

High Performance Computing and Data Infrastructure

Energy problems for HPC and DATA infrastructure



DATA SCIENCE &
ARTIFICIAL INTELLIGENCE



SCIENTIFIC &
DATA-INTENSIVE COMPUTING

2024-2025 @ Università di Trieste

Agenda

- Energy problem in HPC
 - Green500
 - Available technology on modern CPUs
 - What can we measure
 - How can we measure it ?
-
- AIM:
 - Give you the feeling how much is important the energy problem in the HPC arena right now

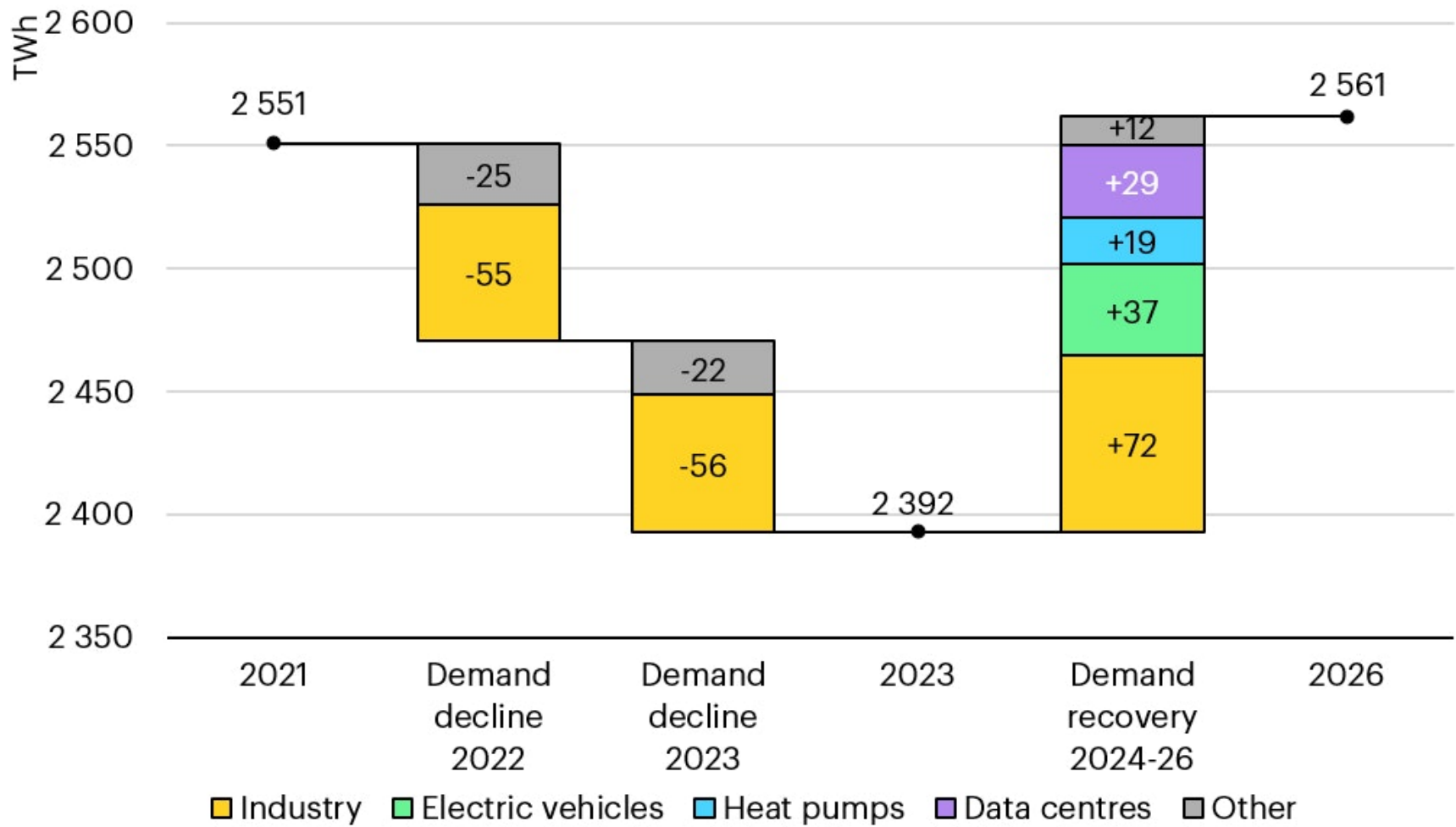
Energy requirements from ICT

- **Electricity consumption from data centres, artificial intelligence (AI) and the cryptocurrency sector could double by 2026.**
 - Data centres are significant drivers of growth in electricity demand in many regions. After globally consuming an estimated 460 terawatt-hours (TWh) in 2022, data centres' total electricity consumption could reach more than 1 000 TWh in 2026.
 - This demand is roughly equivalent to the electricity consumption of Japan.
 - Updated regulations and technological improvements, including on efficiency, will be crucial to moderate the surge in energy consumption from data centres.

Energy problem..

- “A typical supercomputer consumes anywhere between 1 to 10 megawatts of power on average, which is equal to the electricity needs of almost 10,000 homes”
- the electricity bill paid by the RIKEN institute in 2020 for their (energy-efficient) Fugaku supercomputer was nearly \$60 millions

- Estimated drivers of change in electricity demand in the European Union, 2021-2026

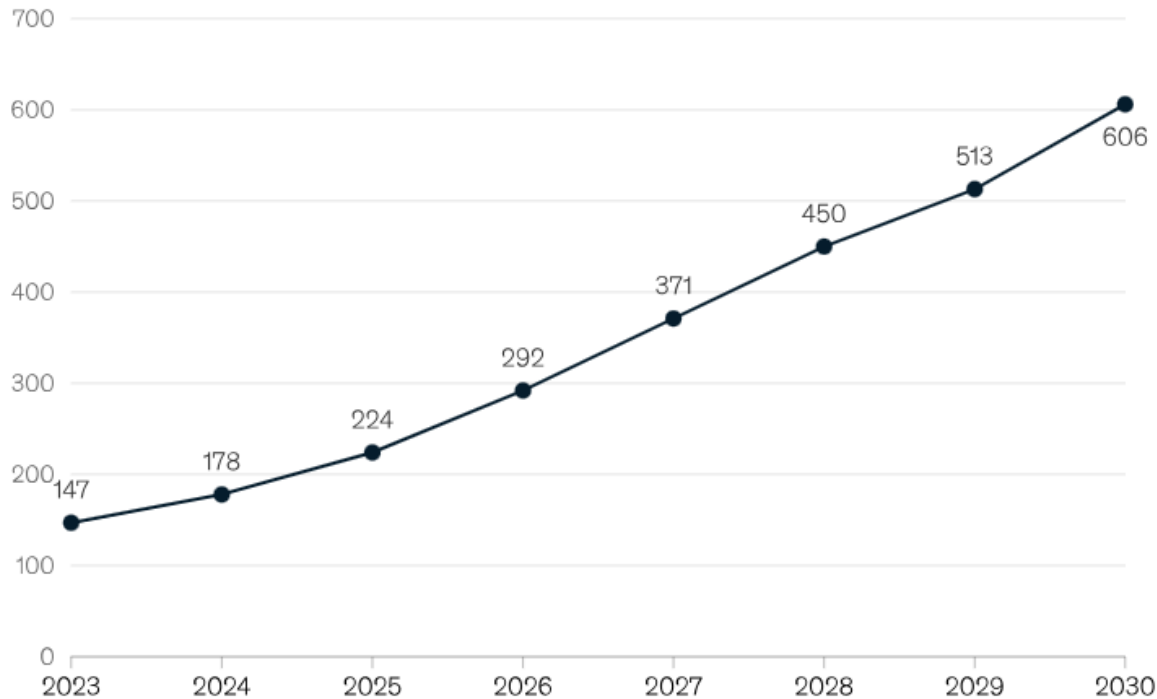


USA numbers..

Demand for power for data centers is expected to rise significantly in the United States.

Terawatt-hours (TWh) of electricity demand, medium scenario

**US data center energy
consumption, TWh**



**Share of total
US power
demand, %**

3.7 4.3 5.2 6.5 8.0 9.3 10.3 11.7

Electricity bill of TOP5 in 2022

| Machine | Peak Perf. | Power | \$/KWh | Total(K\$) |
|----------------|-------------------|--------------|---------------|-------------------|
| FRONTIER | 1.685 EFLOPS | 21.1MW | 0.150 | 3.165 |
| FUGAKU | 537.2 PFLOPS | 29.9MW | 0.219 | 6.548 |
| LUMI | 428.7 PFLOPS | 6.02MW | 0.198 | 1.192 |
| LEONARDO | 255.7 PFLOPS | 5.61MW | 0.561 | 3.147 |
| SUMMIT | 200.8 PFLOPS | 10.1MW | 0.150 | 1.515 |

Table 1: Electricity cost per hour for the top five supercomputers.

[Taken from: Energy Concerns with HPC Systems and Applications](#)

From Time to Energy..

- Problem is generally formulated as the need to reach a good trade-off between time-to-solution and energy-to-solution.
- Different approaches to solve this problem:
 - vendors work on power-efficient processors
 - software developers understand how to use them at the best
- effective solution is possible only by properly managing all layers of the system from the software stack to the cooling system.

Energy vs carbon footprint

- a more general concern currently in the spotlight.
- energy can be turned into carbon emission by multiplying it with the **carbon intensity** of the energy supply.
- If the power consumption of most hardware components is well known or can be measured accurately, it not the case with carbon emission, which has to (roughly) estimated by specific means.

