Top Green500 - Contexto e desafios

Anahí Coimbra Maciel
Eduardo Berthold Kapp
Gabriel Costa R T Almeida

Sumário

1. Contexto: o que é o top500 Green
2. Atual top 10
3. Análise do histórico recente
4. Soluções para eficiência energética
5. Desafios e futuro

Contexto

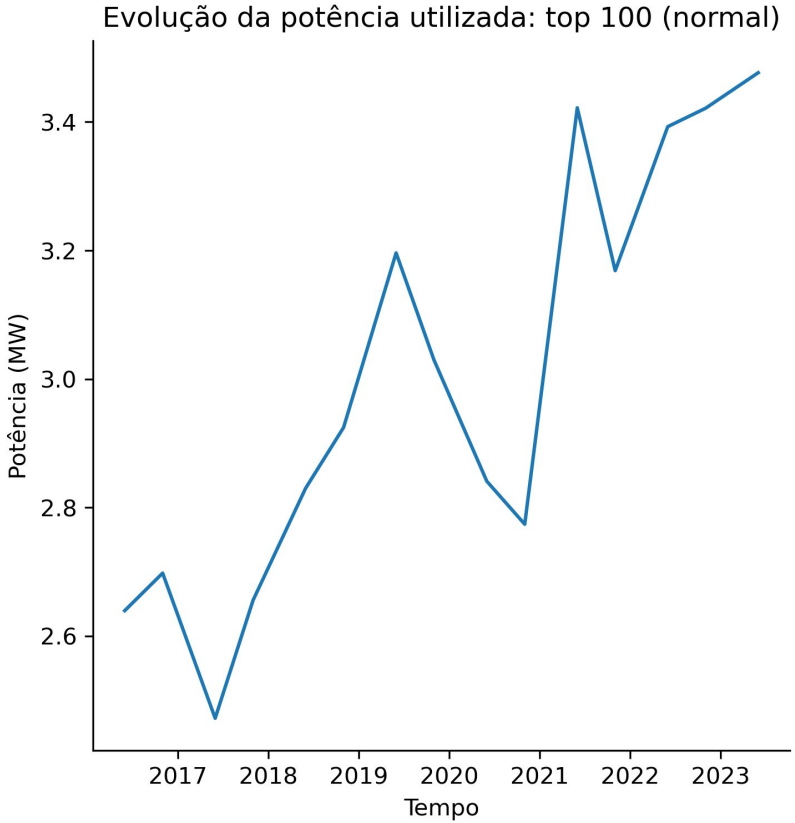
Top Green 500?

- **O que é**: uma versão da lista de top 500 supercomputadores, ordenada pela métrica flops/W (floating point operations per second per watt)
- **Origem**: 2007, Virginia Tech. Kirk Cameron, Wu Feng
- **Propósito**: encorajar a evolução de supercomputadores considerando não apenas a performance pura como a eficiência energética.

Contexto

Obs: para efeito de escala, uma cidade pequena no estado de SP tem consumo médio entre 3~10MWh por ano.

https://dadosenergeticos.energia.sp.gov.br//PortalCEv2/Municipios/Eletricidade/Detalhes_RA_Eletro.asp?ano=2019&ra=4&nome=CAMPINAS Acessado em 02/07/23



Contexto

O aumento contínuo de consumo energético é insustentável, tanto ambientalmente quanto financeiramente.

Assumindo um supercomputador com demanda média de 5 MW, taxa de utilização de 70% e custo médio no mercado livre de energia brasileiro (R$60~R$700 por MWh):

- Custo mensal: **R$1.2M**
- Custo Anual: **R$15.12M**

No BR, o supercomputador Santos Dumont teve de ser desativado justamente pelo alto custo com energia (cerca de 500k/mês).
https://revistagalileu.globo.com/amp/Tecnologia/noticia/2016/06/sem-dinheiro-para-conta-de-luz-supercomputador-brasileiro-e-desativado.html

