

A Supercomputação na Era da Inteligência Artificial

Alba C. M. A. Melo¹, Philippe O. A. Navaux², Lucia M. A. Drummond³,
Cristina Boeres³, Vinod Rebello³, Alfredo Goldman⁴, Márcio Castro⁵

¹Departamento de Ciência da Computação – Universidade de Brasília (UnB)
Brasília – Brasil

²Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – Brasil

³Instituto de Computação – Universidade Federal Fluminense (UFF)
Niterói – Brasil

⁴Instituto de Matemática e Estatística – Universidade de São Paulo (USP)
São Paulo – Brasil

⁵Depto. de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Florianópolis – Brasil

Abstract. *This paper discusses the main challenges for the evolution of supercomputers in the era of Artificial Intelligence (AI) over the next 10 years. It analyzes issues related to architecture, storage, integration with cloud computing, and energy sustainability. Considering the exponential growth trend in computational demand driven by AI models, the need for new approaches in hardware, software, and infrastructure to meet this demand is also highlighted.*

Resumo. *Este artigo discute os principais desafios para a evolução dos supercomputadores na era da Inteligência Artificial (IA) para os próximos 10 anos. São analisadas questões relacionadas à arquitetura, armazenamento, integração com a computação em nuvem e sustentabilidade energética. Considerando-se a tendência de crescimento exponencial da demanda computacional por modelos de Inteligência Artificial, destaca-se ainda a necessidade de novas abordagens em hardware, software e infraestrutura para atender a esta demanda.*

1. Introdução e Contextualização

Os supercomputadores são máquinas utilizadas para resolver problemas extremamente complexos, inviáveis de serem resolvidos por computadores de uso comum, pois exigem Computação de Alto Desempenho (*High Performance Computing* - HPC). São considerados supercomputadores as máquinas mais rápidas existentes em um dado momento no tempo [Bell 2015]. O Cray-1, lançado em 1976, foi um dos primeiros supercomputadores, o qual tinha capacidade de processamento vetorial, atingindo 250 milhões de operações de ponto flutuante por segundo (MFLOPS).

Desde então, temos observado um crescimento impressionante no desempenho dos supercomputadores, rompendo as barreiras de GFLOPS (10^9 FLOPS), TFLOPS (10^{12} FLOPS), PFLOPS (10^{15} FLOPS) e EFLOPS (10^{18} FLOPS) a aproximadamente cada 11

anos. Em Novembro/2024, o *El Capitan*, contendo nodos de computação compostos por CPUs e GPUs (aceleradores), localizado nos Estados Unidos, é o supercomputador listado como mais rápido do mundo. Combinando processadores AMD EPYC de 24 núcleos (4ª geração) e GPUs AMD Instinct MI300A (11.039.616 núcleos de processamento no total), este supercomputador é capaz atingir 1,7 EFLOPS no *High-Performance Linpack* benchmark usado para classificar os sistemas. Considerando-se a taxa de evolução dos computadores recente, espera-se portanto que ultrapassemos a barreira dos ZFLOPS (10^{21} FLOPS) na década de 2030 [Matsuoka et al. 2023].

Tradicionalmente, os supercomputadores tem sido utilizados de forma efetiva para resolver problemas complexos, tais como dobramento acurado de proteínas [Streit et al. 2024], simulação de materiais para eletrônica de alto desempenho [Lyu et al. 2024], clima e previsão do tempo [Govett et al. 2024], simulação de energias renováveis [Veers et al. 2023], dentre outros. Mas, na presente década, temos observado um crescimento sem precedentes da Inteligência Artificial. Isso se dá principalmente devido à disponibilidade de conjuntos imensos de dados, que possibilitam respostas acuradas para modelos com uma quantidade enorme de parâmetros baseados em Redes Neurais do tipo *transformer* [Guo et al. 2019]. Tais arquiteturas deram origem à área de Inteligência Artificial Generativa e diversas soluções existentes, tais como o ChatGPT (OpenAI), são utilizadas milhões de vezes a cada dia ao redor do mundo. O treinamento de modelos com 1 trilhão de parâmetros (GPT-5) requer supercomputadores, com alta capacidade de poder de processamento, memória, interconexão e sistema de armazenamento em disco.