Suggested reading: [A review on the decarbonization of high-performance computing centers – ScienceDirect](#)

Carbon intensity:

- Carbon intensity (CO_2e per Wh) is the amount of carbon dioxide (CO_2e) that is released to produce a watt-hour of electricity.
- The average data-center carbon emissions in 2020 was $0.429\ tCO_2e$ (ton of carbon dioxide equivalent emissions) per MWh (Megawatt hour),
- However: the gross tCO_2e per MWh can be 5x lower in some specific data-centers

(see this paper:

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9810097>

carbon footprint for top ranked supercomputers (2022)

| Machine | Peak Perf. | Power | Kg(CO ₂)/KWh | CO ₂ (kg\$) |
|----------|--------------|--------|--------------------------|------------------------|
| FRONTIER | 1.685 EFLOPS | 21.1MW | 0.379 | 7 997 |
| FUGAKU | 537.2 PFLOPS | 29.9MW | 0.479 | 14 322 |
| LUMI | 428.7 PFLOPS | 6.02MW | 0.132 | 795 |
| LEONARDO | 255.7 PFLOPS | 5.61MW | 0.372 | 2 087 |
| SUMMIT | 200.8 PFLOPS | 10.1MW | 0.379 | 3 828 |

Table 2: CO₂ per hour for the top five supercomputers.

- Observations:
 - floating-point performance and the necessary (CO_2e) are not directly correlated.
 - the hardware profile of the machines is a key factor.

Green500 list

- The Green500 list is published alongside the TOP500 list of the world's most powerful supercomputers.
- It only contains systems that feature in the TOP500 ranking.
- Ultimate goal: raise awareness of energy efficiency in HPC
- Green500's yardstick: performance per watt, i.e. efficiency
 - Performance: R_{max} the maximum performance achieved with Linpack on all the machine.
 - Power: average power consumption during the execution of Linpack on the problem that delivers R_{max} .

Top5 of the Green500

Green500 Data

| Rank | TOP500 Rank | System | Cores | Rmax (PFlop/s) | Power (kW) | Energy Efficiency (GFlops/watts) |
|------|-------------|--|--------|----------------|------------|----------------------------------|
| 1 | 189 | JEDI - BullSequana XH3000, Grace Hopper Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, ParTec/EVIDEN EuroHPC/FZJ Germany | 19,584 | 4.50 | 67 | 72.733 |
| 2 | 128 | Isambard-AI phase 1 - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE University of Bristol United Kingdom | 34,272 | 7.42 | 117 | 68.835 |
| 3 | 55 | Helios GPU - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cyfronet Poland | 89,760 | 19.14 | 317 | 66.948 |
| 4 | 328 | Henri - ThinkSystem SR670 V2, Intel Xeon Platinum 8362 32C 2.8GHz, NVIDIA H100 80GB PCIe, Infiniband HDR, Lenovo Flatiron Institute United States | 8,288 | 2.88 | 44 | 65.396 |
| 5 | 71 | preAlps - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Swiss National Supercomputing Centre (CSCS) Switzerland | 81,600 | 15.47 | 240 | 64.381 |

Taken for Green500 site

- The No. 1 spot on the GREEN500 was claimed by JEDI - JUPITER Exascale Development Instrument, a new system from EuroHPC/FZJ in Germany. Taking the No. 190 spot on the TOP500, JEDI achieved an energy efficiency rating of 72.73 GFlops/Watt while producing an HPL score of 4.5 PFlop/s. JEDI is a BullSequana XH3000 machine with a Grace Hopper Superchip 72C. It has 19,584 total cores.
- The Isambard-AI machine out of the University of Bristol in the U.K. claimed the No. 2 spot with an energy efficiency rating of 68.83 GFlops/Watt and an HPL score of 7.42 PFlop/s. Isambard-AI achieved the No. 129 spot on the TOP500 and has 34,272 total cores.
- The No. 3 spot was claimed by the Helios system from Cyfronet out of Poland. The machine achieved an energy efficiency score of 66.95 GFlops/Watt and an HPL score of 19.14 PFlop/s.
- Like the last list, the Frontier system deserves an honorable mention when discussing energy efficiency. Frontier achieved an exascale HPL score of 1.206 EFlop/s while also earning an energy efficiency score of 56.97 GFlops/Watt. This places the system at No. 11 on the GREEN500 in addition to its No. 1 spot on the TOP500.

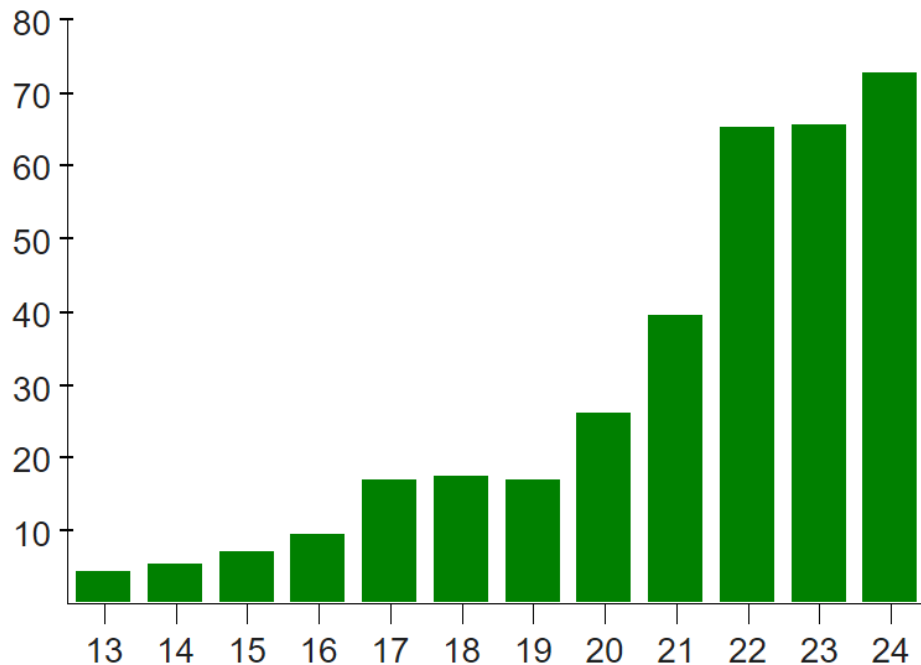
Green500

Efficiency over the years.

Historical development [\[edit\]](#)

Energy efficiency of top-ranked computers (gigaflops/watt)

(from 2013 to 2024)



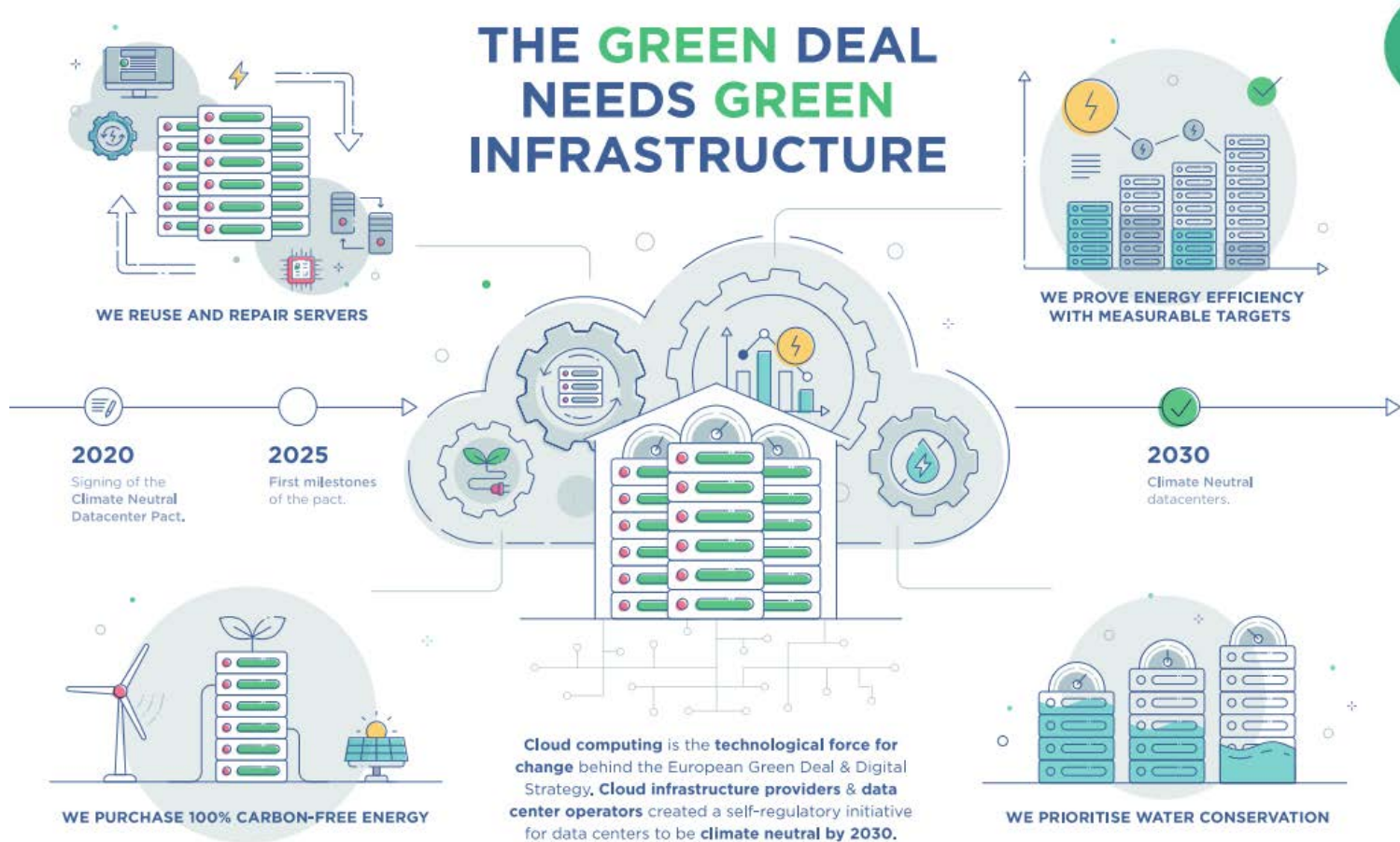
Picture taken from Wikipedia

Climate Neutral Data Centre Pact

- Launched Jan. 21st 2021
- Developed by the industry with the support of the European Commission (EVP Timmermans, DG Connect...)
- Engages Pact Operators (data centres and cloud infrastructure providers) towards Climate Neutrality by 2030 with clear metrics to achieve in 2025 and 2030 on:
 - Energy efficiency
 - Clean energy,
 - Water conservation,
 - Circular economy
 - Circular energy systems
- Engagements will be monitored by independent third-party audits
- Supported by biggest players of the industry operating in Europe (European and nonEuropean companies) as well as SMEs

