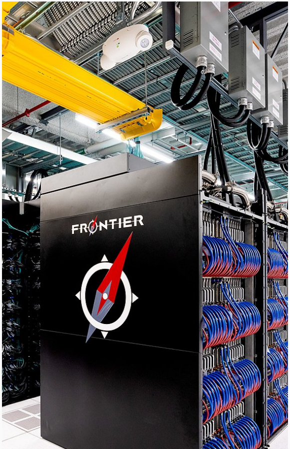
2024/02/01, LIX sustainable computing seminar

Université de Versailles, Université Paris-Saclay, Li-PaRAD

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)



Environmental impact of computation

- The ICT sector consumes $\approx$ 5% of the energy wordwide
- 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 emmissions 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 ressource
- Decarbonation $\rightarrow$ huge increase in electricity demand
  - Heating, Transportation, Industry
  - Computing will compete for low-carbon electricity.

Outline

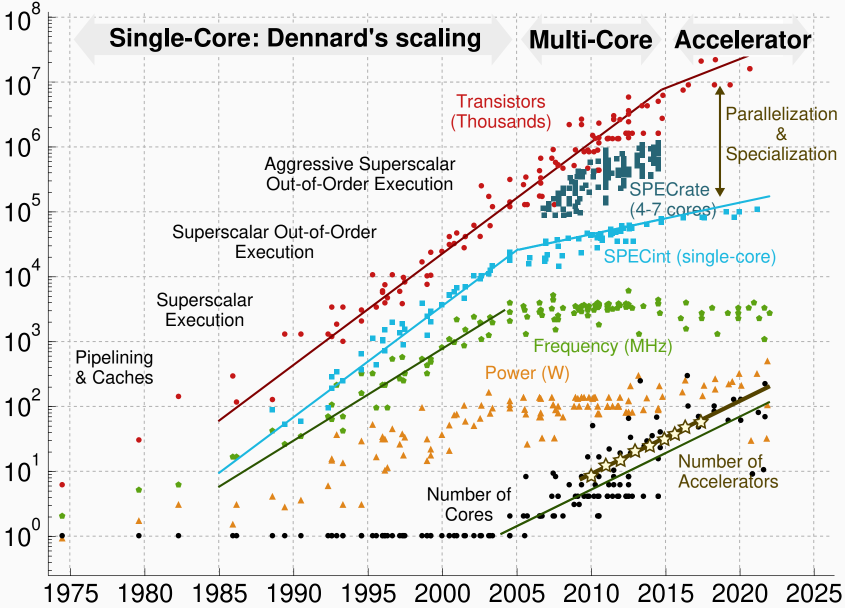
Energy consumption of HPC

AI energy and computation costs

More frugal computing?

Energy consumption of HPC

Evolution of processing units [Batten, 2023]



C. Batten, M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, K. Rupp & [Y. Shao, IEEE Micro'15] & [C. Leiserson, Science'20]

Dennard's scaling 1970-2005

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

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

Multicore 2005-2020

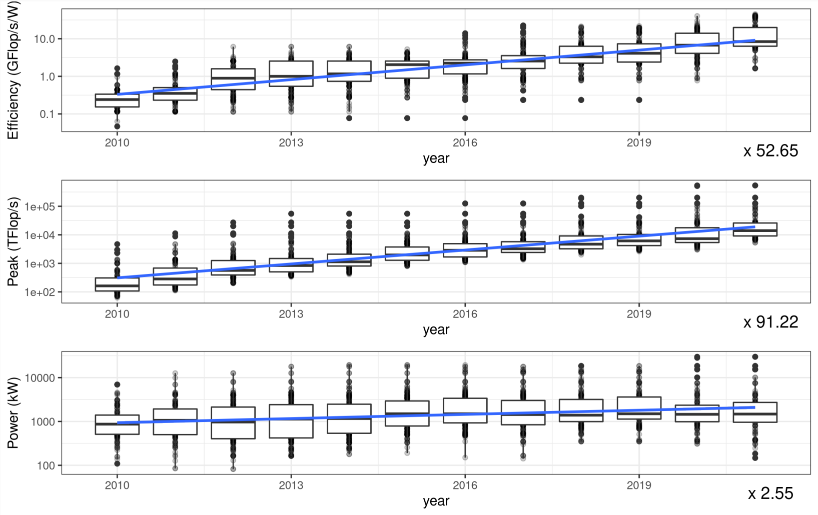
- At current scale, leak currents start increasing ($P_{\text{static}} \nearrow$).
  Power wall slows Dennard's scaling.
- Computing demand $\rightarrow$ parallelism and specialization.
- Number of cores increases exponentially since 2005.