Contexto

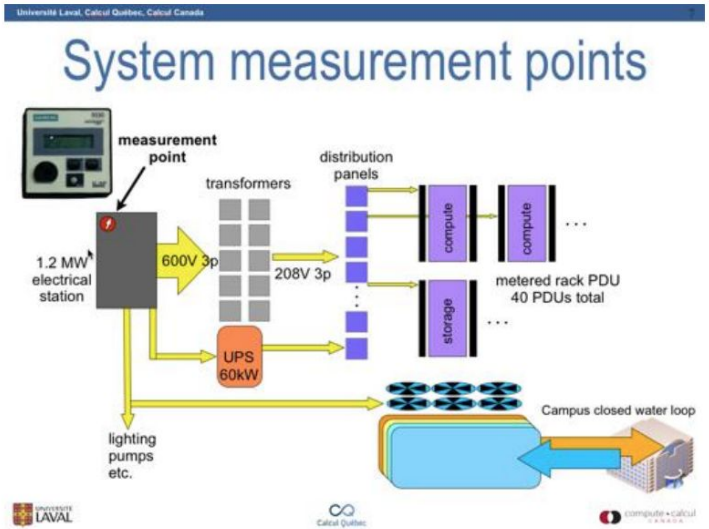
Top 10 Junho 2023

rank	top500 rank	name	country	year	architecture	processor	interconnect	rmax [tflop/s]	power (kw)	(tflop/s)/kw
1	255	Henri	United States	2022	Cluster	Intel Xeon Platinum 8362 32C 2.8GHz	Infiniband HDR	2882.0	44.07	65.3959609711822
2	34	Frontier TDS	United States	2021	MPP	AMD Optimized 3rd Generation EPYC 64C 2GHz	Slingshot-11	19200.0	306.3	62.68364348677767
3	12	Adastra	France	2022	MPP	AMD Optimized 3rd Generation EPYC 64C 2GHz	Slingshot-11	46100.0	791.78	58.22324383035692
4	17	Setonix – GPU	Australia	2022	MPP	AMD Optimized 3rd Generation EPYC 64C 2GHz	Slingshot-11	27160.0	476.63	56.98340431781466
5	77	Dardel GPU	Sweden	2022	MPP	AMD Optimized 3rd Generation EPYC 64C 2GHz	Slingshot-11	8259.0	146.2	56.49110807113544
6	1	Frontier	United States	2021	MPP	AMD Optimized 3rd Generation EPYC 64C 2GHz	Slingshot-11	1194000.0	22703.0	52.59216843588953
7	3	LUMI	Finland	2023	MPP	AMD Optimized 3rd Generation EPYC 64C 2GHz	Slingshot-11	309100.0	6015.77	51.38161864565966
8	483	amplitUDE (GPU Partition)	Germany	2023	Cluster	Xeon Platinum 8480+ 56C 2GHz	Infiniband NDR	1950.0	37.98	51.34281200631912
9	70	Goethe-NHR	Germany	2023	Cluster	AMD EPYC 7452 32C 2.35GHz	Mellanox InfiniBand EDR	9087.0	195.24	46.54271665642286
10	187	ATOS THX.A.B	France	2022	Cluster	Xeon Platinum 8358 32C 2.6GHz	Quad-rail NVIDIA HDR100 Infiniband	3505.0	84.64	41.41068052930056

https://www.top500.org/lists/green500/2023/06/ Acessado em 02/07/23

Contexto

- O ponto chave para o Top Green 500 é a obtenção da taxa (Tflop/s)/W, sendo que a potência utilizada é geralmente a potência média (ou demanda média) mensurada durante a execução do workload associado ao Linpack.