<https://www.climateneutraldatacentre.net>

Climate Neutral data center Pact



Some interesting reading:

- Cost of AI models [Decoding the Energy Footprint of AI - Beyond Entropy \(substack.com\)](#)
- [The growing energy footprint of artificial intelligence - ScienceDirect](#)
- [European Exascale Supercomputer JUPITER Sets New Energy Efficiency Standards with #1 Ranking on Green500](#)
- [A review on the decarbonization of high-performance computing centers – ScienceDirect](#)
- [Energy Concerns with HPC Systems and Applications \(arxiv.org\)](#)

Energy measurement on HP/Data infrastructure

- two groups:
 - Out-of-band (e.g. power meters)
 - Out-of-band measurement is the easiest approach to consider.
 - It uses an external device to measure power consumption without a little to no interference in the computational performance.
 - in-band (e.g. RAPL counters).
 - requires some technical information about the target hardware
 - can access specific registers in a programmatic manner.
- Both types of measurement can be enhanced with **an application-level profiling**.
- However, it might be difficult to assess the type and detail of the measurements that are needed to obtain satisfactory insights from the energy profiling of the application.
- This is a major concern with the Out-of-band measurement, which uses an external device whose output data cannot be directly obtained within a program.

Physics 101

- Energy is in joule
- Power is in Watt
- Time is in second
 - Joule = Watt x second
- Watt-hour is a typical unit of energy
 - 1 Wh = 3600Joule

Standard Consumption metrics

$$E = \sum_{i=0}^n P_i \delta_i,$$

- In the above formula we assume a constant power P_i for time period δ_i , with $\delta_0 + \delta_1 + \dots + \delta_n = T$, where T is the overall time period considered.
- We can simplify considering $E = P \times T$ where P is the average power over the different interval of the previous formula.
- The variations of P and T are quite opposite
- the energy optimization of an HPC system is a matter of a good trade-off between the execution time and the consumed power.
- The goal is to optimize one while keeping the other at an acceptable level

Some concepts and their
definitions

TDP

- thermal design power (TDP) a.k.a thermal design point, is the maximum amount of generated heat (by a computer chip or component) that the cooling system is designed to dissipate.
- The purpose of the TDP is to provide system designers with a power target so as to guide the selection of a convenient thermal solution.
- The power rating (highest power input allowed) for a microprocessor is generally 1.5 times the TDP.
- Under a steady workload, the TDP is the maximum power consumption of the processors.

PUE

- Power usage effectiveness (PUE)

is a metric used to determine the energy efficiency of a data center.

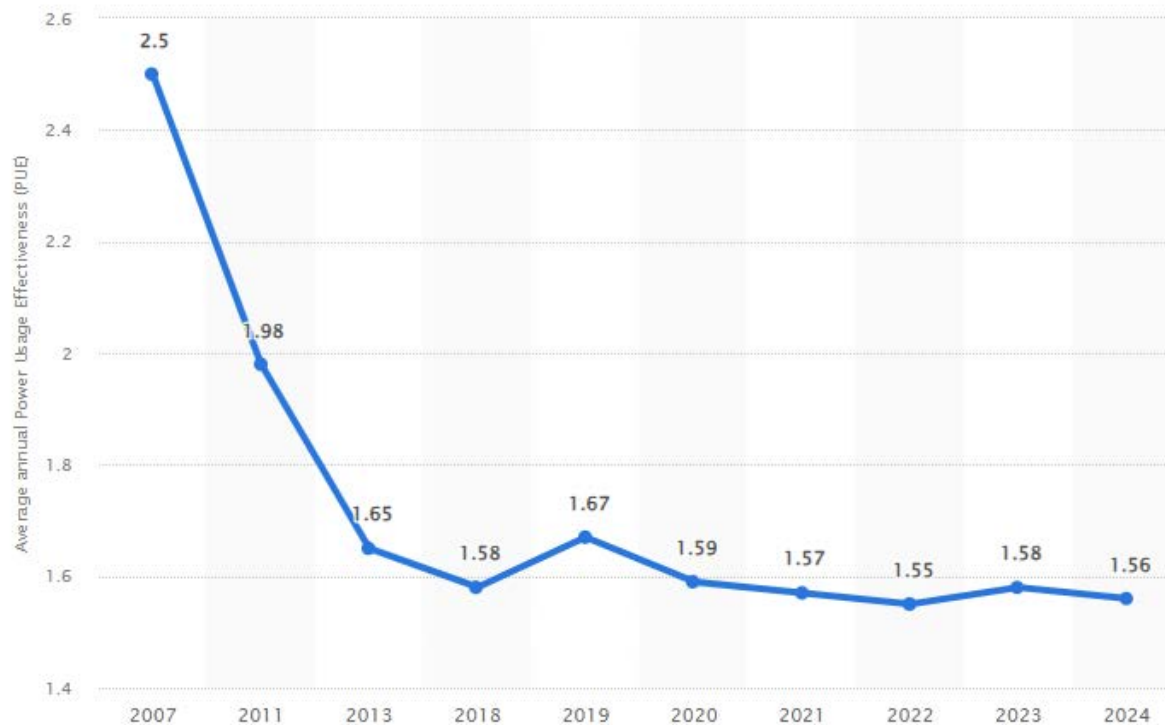
It is determined by dividing the total amount of incoming power by the consumed power.

$$PUE = \frac{Total_Facility_Energy}{IT_Equipment_Energy} = 1 + \frac{Non_IT_Facility_Energy}{IT_Equipment_Energy}$$

PUE (2)

- **Total Facility Power:** This encompasses all the power consumed by the data center facility. It includes power used for IT equipment (servers, storage, networking), cooling systems, lighting, power distribution units (PDUs), and any other supporting infrastructure.
- **IT Equipment Power:** This refers to the power consumed solely by the IT equipment within the data center. It's the energy that directly fuels the computational processes, storage operations, and network communications that drive the data center's purpose.

Some PUE numbers



© Statista 2024

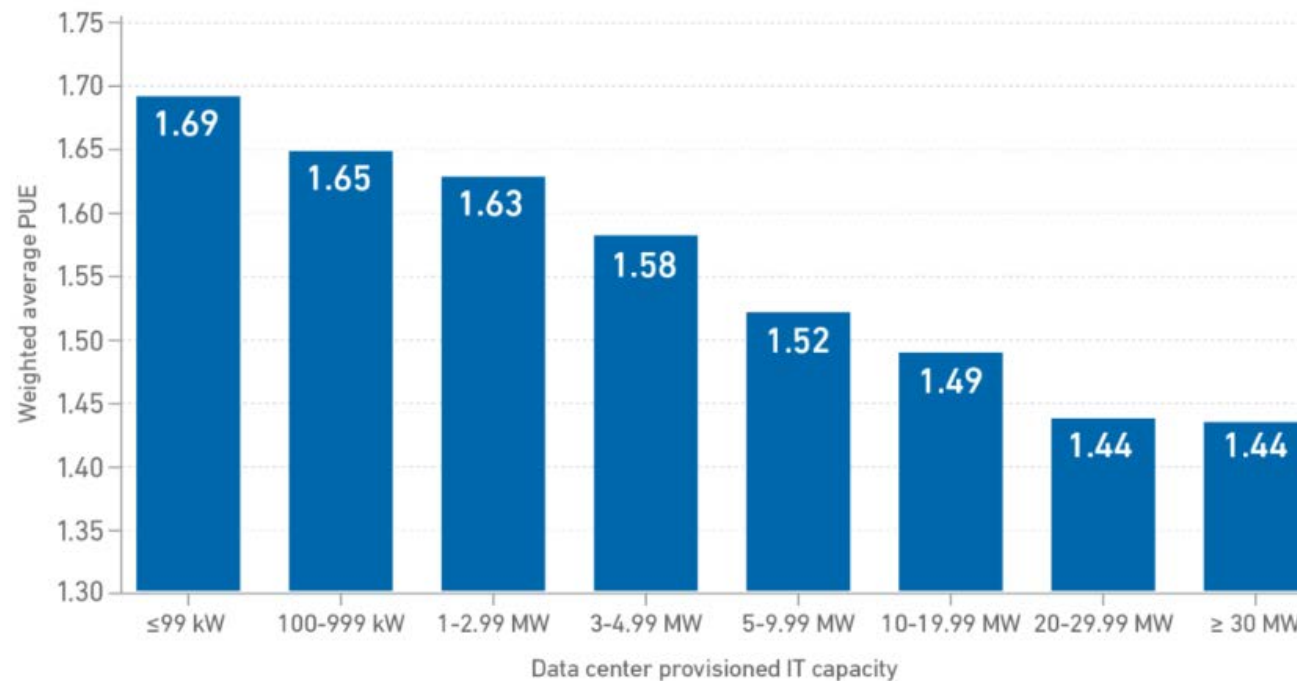
Show source

[Additional Information](#)