- Power efficiency still improving:
  - selectively turning-off inactive transistors;
  - architecture design optimizations;
  - software optimizations.

AI Accelerators 2020-2024

- 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



Efficiency and Peak computation exponential increase.

Rebound effects

- In 1865, Jevons shows that steam engine improvements translate into increased coal consumption.

- In HPC, efficiency gains contribute to the rising computation demand.
  - $\rightarrow$ net increase of the total power consumption.
- Rebound effects for data-centers [Masanet, 2020]
  - $\rightarrow$ 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.

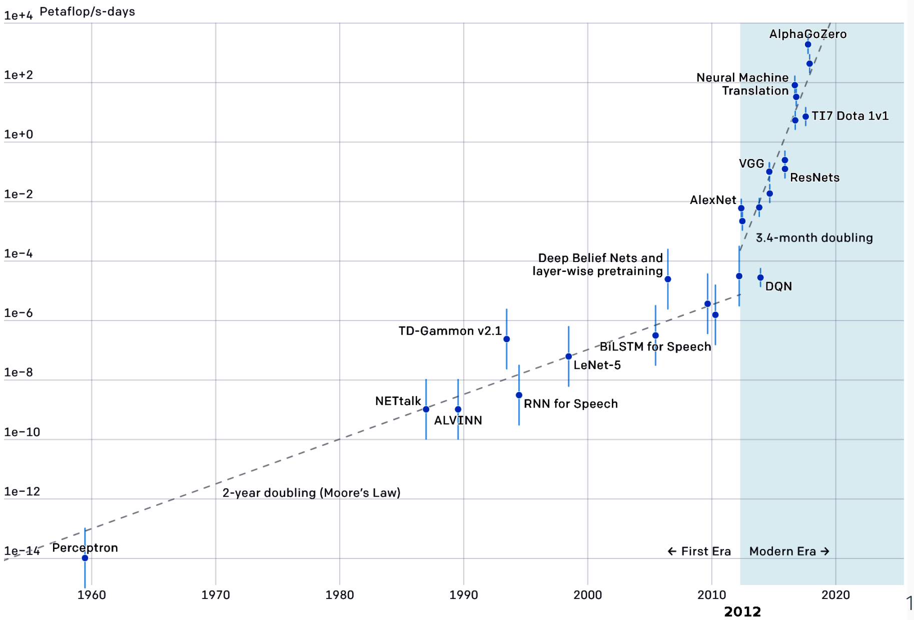
AI energy and computation costs

Artificial Intelligence

- 2012: AI renaissance brought by increased data availability and computation ressources
    - 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

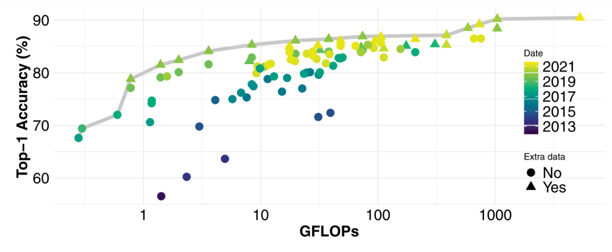
Training cost doubles every 3.4 months [OpenAI, 2020]

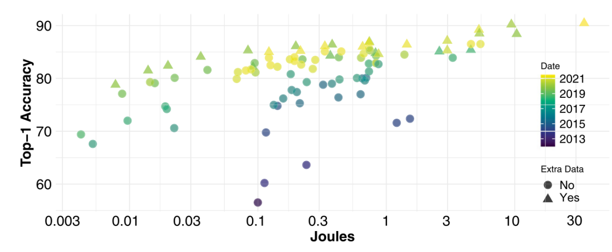


Should we study training or inference?

- Training: huge cost but done once
  - GPT3, 175 billion parameters, $\approx$ 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