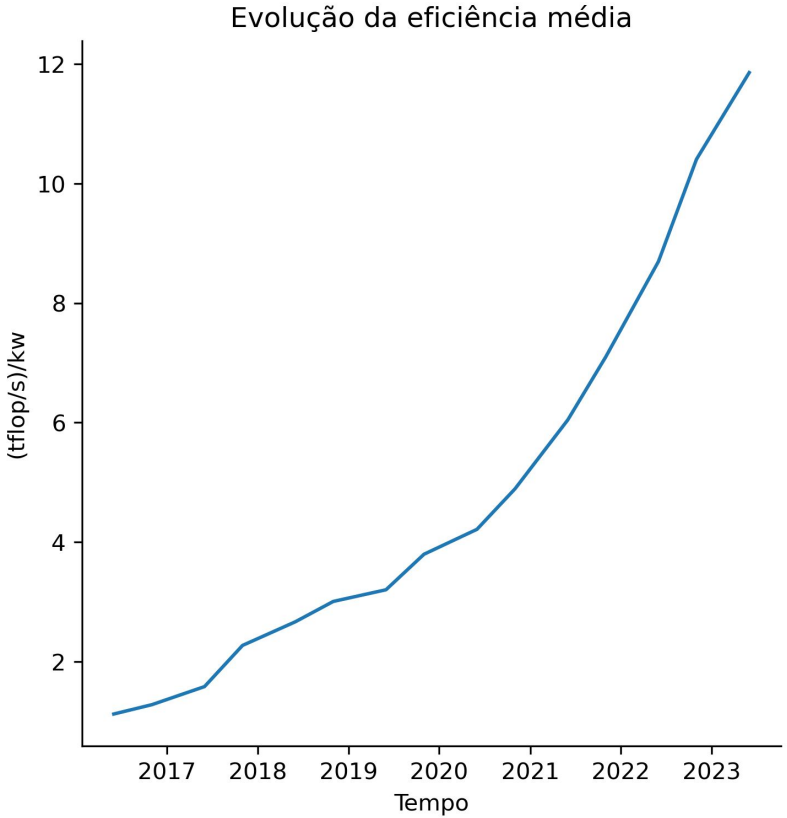
https://www.top500.org/static/media/uploads/methodology-2.0rc1.pdf Acessado em 02/07/23

Contexto

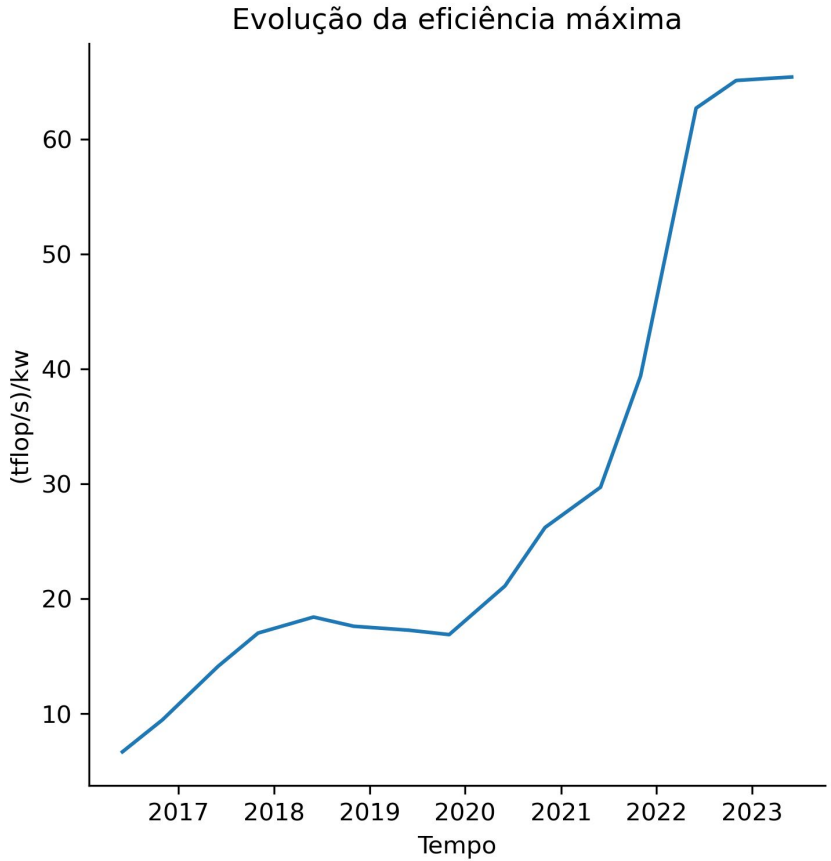
Algumas das **limitações** da lista Top 500 Green:

- A lista é baseada em única métrica (Tflops/W)
- Geralmente os dados de consumo são disponibilizados pelos próprios fornecedores
- O consumo avaliado é apenas da parte responsável pela computação, não considerando outras partes da infraestrutura ou sistemas de refrigeração, por exemplo.
- O workload avaliado é apenas o Linpack (mesmo problema da Top 500).