[Taken from: Data center average annual PUE worldwide 2024 | Statista](#)

Some PUE numbers

Figure 1. Weighted average PUE by data center IT capacity

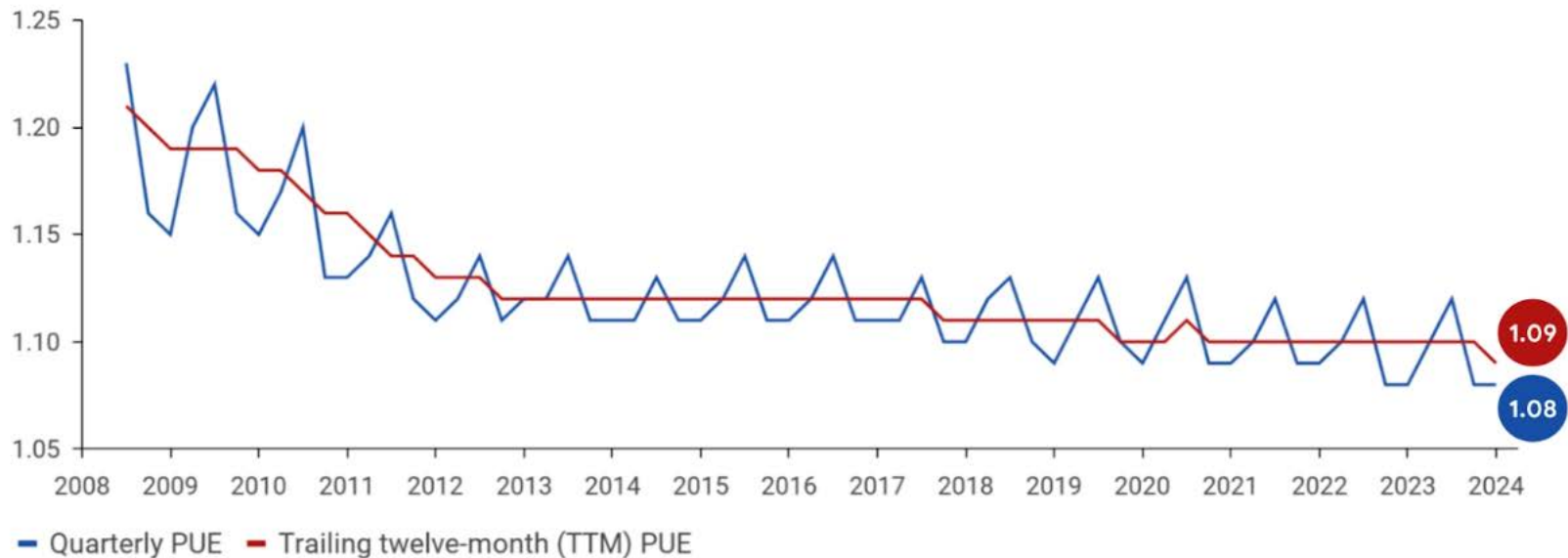


[Large data centers are mostly more efficient, analysis confirms - Uptime Institute Blog](#)

Some PUE numbers (3)

Continuous PUE Improvement

Average PUE for all data centers



Google data centers..

Google approach

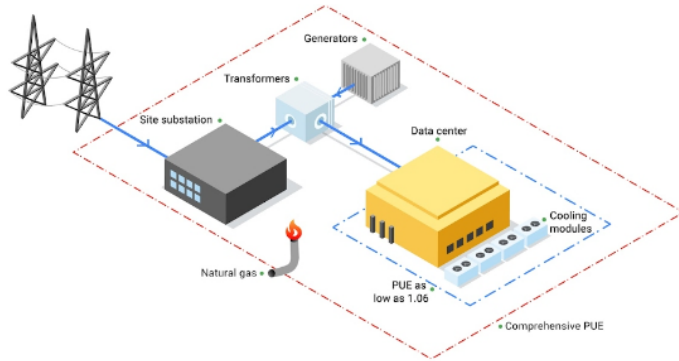


Figura 1: limiti di misurazione del PUE per i data center di Google. Il valore PUE medio per tutti i data center di Google è di 1,10; potremmo vantare un PUE di addirittura 1,06 adottando criteri più limitati.

Adottiamo l'approccio più completo alla misurazione del PUE (Power Usage Effectiveness)

I calcoli includono le prestazioni di tutti i nostri data center nel mondo e non solo quelle delle nostre strutture migliori e più recenti. Le misurazioni si susseguono con regolarità per tutto l'anno, includendo anche le stagioni più calde.

Inoltre, includiamo tutte le fonti di energia infrastrutturale nella nostra metrica relativa all'efficienza. Se assumessimo l'interpretazione meno rigida degli [standard di misurazione del PUE di Green Grid](#), potremmo inserire valori sensibilmente inferiori nel report. Di fatto, il nostro sito migliore potrebbe vantare un PUE inferiore a 1,06 se usassimo l'interpretazione comunemente adottata nel settore. Tuttavia, riteniamo che sia meglio misurare e ottimizzare tutto il sito e non solo una sua parte, pertanto preferiamo attenerci allo standard più elevato. Di conseguenza, nel report dichiariamo un PUE complessivo per i dodici mesi precedenti (TTM) di 1,10 per tutti i nostri data center su larga scala (una volta raggiunta un'operatività stabile) in tutte le stagioni, comprese tutte le fonti di energia infrastrutturale.

Other Green metric definitions .

| Metric Definition | Objective | Proposer | Ideal |
|---|---|---|--------|
| $PUE = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$ | Characterize the total energy efficiency of a data center | Green Grid, 2008 | →1 |
| $DCiE = \frac{1}{PUE} = \frac{\text{IT Equipment Power}}{\text{Total Facility Power}}$ | Characterize the total energy efficiency of a data center | Green Grid, 2008 | →1 |
| $DCeP = \frac{\text{Useful Work Produced}}{\text{Total Energy Consumed to Perform that Work}}$ | Characterize the IT computing efficiency | Green Grid, 2008 | Larger |
| $HVAC = \frac{\text{IT Electrical Energy}}{\text{HVAC} + (\text{Fuel} + \text{Steam} + \text{Chilled}) \times 293}$ | Characterize the energy efficiency of the HVAC system | Lawrence Berkeley National Laboratory, 2009 | Higher |
| $CEF = \frac{\text{Total CO}_2 \text{ Emissions}}{\text{Total Facility Energy}}$ | Assess the carbon emissions per unit of energy used | Green Grid, 2010 | Lower |
| $CUE = CEF \times PUE = \frac{\text{Total CO}_2 \text{ Emissions}}{\text{IT Equipment Energy}}$ | Characterize the overall efficiency of the cooling system | Green Grid, 2010 | Lower |
| $CSE = \frac{\text{Average Cooling System Power Usage}}{\text{Average Cooling Load}}$ | Represent the carbon emission efficiency of IT energy use | Lawrence Berkeley National Laboratory, 2009 | Lower |
| $AEU = \frac{\text{Air Economizer Hours}}{24 \times 365}$ | Measure the percentage of hours in a year that an air-side economizer system is used to provide “free cooling” | Lawrence Berkeley National Laboratory, 2009 | →100% |
| $WEU = \frac{\text{Water Economizer Hours}}{24 \times 365}$ | Measure the percentage of hours in a year that a water-side economizer system is used to provide “free cooling” | Lawrence Berkeley National Laboratory, 2009 | →100% |