As empresas da área de Inteligência Artificial estão trabalhando para produzirem supercomputadores mais adequados a esse tipo de processamento. A empresa *Cerebras*, por exemplo, desenvolveu o *Cerebras Wafer-Scale Cluster*, com *design* otimizado, desacoplando memória e processamento, com acesso remoto à memória, interconexão de rede com nova topologia, dentre outros recursos. Com isso, foi lançado no final de 2023 o supercomputador Condor Galaxy 1, com capacidade de processamento de 4 EFLOPS (meia precisão) [Lie 2023] e o 8-EFLOPS Galaxy 3 já está em desenvolvimento.

Os próximos Modelos de Linguagem de Grande Escala (*Large Language Models* - LLMs) necessitarão de ainda mais os avanços no projeto de supercomputadores e, além de processamento, memória, interconexão, sistema de armazenamento secundário, o consumo de energia deverá ser considerado. Sendo assim, novas arquiteturas serão necessárias para que os supercomputadores do futuro possam processar tais modelos sem aumentar significativamente o consumo de energia. O consumo de energia associado a um supercomputador é enorme, sendo que quatro dos equipamentos entre os mais rápidos do mundo tem um consumo de mais de 10 MegaWatts, o que corresponde a cerca de três vezes o consumo de uma cidade como Campinas ou Guarulhos¹.

Para abordar o desafio da supercomputação na era da IA, devemos determinar uma maneira viável e sustentável de se ultrapassar a barreira dos ZFLOPS na década de 2030. As métricas objetivas e claras para avaliar o progresso das soluções são desempenho em FLOPS e a eficiência energética (GFLOPS/Watts) atingidas por cada supercomputador.

¹https://dadosenergeticos.energia.sp.gov.br/PortalCEV2/Municipios/Eletricidade/M_Eletricidade.asp

2. Desafios para os Próximos 10 Anos

Nesta seção, partiremos dos supercomputadores mais avançados atualmente (1 a 4 EFLOPS) e discutiremos as principais questões para endereçarmos o desafio de supercomputadores na era da IA para os próximos dez anos.

2.1. Qual Deverá ser a Arquitetura do Supercomputador do Futuro?

Serão necessários grandes avanços em nível arquitetural para ultrapassar a barreira dos ZFLOPS, tais como: novas tecnologias de interconexão entre os nodos, novas tecnologias para resfriamento de supercomputadores de maneira efetiva e mais sustentável, novos mecanismos em software e hardware para tratar a falha de componentes, entre outros. Defendido por vencedores do Prêmio Turing [Hennessy and Patterson 2019], uma alternativa é projetar arquiteturas adaptadas a um domínio de problema específico. Arquiteturas específicas de domínio (DSAs) podem fornecer ganhos significativos de desempenho (e eficiência) e, dado que os acessos à memória se tornaram muito mais caros do que os cálculos aritméticos, podem ajudar fazer uso mais eficaz da hierarquia de memória. Portanto, supercomputadores do futuro deverão possuir tanto arquiteturas convencionais (processadores de propósito geral e aceleradores, incluindo aceleradores específicos para IA), como arquiteturas alternativas do tipo *Processing in Memory* (PIM). Há um interesse crescente em plataformas de hardware que implementem diferentes arquiteturas (ISAs) alternativas com desempenhos superiores por watt, por exemplo, a ISA aberta proposta pela iniciativa RISC-V² (da qual o Brasil faz parte), e que por ser aberta evita a dependência do fabricante do *chip*. Em termos de montagem, os fabricantes esperam que *chipllets* permitirão que eles abordem melhor três dos aspectos de eficiência computacional: energia computacional, energia de comunicações e energia de memória. Em um futuro próximo, a computação quântica deverá se aliar à computação de alto desempenho para, em conjunto, viabilizar a ultrapassagem da barreira dos ZFLOPS.

2.2. Como Será o Armazenamento dos Dados na Era da Supercomputação para IA?

O volume de dados utilizados em algoritmos de IA em larga escala é imenso. O primeiro desafio a ser superado é como armazenar, de maneira compacta e eficiente, volumes de dados na ordem de PetaBytes em constante (e rápido) crescimento. Técnicas de compressão de dados e de redução de dimensionalidade devem ser consideradas, dentre outras. Tendo os dados armazenados, o segundo desafio é recuperá-los de maneira rápida para processamento. Ou seja, deve ser definido como será a interconexão entre os nodos e com o sistema de arquivos e também como se dará o processamento (e.g., processamento com o dado comprimido).