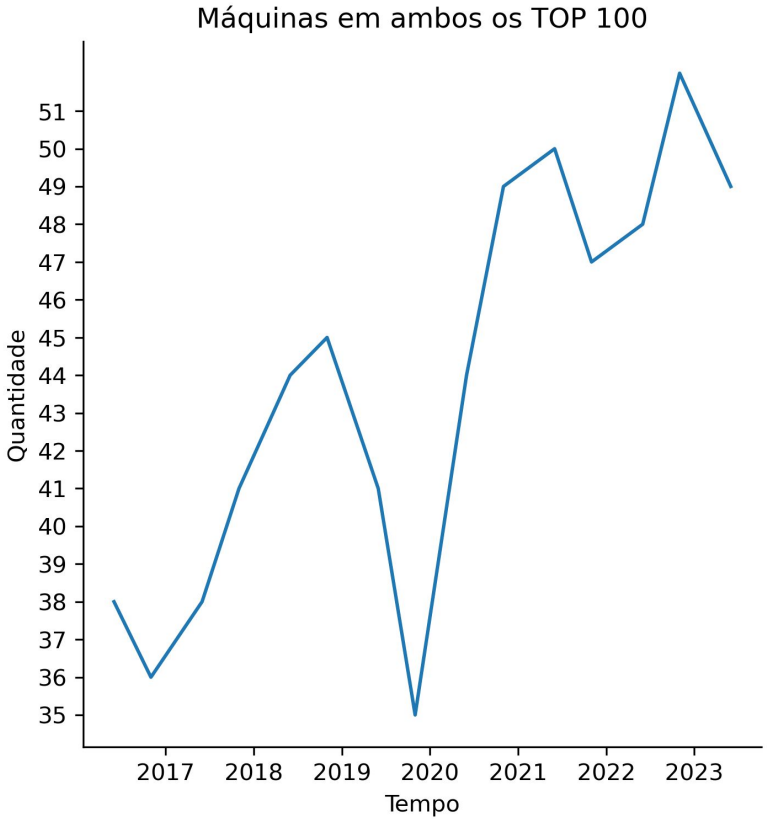
Análise Histórico Recente



Análise Histórico Recente



Análise Histórico Recente



Atual top 1 - Henri

Especificações:

- **Ano de fabricação**: 2022
- **Fabricante**: Lenovo
- **Site**: Flatiron Institute
- **Segmento**: Pesquisa
- **Arquitetura**: Cluster
- **Processador**: Intel Xeon Platinum 8362 32C 2.8GHz
- **Nº Cores**: 8288
- **Acelerador**: NVIDIA H100 PCIe (!)
- **Nº Cores no acelerador**: 7392
- **Interconnect**: Infiniband HDR
- **TFlop/s**: 2882.0
- **Potência (kW)**: 44.07

https://www.top500.org/lists/green500/2023/06/ Acessado em 02/07/23

Atual Top 1 - Henri

Pontos interessantes:

- Uso de acelerador NVIDIA H100 PCIe: nova geração de GPU, lançada em 2022. Tecnologia de fabricação TSMC 4N, desempenho cerca de 30x maior que a geração A100.
- O supercomputador ainda não está totalmente otimizado, tendo uma eficiência de ~37.6% no Linpack, ou seja, só atinge 37.6% da performance teoricamente possível.

https://www.hpcwire.com/2022/11/14/nvidias-h100-debuts-in-henri-topping-the-green500-list/ Acessado em 02/07/23

https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet Acessado em 02/07/23