Steps in power/energy management

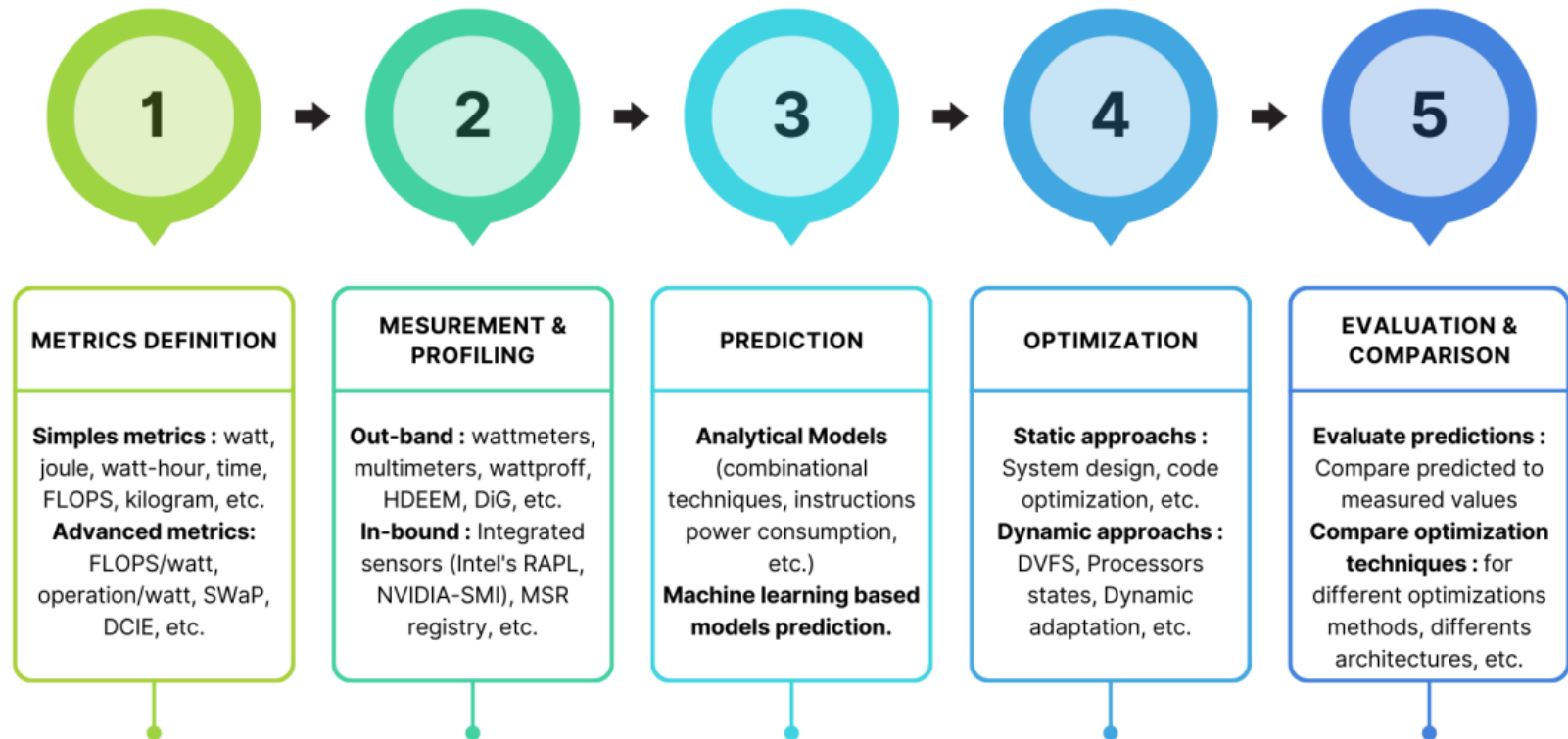


Figure 1: Taxonomy of power/energy management solutions with related approaches and tools.

[Taken from: Energy Concerns with HPC Systems and Applications \(arxiv.org\)](#)

Energy and HW architecture

- Accelerators:
 - GPU
 - TPU
 - FPGA
- CPU
 - x86 based processors
 - ARM based processors

GPUs

- Very energy intensive
- Very efficient: they show better performance-per-watt than standard CPUs for specific workloads (i.e. HPL)
- If you want Top500 buy GPUs otherwise the electricity bill skyrockets

Accelerators in 2022

| Name | RAM (GB) | core frequency (GHz) | TDP (w) | Peak TOPS | peak TFLOPS(fp32) | performance/watt (INT8) |
|---------------------|----------|----------------------|------------|-----------------------|-------------------|-------------------------|
| Tesla A100 SXM4 | 80 | 1.41 | 400 | 312(bf16)/624(int8) | 19.5 | 1.56 TOPS/W |
| Tesla H100 SXM5 | 80 | 1.98 | 700 | 1000(bf16)/2000(int8) | 60 | 3.33 TOPS/W |
| AMD Instinct MI250X | 128 | 1.7 | 560 | 383 (bf16 or int8) | 95.7 | 0.68 TOPS/W |
| Intel Ponte Vecchio | 128 | 1.6 | 600 | 720(bf16)/1440(int8) | 45 | 2.40 TOPS/W |
| Google TPU v4 | 32 | 1.05 | 192 (idle) | 275 (bf16 or int8) | / | 1.43 TOPS/W |

Table 3: SOTA accelerators systems characteristics

Accelerators in 2024

| NVIDIA Flagship Accelerator Specification Comparison | | | |
|--|-----------------------------------|----------------------------------|----------------------------------|
| | B200 | H100 | A100 (80GB) |
| FP32 CUDA Cores | A Whole Lot | 16896 | 6912 |
| Tensor Cores | As Many As Possible | 528 | 432 |
| Boost Clock | To The Moon | 1.98GHz | 1.41GHz |
| Memory Clock | 8Gbps HBM3E | 5.23Gbps HBM3 | 3.2Gbps HBM2e |
| Memory Bus Width | 2x 4096-bit | 5120-bit | 5120-bit |
| Memory Bandwidth | 8TB/sec | 3.35TB/sec | 2TB/sec |
| VRAM | 192GB (2x 96GB) | 80GB | 80GB |
| FP32 Vector | ? TFLOPS | 67 TFLOPS | 19.5 TFLOPS |
| FP64 Vector | ? TFLOPS | 34 TFLOPS | 9.7 TFLOPS (1/2 FP32 rate) |
| FP4 Tensor | 9 PFLOPS | N/A | N/A |
| INT8/FP8 Tensor | 4500 T(FL)OPS | 1980 TOPS | 624 TOPS |
| FP16 Tensor | 2250 TFLOPS | 990 TFLOPS | 312 TFLOPS |
| TF32 Tensor | 1100 TFLOPS | 495 TFLOPS | 156 TFLOPS |
| FP64 Tensor | 40 TFLOPS | 67 TFLOPS | 19.5 TFLOPS |
| Interconnect | NVLink 5 18 Links (1800GB/sec) | NVLink 4 18 Links (900GB/sec) | NVLink 3 12 Links (600GB/sec) |
| GPU | "Blackwell GPU" | GH100 (814mm2) | GA100 (826mm2) |
| Transistor Count | 208B (2x104B) | 80B | 54.2B |
| TDP | 1000W | 700W | 400W |
| Manufacturing Process | TSMC 4NP | TSMC 4N | TSMC 7N |
| Interface | SXM | SXM5 | SXM4 |
| Architecture | Blackwell | Hopper | Ampere |

DATA SHEET

AMD INSTINCT™ MI325X ACCELERATOR

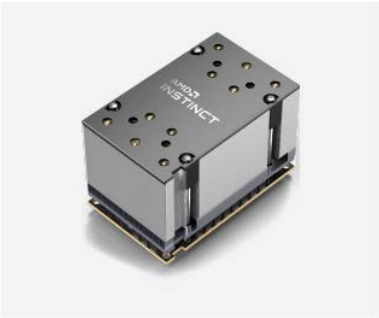
Leading-Edge, industry-standard accelerator module for generative AI, inference, training, and high performance computing


Designed to Accelerate Modern Workloads

The increasing demands of generative AI, large-language models, inference, and machine learning training puts next-level demands on GPU accelerators. The discrete AMD Instinct MI325X GPU delivers superior performance on a broad set of data types needed for AI software, including FP16, BF16, FP8, and INT8 used in both high-precision inference and training.^{[MI325-004](#)} An industry-leading 256 GB of HBM3E memory ^{[MI325-004](#)} and 6 TB/s bandwidth enables a single accelerator to contain and process a one-trillion parameter model while reducing total cost of ownership for select large-language models. ^{[MI325-003](#)}

Support for matrix sparsity further economizes memory use and boosts computational speed, helping enable sustainable scaling of AI solutions across data centers, speeding time to market and enhancing performance.

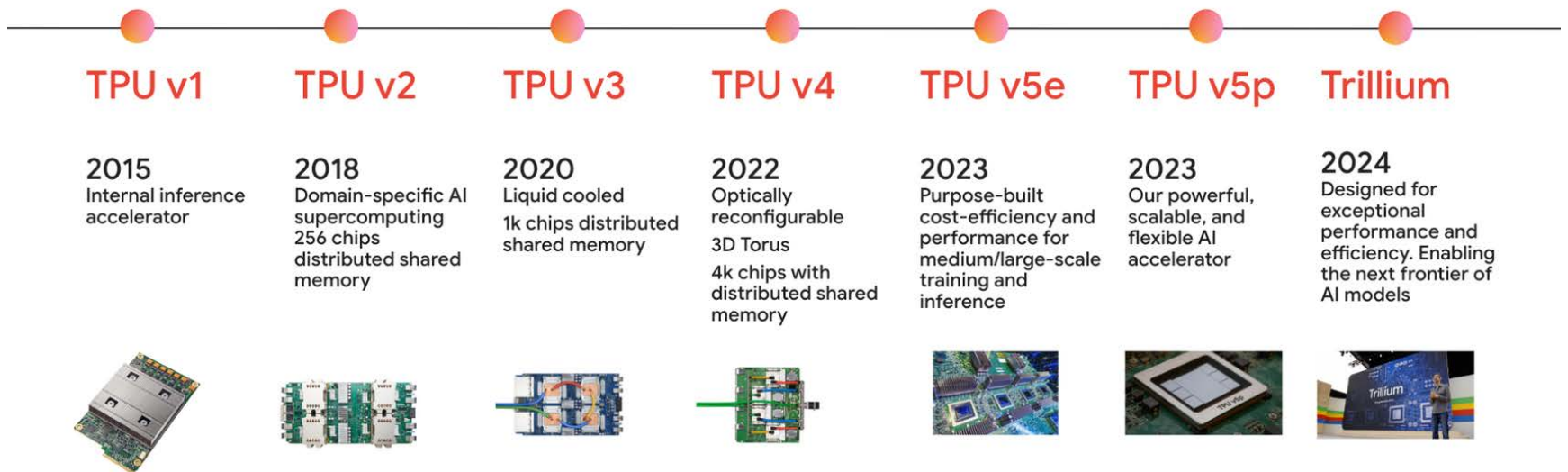
Integrated with AMD ROCm™ software, the accelerator supports key AI and HPC frameworks, simplifying deployment. Seamless, drop-in compatibility with the AMD Instinct MI300X Platform comes through strong support from our OEM partners, industry-leading frameworks, and thousands of large-language models.