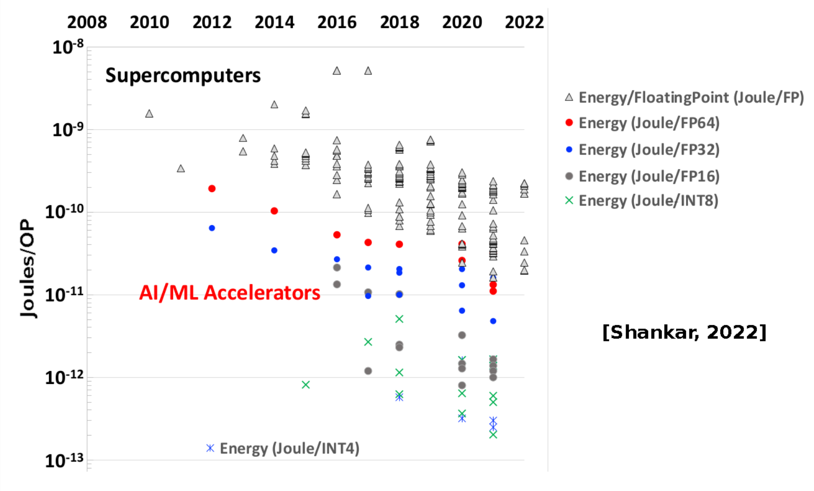




Exponential increase in compute for linear accuracy gain
[Desislavov, 2023 / Schwartz, 2019]

More frugal computing?

Smaller precision / Smaller models for AI



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

Tradeoff: Model complexity - Cost - Explainability

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

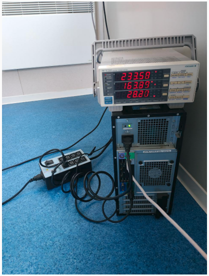
Two case-studies in HPC

- Computing slower: DVFS for LU decomposition in KNM architectures

- Computing less precisely: mixed precision in YALES2 solver

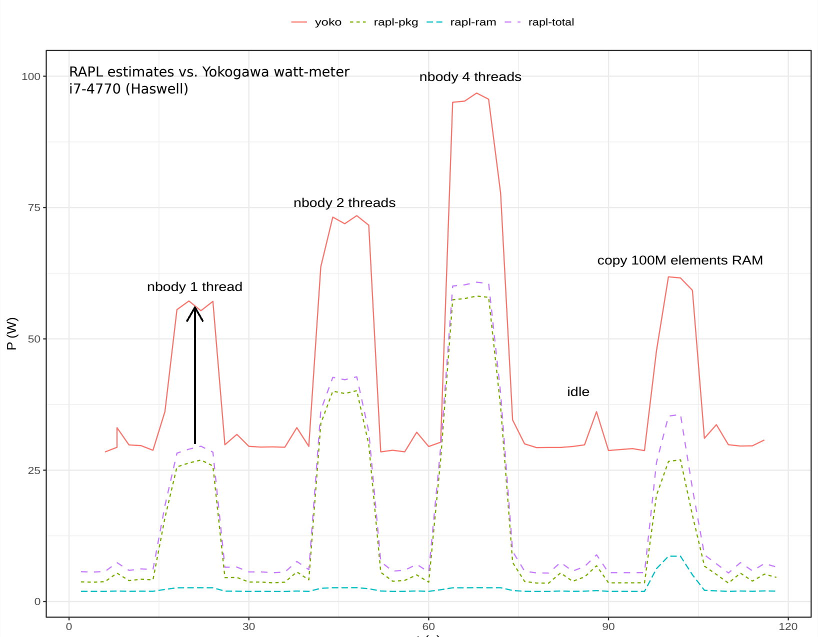
Measuring the energy?

- Wall watt-meters
  - precisely measure the total consumption
  - slow sampling resolution ($\approx$ 1s)
  - hard to use within a super-computer

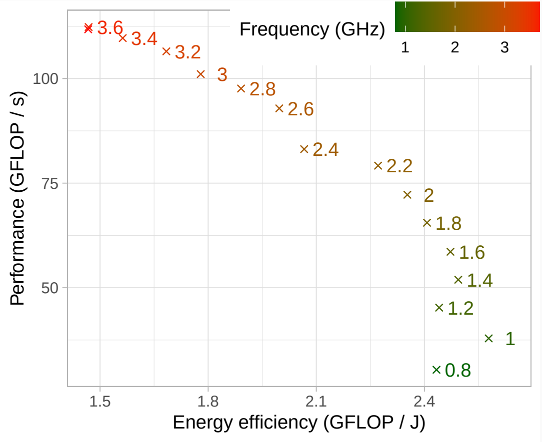


- Manufacturers energy-counters (rapl, nvml, ...)
  - easy to access and high sampling rate
  - do not capture the whole system consumption
  - use power estimate models