Soluções para eficiência energética

Hardware and Software Solutions for Energy-Efficient Computing in Scientific Programming

Daniele D'Agostino,[1] Ivan Merelli,[2] Marco Aldinucci,[3] and Daniele Cesini[4]

Show more

Academic Editor: Cristian Mateos

Received	Accepted	Published
25 Jan 2021	28 May 2021	09 Jun 2021

Link: https://www.hindawi.com/journals/sp/2021/5514284/, acessado em 02/07/2023

Hardware - Técnicas de economia de energia

Exploram características de hardware para reduzir consumo de energia.

São implementadas no chip para controlar a frequência e a tensão dentro de um intervalo aceitável.

- Dynamic frequency scaling (DFS)
- Dynamic voltage scaling (DVS)

Modular o consumo de energia do processador e memória a partir da alteração do clock, de acordo com o uso.

Redução de tensão aplicada de forma diferente para dispositivos diferentes.

Hardware - Técnicas de economia de energia

- Dynamic voltage and frequency scaling (DVFS)

HW implementado com pequena camada de SW e pouco envolvimento do SO. Exemplo: CPUfreq subsystem do Linux

- Advanced Configuration and Power Interface (ACPI)

API para controle de potência via sistema operacional. Por exemplo, ACPI define até 16 estados, do P0 até P15, em que P0 (máxima tensão e frequência) e P15 a menor.

CPU performance scaling enables the operating system to scale the CPU frequency up or down in order to save power or improve performance. Scaling can be done automatically in response to system load, adjust itself in response to ACPI events, or be manually changed by user space programs.

Link: https://wiki.archlinux.org/title/CPU_frequency_scaling, acessado em 02/07/2023

Hardware - Técnicas de economia de energia

- Near-threshold voltage (NTV)

Diminuição da tensão operacional faz com que processadores operem em tensões muito baixas.

O sistema fica mais exposto a erros ocasionados por interferência de radiação atmosférica, gerando aumento da quantidade de erros de cálculo.

Assim, é necessário adicionar checagens e correções às aplicações.

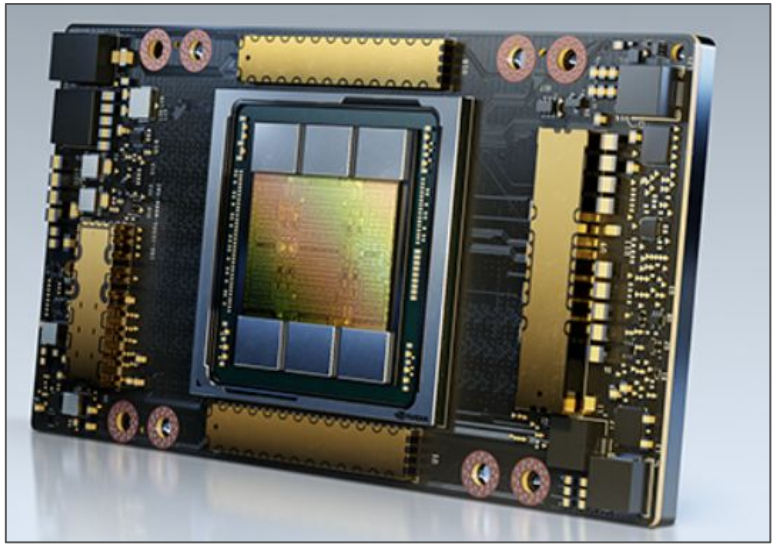
Hardware - Dispositivos específicos

GPU

Alta eficiência energética. Exemplos:

- MI100 (AMD) - 38,33 GFlops/W
- A100 (NVIDIA) - 24,25 GFlops/W

Necessitam de cuidados ao programar e otimizar para alcançar essa performance.



Link: https://www.nvidia.com/pt-br/data-center/a100/, acessado em 02/07/2023

Hardware - Dispositivos específicos

System on Chip (SoC)

Plataformas de HW com processadores de baixo consumo, originalmente projetados para embarcados

Aumento do poder computacional com baixo custo e baixo consumo energético

Exemplo: A64FX de 48 cores, com performance energética comparável a GPU.