| | | | | | |
|--|---|----------------------------------|-------------------------------------|---|------------|
|  | AI PEAK THEORETICAL PERFORMANCE | | SPECIFICATIONS | | |
| | | with sparsity | Form factor | OAM module | |
| | TF32 (TFLOPs) | 653.7 | 1307.4 | Lithography | 5nm FinFET |
| | FP16 (TFLOPs) | 1307.4 | 2614.9 | Active interposer dies (AIDs) | 6nm FinFET |
| | BFLOAT16 (TFLOPs) | 1307.4 | 2614.9 | GPU compute units | 304 |
| | INT8 (TOPS) | 2614.9 | 5229.8 | Matrix cores | 1216 |
| | FP8 (TFLOPs) | 2614.9 | 5229.8 | Stream processors | 19,456 |
| HPC PEAK THEORETICAL PERFORMANCE (TFLOPS) | | | Peak engine clock | 2100 MHz | |
| FP64 vector | 81.7 | Memory capacity | Up to 256 GB HBM3E | | |
| FP32 vector | 163.4 | Memory bandwidth | 6 TB/s max. peak theoretical | | |
| FP64 matrix | 163.4 | Memory interface | 8192 bits | | |
| FP32 matrix | 163.4 | AMD Infinity Cache™ (last level) | 256 MB | | |
| DECODERS AND VIRTUALIZATION | | | Memory clock | Up to 6.0 GT/s | |
| Decoders¹ | 4 groups for HEVC/H.265, AVC/H.264, VP9, or AV1 | | Scale-up AMD Infinity Fabric™ Links | 7x 128 GB/s | |
| JPEG/MJPEG CODEC | 32 cores, 8 cores per group | | I/O to host CPU | 1 PCIe® Gen 5 x16 (128 GB/s) | |
| Virtualization support | SR-IOV, up to 8 partitions | | Scale-out network bandwidth | PCIe Gen 5 x16 (128 GB/s) | |
| ¹Video codec acceleration (including at least the HEVC (H.265), H.264, VP9, and AV1 codecs) is subject to change and not operable without inclusion/installation of compatible media players. GD-176 | | | RAS features | Full-chip ECC memory, page retirement, page avoidance | |
| | | | Maximum TBP | 1000W | |

Google TPU

TPU AI accelerators



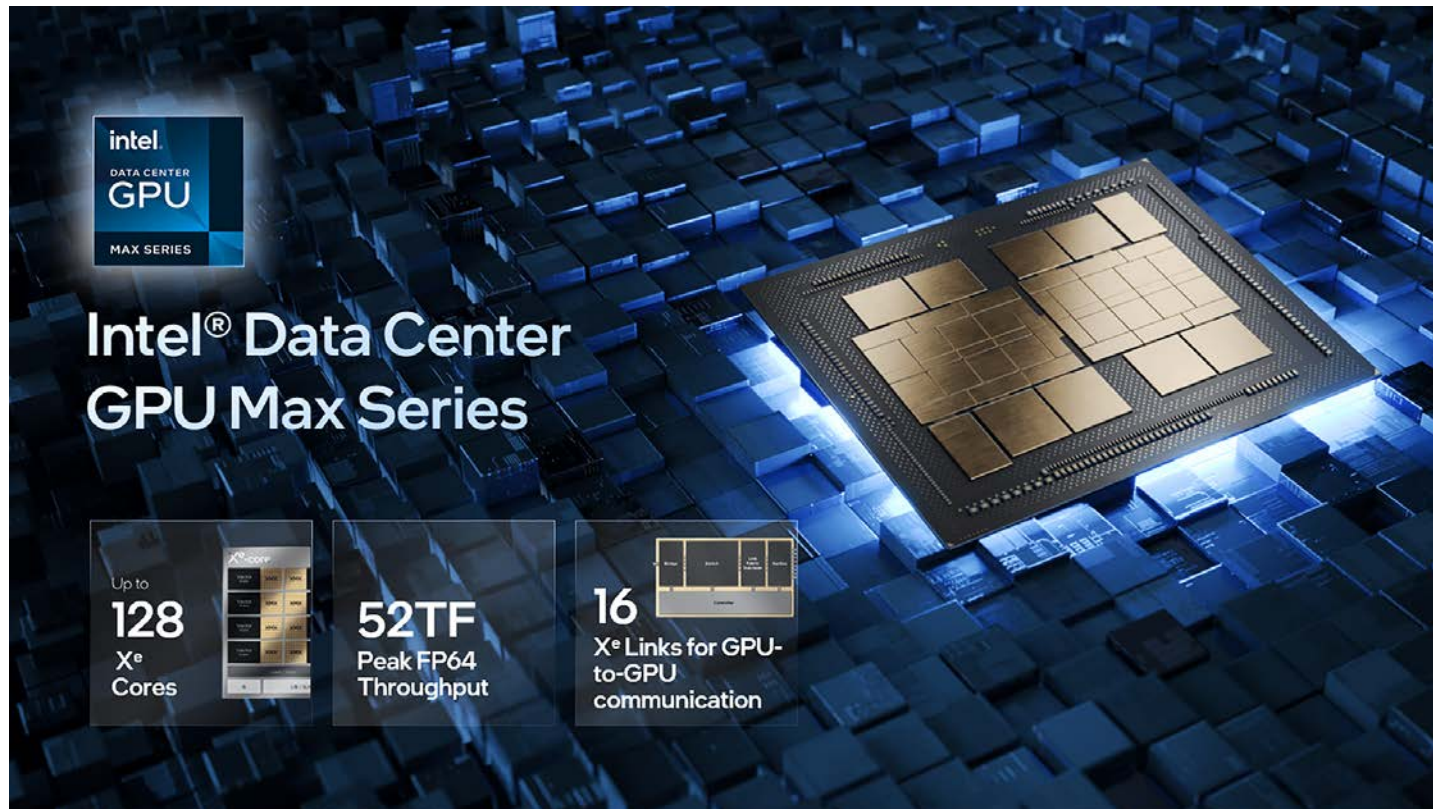
Some more details on TPU..

| Google TPU Compute Engines | TPU v1 | TPU v2 | TPU v3 | TPU v4i | TPU v4 | TPU v5p | TPU v5e | "Trillium" TPU v6 |
|----------------------------|---------------------|---------------------|---------------------|---------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| First Deployed | Q2 2015 | Q3 2017 | Q4 2018 | Q1 2020 | Q4 2021 | Q4 2023 | Q3 2023 | Q4 2024 |
| ML Inference | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| ML Training | No | Yes | Yes | No | Yes | Yes | Yes | Yes |
| Chip Process | 28 nm | 16 nm | 16 nm | 7 nm | 7 nm | 5 nm | 5 nm | 4 nm |
| Transistors | 3 B | 9 B | 10 B | 16 B | 31 B | ??? | ??? | ??? |
| Die Size | 330 mm ² | 625 mm ² | 700 mm ² | 400 mm ² | 780 mm² | 700 mm² | 350 mm² | 790 mm² |
| Clock Speed | 700 MHz | 700 MHz | 940 MHz | 1,050 MHz | 1,050 MHz | 2,040 MHz | 1,750 MHz | 2,060 MHz |
| TensorCores Per Chip | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 |
| MXU Matrix Size/Core | 1 * 256x256 | 1 * 128x128 | 2 * 128x128 | 4 * 128x128 | 4 * 128x128 | 4 * 128x128 | 4 * 128x128 | 4 * 256x256 |
| Dataflow SparseCores | - | - | - | - | 4 | 4 | 2 | 4 |
| On Chip Cache Memory | 28 MB | 32 MB | 32 MB | 144 MB | 32 MB | 48 MB | 112 MB | ??? |
| Off Chip HBM Memory | 8 GB | 16 GB | 32 GB | 8 GB | 32 GB | 95 GB | 16 GB | 32 GB |
| HBM Memory Bandwidth | 300 Gb/sec | 700 GB/sec | 900 GB/sec | 300 GB/sec | 1,228 GB/sec | 2,765 GB/sec | 819 GB/sec | 1,640 TB/sec |
| INT8 Peak Teraflops | 92 | - | - | 138 | 275 | 918 | 394 | 1,852 |
| BF16 Peak Teraflops | - | 46 | 123 | 69 | 137.5 | 459 | 197 | 926 |
| Precision | INT8 | BF16 | BF16 | BF16/INT8 | BF16/INT8 | BF16/INT8 | BF16/INT8 | BF16/INT8 |
| ICI Links * Speed Gb/sec | - | 4 * 496 | 4 * 656 | 2 * 400 | 6 * 448 | 6 * 800 | 4 * 400 | 4 * 800 |
| Interconnect Topology | - | 2D Torus | 2D Torus | - | 3D Torus | 3D Torus | 2D Torus | 2D Torus |
| Chip Idle Watts | 28 | 53 | 84 | 55 | 170 | ??? | ??? | ??? |
| Max Measured Watts | ??? | ??? | 262 | ??? | 192 | ??? | ??? | ??? |
| Chip TDP Watts | 75 | 280 | 450 | 175 | 300 | ??? | ??? | ??? |

Taken from: [Lots Of Questions On Google's "Trillium" TPU v6, A Few Answers](#)

Intel GPU

- See [Intel Data Center GPU Max Series](#)