RAPL vs. Yokogawa watt-meter



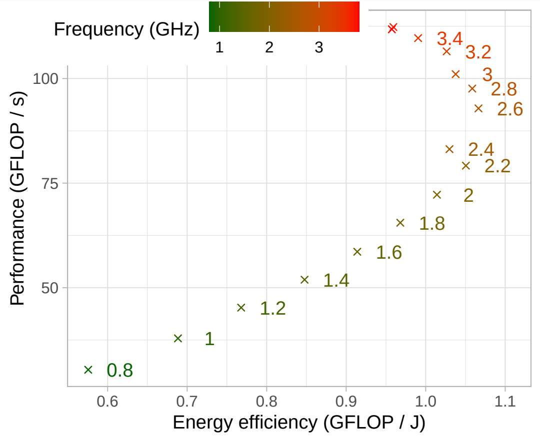
DVFS study of LU decomposition



- Knights Mill 72 cores
- Intel MKL dgetrf
- $n \in [1000, 3000]$
- RAPL estimation

Save energy by computing slower: 1GHz

When accounting for the whole system

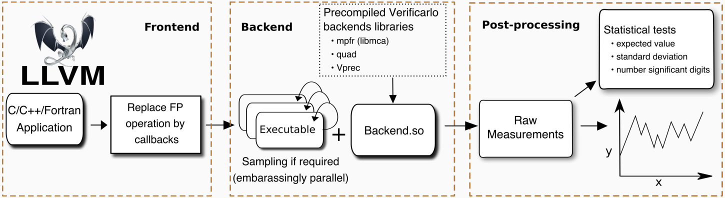


- Model : RAPL + 40W

- Optimal 2.6 GHz : compute faster and turn off machine
- Saves idle power (race to idle)

`github.com/verificarlo/verificarlo`

- Based on the LLVM compiler
- Active open source project with 15 contributors
- Backends: debugging (MCA, Cancellation) + mixed-precision (Vprec)
- MCA overhead from $\times 6$ (binary32) to $\times 160$ (binary64).



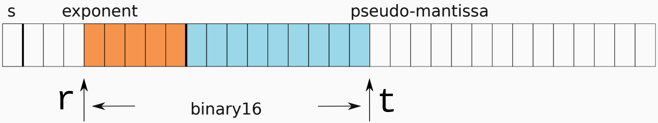
Verificarlo: Checking Floating Point Accuracy through Monte Carlo Arithmetic.

Denis, de Oliveira Castro, Petit. IEEE Symp. on Computer Arithmetic 2016

VPREC for mixed precision

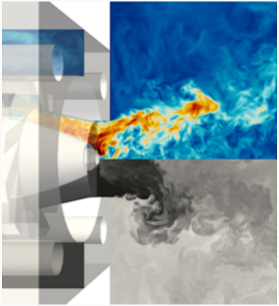
- Estimate numerical effect of bfloat16, tensorflow32, fp24 on standard IEEE-754 hardware (before paying the porting cost)

- VPREC emulates any range and precision fitting in original type
  - Uses native types for storage and intermediate computations
  - Handle overflows, underflows, denormals, NaN, $\pm\infty$
  - Rounding to nearest (faithful)
  - Fast: $\times$ 2.6 to $\times$ 16.8 overhead



YALES2 application

- Computational Fluid Dynamics solver from Coria-CNRS

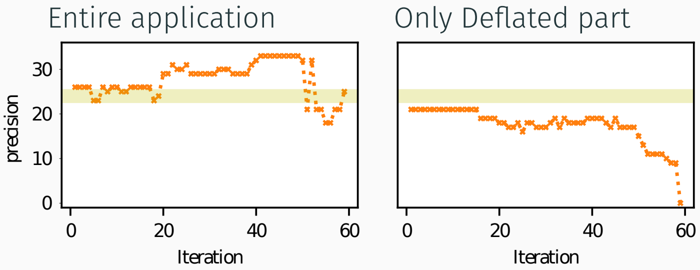


- Deflated Preconditioned Conjugate Gradient
- CG iterations alternate between a:
  - Deflated coarse grid
  - Fine grid

VPREC: Find minimal precision over iterations that preserves convergence (dichotomic exploration)

Automatic exploration of reduced floating-point representations in iterative methods. Chatelain, Petit, de Oliveira Castro, Lartigue, Defour. Euro-Par 2019

Mixed-precision on Yales2



Minimal precision that preserves convergence.

Energy	16% gain on the deflated part
Communication	28% gain on communication volume
Time	10% speedup on CRIANN cluster (560 nodes)

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

Exemple: e-health solution in Tanzania [d'Acremont, 2021]

Treatment of febrile children illnesess 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 $\rightarrow$ 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.

D'Acremont presentation: `https://youtu.be/oKcy_cY0QOw`