Fugaku - 10º colocação Green500 (nov/2020), com 15.418 GFlops/W

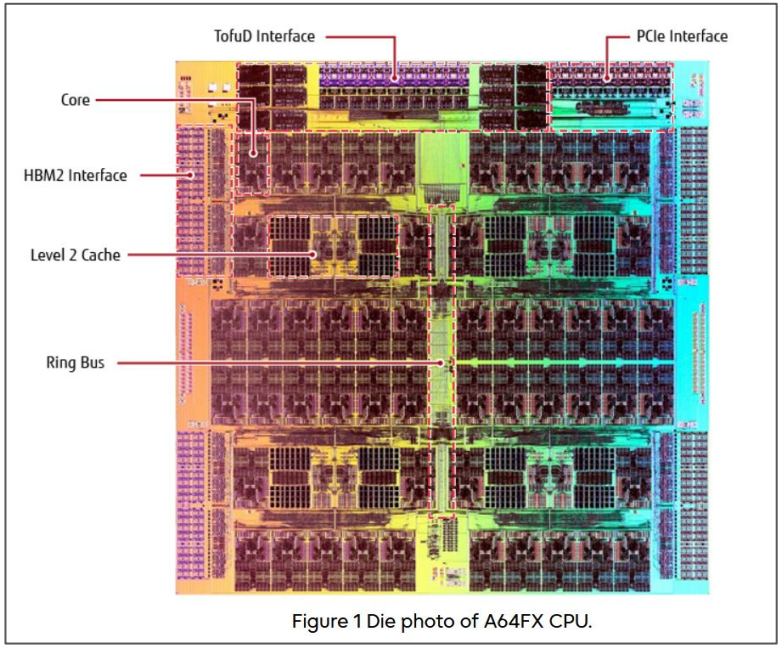


Figure 1 Die photo of A64FX CPU.

Link: https://www.fujitsu.com/global/about/resources/publications/technicalreview/2020-03/article03.html, acessado em 02/07/2023

Hardware - Dispositivos específicos

SoCs baseados em ARM ou FPGAs também são alternativas avaliadas

Ponto negativo: portabilidade de aplicações é complexa

ExaNoDe, ExaNeSt e Ecoscale

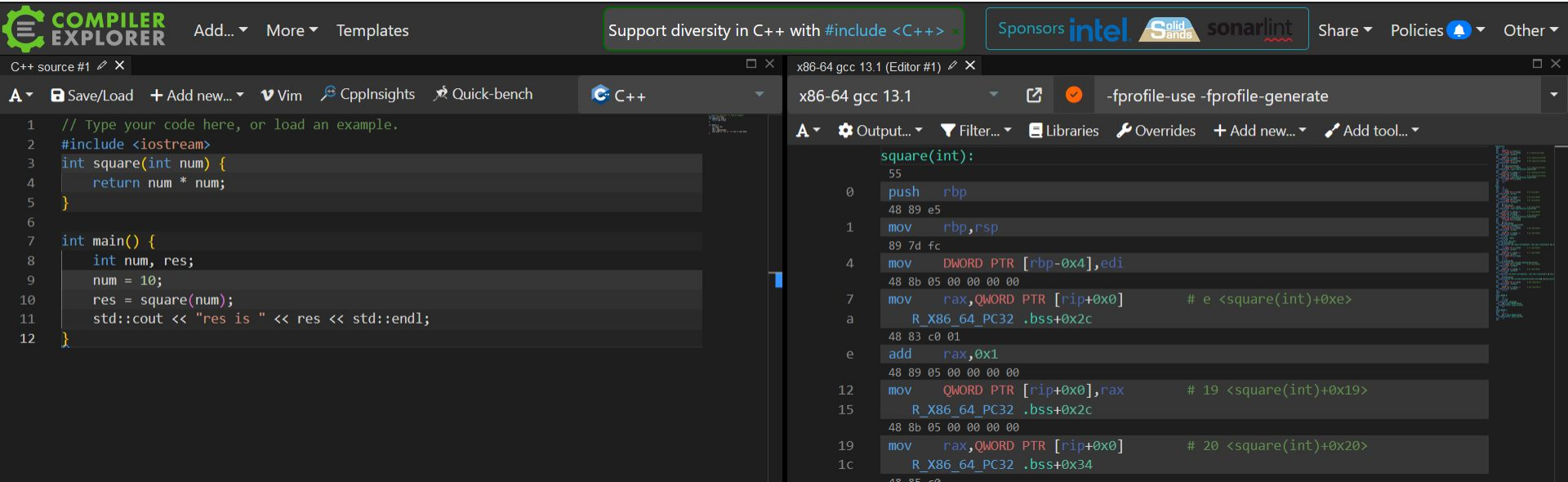
Computadores Exaflop de 50 GFlops/W, com processadores de arquitetura ARM, x86 e FPGA.