The advertisement features a large, detailed image of an Intel Data Center GPU Max Series chip, which is a square, gold-colored integrated circuit with a complex grid of internal components. The chip is set against a dark blue background with a grid of glowing, three-dimensional cubes, creating a sense of depth and technological sophistication. The lighting is dramatic, with bright highlights on the chip's edges and the cubes, and deep shadows elsewhere.

intel.
DATA CENTER
GPU
MAX SERIES

**Intel® Data Center
GPU Max Series**

Up to
128
Xe
Cores

52TF
Peak FP64
Throughput

16
Xe Links for GPU-
to-GPU
communication

CPUs

Intel vs AMD vs ARM

- x86 architecture family for desktop/server
- ARM architecture for mobile categories such as smartphones or tablets.
- The market of CPUs in the HPC landscape and data centers is still dominated today by x86 CPUs.
- AMD claim that its EPYC processors power the most energy efficient x86 servers, delivering exceptional performance with lower energy consumption.
- AMD EPYC 9654 servers shall use up to 29% less annual power than Intel Xeon Platinum 8490H servers at the same performance.
- TDP:
 - AMD 360 Watt
 - INTEL 350 Watt

ARM in HPC

- Several international collaborative projects started promoting ARM processors in HPC Area
- European Mont-Blanc: **The Mont-Blanc Project: A Strategy for Enhancing Computer and Energy Efficiency**
- On November 2018, for the first time, an ARM-based system was listed in the Top500 ranking. It was the Astrasupercomputer powered by Marvell's ThunderX2 ARM CPU and hosted at the Sandia National Laboratories (USA).
- Fugaku, with its A64FX ARM architecture based CPU was the #1 in TOP500 June 2020

A64FX

| General information | |
|---------------------------------|---|
| Launched | 2019 |
| Marketed by | Fujitsu |
| Designed by | Fujitsu |
| Common manufacturer | TSMC |
| Architecture and classification | |
| Technology node | 7 nm |
| Microarchitecture | In-house |
| Instruction set | ARMv8.2-A with SVE and SBSA level 3 |
| Physical specifications | |
| Cores | 48 per CPU ^[1] plus optional assistant cores ^{[2][3]} |
| History | |
| Predecessor | SPARC64 V |

A comparison among them

| Name | memory(GB) | core frequency (GHz) | TDP (w) | Peak TOPS | Peak GFLOPS (fp64) | performance/watt |
|----------------------|------------|----------------------|---------|------------|--------------------|------------------|
| Intel Platinum 8490H | 4000 | 1.9 | 350 | / | 3 648 | 10.42 GFLOPS/W |
| AMD EPYC 9654 | 6000 | 2.4 | 360 | 7763(int8) | 3 686 | 10.23 GFLOPS/W |
| Fujitsu A64FX | 32 | 2.6 | 150 | 3.4(int8) | 3 400 | 22.66 GFLOPS/W |
| Marvell ThunderX2 | 512 | 2.2 | 180 | / | 563 | 3.12 GFLOPS/W |

Table 5: SOTA General purpose computers characteristics

Energy Management Tools for GPUs

- GPU
 - Nvidia-smi
- The tool can be used to set the power range (max and min, inWatt) of the execution of a given application. Its GPU Operation Mode (GOM) allows to reduce the power usage and optimize the GPU throughput by disabling some features accordingly. It also implements a power scaling algorithm to dynamically reduce the clock frequency when the GPU is consuming too much power.

Energy Management Tools for CPUs

- Intel RAPL (Running Average Power Limit Interface)
 - RAPL measurements ignore a large part of the power consumption of servers because they focus on CPU and RAM.
 - Some experiments on Intel processor show that it just represent 42% of the overall servers consumption
- AMD RAPL counters
 - The implementation is similar to the corresponding Intel's RAPL, but uses different control registers.
 - While Intel typically provides multiple domains and the option to limit power consumption over various time frames, AMD only considers registers for memory reads and core power consumption

Energy Management Tools for CPUs

- Model-Specific Register (MSR)
 - is any of the various control registers in the x86 architecture used for debugging, program execution tracing, computer performance monitoring, and toggling certain CPU features
- ACPI (Advanced Configuration and Power Interface)
 - an open standard that the operating system can use to discover and configure the components of the computer, to perform power management, auto configuration, and status monitoring. (see later)

Energy Tools for CPU

| Name | Type | Objective | Techniques | Portability |
|-------------------------------|---------------|---|---|---|
| RAPL counters[65] | hardware | power management | Dynamic Power Management, Power capping, Sampling measurement | x86 CPU |
| ACPI[130] | specification | power management | Dynamic Power Management, Power capping | x86 CPU |
| DT[80] | specification | power management | Dynamic Power Management, Power capping | ARM CPU |
| Perf tools[38] | software | performance and energy management | interface to hardware counters | Linux with Intel devices for energy |
| PAPI[88] | software | performance and energy management interface | interface to hardware counters | All Linux systems |
| Likwid-powermeter[129] | software | power profiling | query RAPL counters | Linux devices with Intel processor |
| PowerTOP[62] | software | energy monitoring | query Intel RAPL | Linux with AMD or Intel devices |
| PyJoules[59] | software | energy monitoring | query RAPL and Nvidia SMI interfaces | Linux with AMD, Nvidia or Intel devices |
| Powerstat[26] | software | measure energy consumption | query Intel RAPL | Linux on Intel PCs |
| Power Gadget and PowerLog[61] | software | energy/power and temperature monitoring | query Intel RAPL | Mac or Windows on Intel PCs |
| tx2mon[87] | software | energy/power and temperature monitoring | query hardware counters | Marvell ThunderX2 |

Table 7: SOTA General purpose computers tools for energy/power management

Energy tool for HPC system

| Name | Type | Objective | Technique | Portability |
|------------------|----------|--|---|--|
| WattProf[111] | hardware | energy/power measurement | system on-chip based power monitoring board | all server node |
| HDEEM[52] | hardware | power measurement | system on-chip based power monitoring board | all server node |
| DiG[78] | hardware | energy monitoring | system on-chip based power monitoring board | all server node |
| PowerPack[42] | software | energy/power measurement | isolate power consumption of devices in measurement | Linux systems |
| EERT[20] | software | power management | dynamic rescheduling, core usage maximization | Linux HPC systems |
| Phase-TA[123] | software | energy profiling | analysing the profiles of HPC applications | Linux systems |
| PMAC[21] | software | power management | DPM and DVFS techniques; web based monitoring | Linux systems |
| BDPO[123] | software | power optimization | DFS on computing cores during workloads execution | Linux x86 systems |
| lo2s[57] | software | performance and energy profiling | Sample hardware counters events | Linux x86 systems |
| EAR[77] | software | energy management | DPM techniques, power capping, On/Off policies | Linux with Intel, AMD and Nvidia devices |
| READEX[98] | software | energy and performance optimization | exploit the dynamic behaviour of application and make resources allocation | Linux x86 and ARM systems |
| MERIC[134] | software | energy management | dynamic application tuning and hardware energy measurement | Linux x86, ARM and Nvidia GPUs systems; HDEEM and DiG supports |
| FIRESTARTER[120] | software | benchmark tests of cooling and maximum power consumption | stress execution units and data transfer between cores and memory hierarchy | x86 CPU and GPU |

Table 9: SOTA Supercomputers systems tools for energy/power management

How to reduce the power consumption of HPC resources?

- policy-based automatic power management (idle nodes into power saving modes, power on/wake nodes for new workload, ...)
- exploit hardware capabilities: DVFS / power-saving states / performance states / turbo mode
- power capping policies (maximum amount of overall admitted power consumption)
- assign workload to highest performance-per-watt resources first
- energy-aware resource management systems and schedulers able to exploit all of the above, implementing out-of-band and unattended energy assessment capabilities

Energy optimization techniques

```
graph TD; A[Energy optimization techniques] --> B[Static approaches]; A --> C[Dynamic approaches]; B --> D[Hardware]; B --> E[Software]; C --> F[Hardware]; C --> G[Software]; D --> D1[• Transistor and Circuit design]; D --> D2[• Multicore/Multithread]; D --> D3[• Hybrid CPU design]; D --> D4[• Energy aware dedicated architectures]; D --> D5[• Heterogeneous architectures]; E --> E1[• Compilation optimization]; E --> E2[• Data structure organization]; E --> E3[• Programming rules]; E --> E4[• Programming languages efficiency]; E --> E5[• Process binding/pinning]; E --> E6[• Energy prediction]; E --> E7[• Power Capping]; F --> F1[• Dynamic Component Deactivation]; F --> F2[• DVS, DFS and DVFS]; F --> F3[• On/Off policies]; F --> F4[• P and C states]; F --> F5[• CPU clock speed variation]; G --> G1[• Dynamic Power Management]; G --> G2[• Jobs rescheduling]; G --> G3[• Dynamics adaptation]; G --> G4[• Workload balancing]; G --> G5[• Workload consolidation techniques]; G --> G6[• Workload peak reduction];
```

Static approaches

Hardware

- Transistor and Circuit design
- Multicore/Multithread
- Hybrid CPU design
- Energy aware dedicated architectures
- Heterogeneous architectures

Software

- Compilation optimization
- Data structure organization
- Programming rules
- Programming languages efficiency
- Process binding/pinning
- Energy prediction
- Power Capping

Dynamic approaches

Hardware

- Dynamic Component Deactivation
- DVS, DFS and DVFS
- On/Off policies
- P and C states
- CPU clock speed variation

Software

- Dynamic Power Management
- Jobs rescheduling
- Dynamics adaptation
- Workload balancing
- Workload consolidation techniques
- Workload peak reduction

Some more detailed definitions

- Thermal Design Power (TDP):
 - is the maximum amount of heat generated by a computer chip or component (often a CPU, GPU or system on a chip) that the cooling system in a computer is designed to dissipate under any workload.
 - The processor's rated frequency assumes that all execution cores are running an application at TDP level
- P-states:
 - During the execution of code, the operating system and CPU can optimize power consumption through different p-states (performance states). Depending on the requirements, a CPU is operated at different frequencies. P0 is the highest frequency (with the highest voltage).
- C-states:
 - Unlike the P-States, which are designed to optimize power consumption during code execution, C-States are used to optimize or reduce power consumption in idle mode (i. e. when no code is executed).