2.3. Como Será a Interação entre Computação em Nuvem e Computação de Alto Desempenho na Era da IA?

Provedores de nuvem têm oferecido arquiteturas especializadas para computação paralela, tais como GPUs, FPGAs e aceleradores específicos para IA (e.g., TPUs da Google e Trainium e Inferentia da AWS e o MAIA da Microsoft) [Reed et al. 2023]. Portanto, a infraestrutura de computação em nuvem tem sido vista como uma forma de estender

²<https://riscv.org/>

uma infraestrutura de supercomputação instalada localmente. Nesse contexto, há diversos desafios relacionados a interoperabilidade entre nuvem e supercomputadores, o compartilhamento de recursos entre diversos usuários na nuvem e a dependência dos pesquisadores em certos serviços oferecidos por provedores de nuvem. Novas estratégias que permitam a redução do custo financeiro no uso de infraestruturas de nuvem também precisarão ser desenvolvidas.

2.4. Como Tratar o Problema de Consumo de Energia e da Sustentabilidade de Maneira Transversal?

Para sustentar os avanços na IA, são necessárias medidas urgentes para resolver o conjunto de gargalos enfrentados pela computação de alto desempenho. Os supercomputadores modernos são construídos usando dezenas de milhares de aceleradores, mas as limitações nas abordagens usadas para agregá-los significam que o desempenho próximo do pico raramente, ou nunca, é alcançado na prática. Isto implica que a quantidade de energia necessária para obter os ganhos de desempenho do ZFLOPS provavelmente excederá proporcionalmente os dos avanços anteriores. Se o desenvolvimento tecnológico continuar em sua trajetória atual, é estimado que um supercomputador da classe ZFLOPS precisaria em torno de 500 megawatts de potência³, pouco menos da quantidade de energia gerada por uma das 20 turbinas da barragem de Itaipu e comparável ao consumo atual dos centros de dados de tamanho médio. A sede, aparentemente insaciável, de poder da IA ganhou atenção considerável dos provedores de nuvem. Num esforço para encontrar fontes de energia para novos centros de dados, os fornecedores de nuvens estão atualmente recorrendo à energia nuclear⁴. Embora não produza poluição do ar (ou dióxido de carbono durante sua operação), esta solução levanta a questão do tratamento de resíduos radioativos. Com os centros de dados de nuvem hoje consumindo de 2% a 3% de toda a energia utilizada, espera-se que o crescimento contínuo da IA faça com que este valor duplique. Isto está causando preocupação aos governos. Por exemplo, a União Europeia (UE) estabeleceu uma meta para reduzir o consumo de energia na Europa em 11,7% até 2030. Em relação à previsão para 2030 feita em 2020, ela está exigindo que as organizações que operam centros de dados nos países da UE apresentem relatórios anuais detalhando o consumo de água e de energia, bem como as medidas tomadas para sua redução⁵. Durante a próxima década, a eficiência computacional se tornará a prioridade número um. O desafio é descobrir como melhorá-lo significativamente numa forma economicamente e ambientalmente sustentável.

3. CVs dos Proponentes

Alba Cristina Magalhaes Alves de Melo é Professora Titular do Departamento de Ciência da Computação da UnB e atualmente é Chefe do Departamento. Atua na área de supercomputação e computação de alto desempenho desde 1996, quando concluiu seu doutorado no INPG, França. A Prof. Alba é pesquisadora Nível 1C do CNPq, Conselheira da SBC e Membro Senior da Sociedade IEEE. É editora associada de vários periódicos de prestígio e organizou diversas conferências internacionais na área de computação de alto desempenho. Coordenou também projetos de cooperação

³<https://www.nextbigfuture.com/2023/02/intel-and-amd-path-to-zettaflop-supercomputers.html>

⁴<https://https://www.datacenterknowledge.com/energy-power-supply/going-nuclear-a-guide-to-smrs-and-nuclear-powered-data-centers>

⁵<https://www.cio.com/article/2100517/eu-moves-toward-regulating-data-center-energy-and-water-use.html>

com o Canadá e Espanha e participou/participa de projetos com a Comunidade Européia, França e Alemanha.

Alfredo Goldman é professor da Universidade de São Paulo e bolsista de produtividade do CNPq. Atualmente é conselheiro da Sociedade Brasileira de Computação e coordenador da comunidade de processamento paralelo na *IEEE Computer Society* (TCPP). Possui doutorado em Informática e Sistemas pelo Instituto Nacional Politécnico de Grenoble. Trabalha com programação paralela e distribuída e desenvolvimento ágil de software. Coordena um projeto temático da FAPESP.