Esforço em criar modelos de programação unificados de diferentes arquiteturas de hardware como OpenCL

Software - Computação com Eficiência Energética

Linguagens de alto nível abstraem a camada de hardware. O mesmo código pode gerar instruções de hardware diferentes.

Compiler Explorer: gera assembly produzido por diferentes compiladores

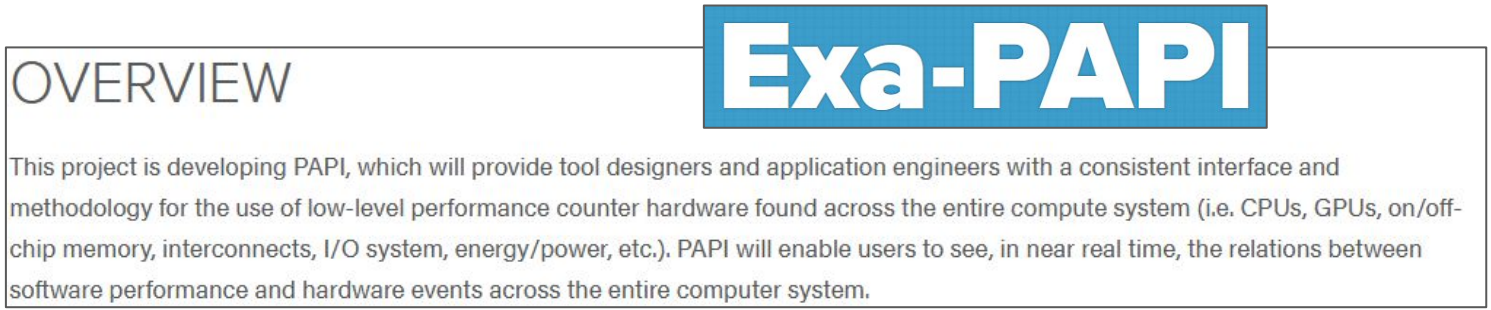


Software - Computação com Eficiência Energética

Profiling

Investigar o comportamento do software coletando informações da execução do programa

Necessário instrumentar a aplicação (no fonte ou no binário) utilizando a ferramenta profiler



Link: https://icl.utk.edu/exa-papi/, acessado em 02/07/2023

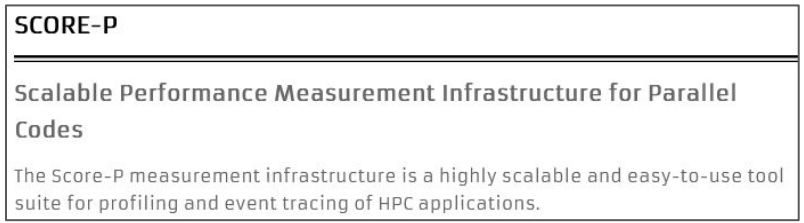
Software - Computação com Eficiência Energética

Profiling

PAPI, Performance API - independente de plataforma e provê interface e metodologia para coleta de dados de performance que o hardware disponibiliza.