Hardware capabilities

- Advanced Configuration and Power Interface (ACPI) defines:
 - sleeping states (S-states)
 - power (core) states (C-states)
 - performance states (P-states)
- Such of the above solutions involves methods like
 - DVFS (Dynamic Voltage and Frequency Scaling)
 - RFTS(Run Fast Then Stop) mechanisms and power/clock gating
 - Turbo Boost technology (INTEL specific)
- The above tricks dynamically configure and monitor the power consumption
- All these features, which are implemented at hardware level by the CPUs, can be enabled by compliant motherboard's BIOS and exposed as a control knob to the operating system for run-time power-optimization

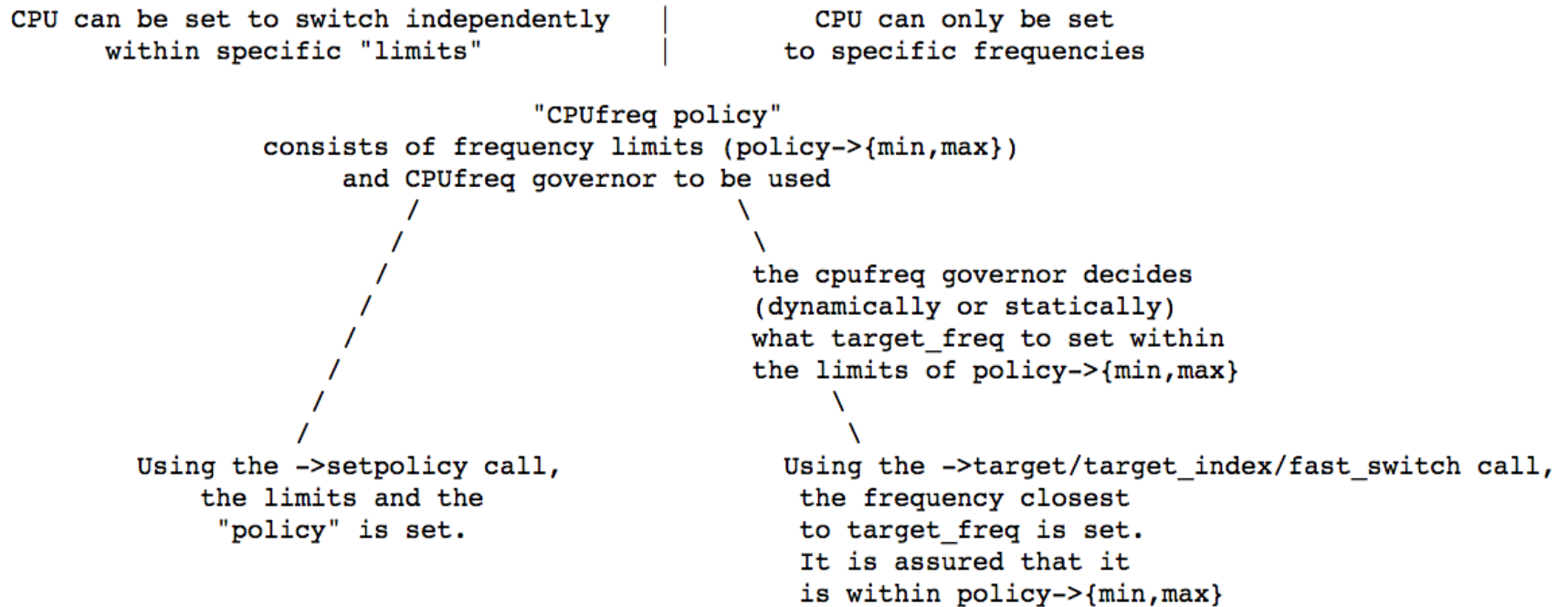
HW tricks: DVFS

- Observation
 - Power \propto voltage² \propto frequency
 - Performance \propto frequency
- Mechanism: Dynamic Voltage & Frequency Scaling (DVFS)
 - Allows changes to CPU voltage & frequency at run time
 - Trades CPU performance for power reduction
- Uses commodity technology[?]
- Policy: DVFS Scheduling
- Determines
 - WHEN to adjust the frequency-voltage setting, and
 - WHAT the new setting should be.

How can I change the frequency?

How to decide what frequency within the CPUfreq policy should be used?
That's done using "cpufreq governors".

Basically, it's the following flow graph:



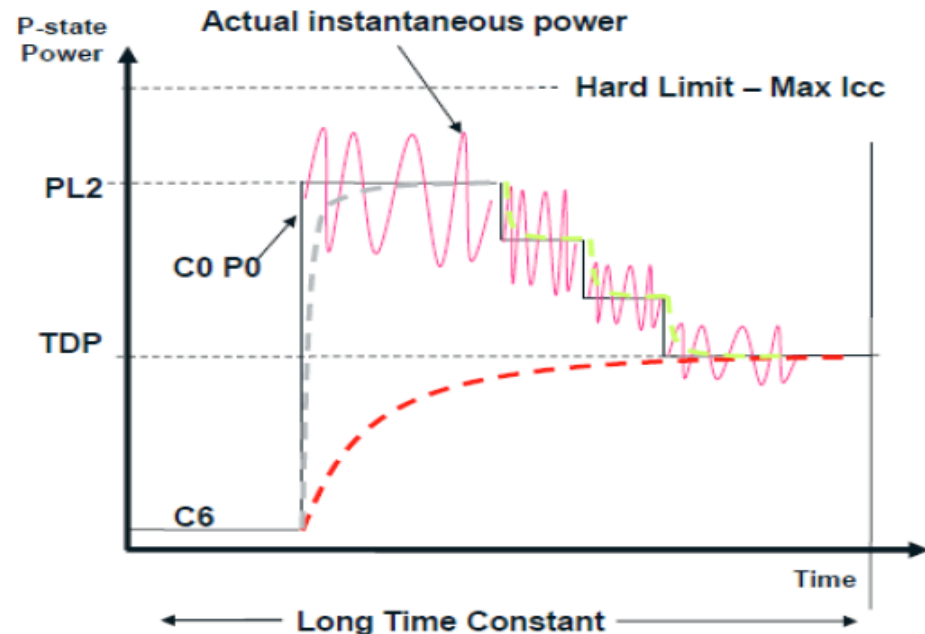
From: <https://www.kernel.org/doc/Documentation/cpu-freq/governors.txt>

HW tricks: intel turbo boost

Intel® Turbo Boost Technology Behavior

Haswell Intel® Turbo Boost Technology uses transient headroom:

- Can briefly exceed TDP for maximum performance.
- Temperatures ramp more quickly, but no impact to “steady state” condition.
- Transient power limited by power delivery capacities of platform (PL2).





EfiMon: A Process Analyser for Granular Power Consumption Prediction

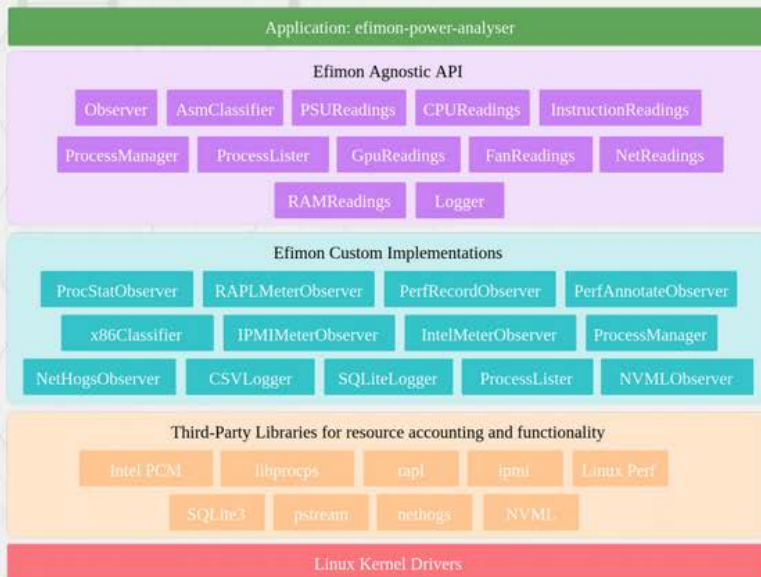
Luis G. León-Vega (1), Niccolò Tosato (1,2), Stefano
Cozzini (2)

University of Trieste, Area Science Park

Efimon requirements

- Energy consumption at the system level: not only the CPU or the socket
- Energy consumption at the process level: capability of isolating the running processes.
- Energy consumption per component: CPU, DRAM, network, I/O and accelerators

EfiMon: Efficiency Monitor



- **Instruments:** Intel PCM, ProcPS, RAPL, IPMI, Linux Perf, NetHogs, NVML.
- **Hexagonal (or interface-adapter) architecture:** Readings, Observer, Logger
- **Daemon-Client structure:** for rootless mode.
- **Characteristics:**
 - Extensible
 - Compilation aware
 - Selective

- Demo/Tutorial on next Wednesday

The end