Cristina Boeres é professora do Instituto de Computação da UFF, e PhD em Ciência da Computação pela Universidade de Edimburgo (Reino Unido). Realiza pesquisas em HPC e gerenciamento de execução em ambientes paralelos e distribuídos. Tem participado de projetos financiados por CNPq, Petrobras e Faperj e coordena projeto CNPq/AWS.

Lúcia M. A. Drummond é Professora Titular da Universidade Federal Fluminense. É Pesquisadora Nível 1 do CNPq. É Cientista do Estado pela FAPERJ. Foi coordenadora da Comissão Especial de Arquitetura de Computadores e Processamento de Alto Desempenho da Sociedade Brasileira de Computação de 2019 a 2023. Atualmente é coordenadora de projetos financiados pela FAPERJ, CNPQ, CAPES e PETROBRAS na área de Computação de Alto Desempenho, envolvendo pesquisadores das Universidades de Sorbonne e Bordeaux (INRIA).

Márcio Castro é professor da Universidade Federal de Santa Catarina (UFSC) e PhD em Ciência da Computação pela *Université Grenoble Alpes* (França). É pesquisador Nível 2 do CNPq, coordenador da CE-ACPAD da SBC e do Programa de Pós-Graduação em Ciência da Computação da UFSC. Realiza pesquisas em HPC e Cloud em projetos financiados pelo CNPq, CAPES, FAPESC e MPSC.

Philippe O. A. Navaux é professor emérito do Instituto de Informática da UFRGS, Doutorado em Computação, INPG, 1979. Pesquisador nível 1 do CNPq nas áreas de Arquitetura de Computadores e HPC. Orientou mais de 100 estudantes de doutorado e mestrado e publicou mais de 400 artigos em periódicos e conferências. Sócio da SBC, SBPC, ACM e IEEE. Foi coordenador do Comitê da Computação da Capes, CNPq e Fapergs. Foi conselheiro da UFRGS, da Fapergs, do CATI/ MCTI, do LNCC. Atualmente membro do conselho do CIEE-RS e presidente do Conselho do SCALAC.

Vinod Rebello é Professor Associado da Universidade Federal Fluminense, tem um PhD em Ciência da Computação pela Universidade de Edimburgo (Reino Unido), e realiza pesquisas na área de HPC, nuvens e sistemas paralelos e distribuídos. Tem participado em várias projetos financiados por agências nacionais e internacionais e na organização de conferências da área.

Referências

- Bell, G. (2015). Supercomputers: The amazing race. Microsoft Technical Report MSR-TR-2015-2, Microsoft Research.
- Govett, M., Bah, B., Bauer, P., et al. (2024). Exascale computing and data handling: Challenges and opportunities for weather and climate prediction. *Bulletin of the American Meteorologic Society*, early access.
- Guo, Y., Zheng, Y., Tan, M., Chen, Q., Chen, J., Zhao, P., and Huang, J. (2019). NAT: Neural architecture transformer for accurate and compact architectures. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, volume 32, pages 1–12. Curran Associates, Inc.

- Hennessy, J. L. and Patterson, D. A. (2019). A new golden age for computer architecture. *Commun. ACM*, 62(2):48–60.
- Lie, S. (2023). Inside the cerebras wafer-scale cluster: Cerebras systems. In *IEEE Hot Chips Symposium (HCS)*, volume 35, pages 1–41. IEEE.
- Lyu, B., Chen, J., Wang, S., et al. (2024). Graphene nanoribbons grown in hbn stacks for high-performance electronics. *Nature*, 628:758–764.
- Matsuoka, S., Domkel, J., Wahib, M., Drozd, A., and Hoefler, T. (2023). Myths and legends in high-performance computing. *Int. Journal of High Performance Computing and Applications*, 37(3-4):245–259.
- Reed, D., Gannon, D., and Dongarra, J. (2023). HPC Forecast: Cloudy and uncertain. *Commun. ACM*, 66(2):82–90.
- Streit, J., Bukvin, I., Chan, S., et al. (2024). The ribosome lowers the entropic penalty of protein folding. *Nature*, 633:232–239.
- Veers, P., Bottasso, C. L., Manuel, L., et al. (2023). Grand challenges in the design, manufacture, and operation of future wind turbine systems. *Wind Energy Science*, 8(7):1071–1131.