PowerPack - conjunto de ferramentas que analisa a performance energética. Executado numa máquina apartada para diminuir a interferência.

Score-P - profiler, tracing de eventos e análise de aplicações HPC, permitindo análise comportamental de aplicações multi-processo, multi-thread ou baseada em acelerador.



Link: https://www.vi-hps.org/projects/score-p, acessado em 02/07/2023

Link: https://ieeexplore.ieee.org/abstract/document/4906989, acessado em 02/07/2023

Software - Computação com Eficiência Energética

Tuning Dinâmico

Ferramentas que visam controlar o consumo energético automaticamente, como o Global Extensible Open Power Manager (GEOPM), focado em HPC.

Configura o HW automaticamente dependendo da atividade e do gerenciador de recursos.

Link: https://geopm.github.io/, acessado em 02/07/2023

Software - Computação com Eficiência Energética

Soluções integradas - Frameworks que realizam profiling e tuning

READEX - Ferramenta instrumentada para auto-tuning visando eficiência energética, baseado em Score-P

LEGaTO - Ferramentas para eficiência energética para computação em hardware heterogêneo composto por CPU, GPU e FPGA. A instrumentação é feita por meio de OmpSs, uma extensão de OpenMP para HW heterogêneos.

Projetos financiadas pelo fundo de pesquisa europeu Horizon 2020

Desafios e perspectivas para o futuro

- Infraestrutura externa ao sistema de HPC é muitas vezes esquecida quando se fala de HPC
- Muito foco é dado para o desenvolvimento de hardware de alta performance e eficiência e de software que permita o gerenciamento dos recursos do sistema de forma eficiente
- No entanto, grandes avanços em sustentabilidade podem ser feitos focando na infraestrutura externa

Infraestrutura energeticamente eficiente

- Técnicas de resfriamento eficientes
  - Resfriamento é fundamental para dissipar energia produzida pelo sistema, permitindo que funcione corretamente
  - Grande parte da energia usada por um supercomputador é utilizada para resfriamento
  - Atualmente, grande parte dos supercomputadores já utilizam resfriamento líquido, que é mais eficiente em relação ao resfriamento por ar

Infraestrutura energeticamente eficiente

- Técnicas de resfriamento eficientes
  - Resfriamento a ar - mais simples, infraestrutura mais barata. Quase não é mais usada no contexto de HPC
  - Resfriamento líquido - técnica mais utilizada na computação de alto nível atualmente; baseia-se no uso de líquido em vez de ar para resfriar o sistema
  - Cada vez mais, computadores de maior performance precisam de técnicas de resfriamento adequadas, capazes de dissipar uma maior quantidade de calor de forma eficiente

Exemplo

Resfriamento por imersão

Exemplo

Resfriamento Thermosyphon

Link: https://www.hpcwire.com/2020/02/24/new-supercomputer-cooling-method-saves-half-million-gallons-of-water-at-sandia-national-laboratories

Infraestrutura energeticamente eficiente

- Diminuição de perda de energia na rede elétrica
- Utilização de fontes de energia sustentáveis
- Reutilização do calor dissipado pelo sistema

Exemplo

BNP Paribas

NEWS

BNP Paribas moves to Swedish datacentre in high-performance computing strategy

BNP Paribas is moving data-intensive workloads to a Swedish datacentre to increase capacity while reducing its carbon footprint

Link: https://www.computerweekly.com/news/365534046/BNP-Paribas-moves-to-Swedish-datacentre-in-high-performance-computing-strategy

Exemplo

Uso do calor dissipado para aquecer edifício

Supercomputer That Uses Waste Heat to Warm Its Building Wins R&D 100 Award

HP Supercomputer at NREL Garners Top Honor

Link: https://www.achrnews.com/articles/128397-supercomputer-that-uses-waste-heat-to-warm-its-building-wins-rd-100-award

Exemplo

Calor dissipado utilizado para aquecer piscina

Link: https://www.datacenterdynamics.com/en/news/uk-data-center-startup-offers-to-heat-britains-swimming-pools-with-waste-heat/