Uma Abordagem do Uso de Risc-V em Computação de Alto Desempenho (HPC)

Leon Barboza[1], **Marcondes Gorgonho**[1], **Evaldo Costa**[1]

[1]Centro de Estudos e Sistemas Avançados do Recife (CESAR)
Recife – PE – Brasil

{lmub,msg,ebc}@cesar.org.br

Abstract. *Architectures for use in high-performance computing (HPC) are under constant development, aiming to process large volumes of data through more efficient computational resources. Among these architectures is RISC-V, whose main goals are openness and low energy consumption. In this article, we present an approach to the use of RISC-V in high-performance computing, highlighting its main characteristics, applications, and current developments.*

Resumo. *Arquiteturas para uso em computação de alto desempenho (HPC) estão em constantes desenvolvimentos, buscando o processamento de grandes quantidades de dados com o uso de recursos computacionais mais eficientes. Entre essas arquiteturas está o Risc-V que tem por principal objetivo o uso de uma arquitetura aberta e baixo consumo de energia. Neste artigo apresentaremos uma abordagem do uso de Risc-V em computação de alto desempenho, apresentando suas principais características, aplicações e o seu desenvolvimento atual.*

1. Introdução

A Computação de Alto Desempenho atravessa uma fase de transformação na era da exaescala, com novos desafios em eficiência energética e sustentabilidade. Neste cenário, a arquitetura aberta RISC-V surge como uma força disruptiva. Após 15 anos de desenvolvimento, desde sua criação em 2010, o RISC-V consolidou-se no mercado de sistemas embarcados e agora, com seus componentes de software e ISA fundamentais já estabelecidos, avança para setores da indústria que demandam maior desempenho, como servidores, Inteligência Artificial e HPC.

A chegada de processadores de alto desempenho, como o SOPHON SG2042 e o mais recente SG2044, demonstra a viabilidade do RISC-V para cargas de trabalho de servidor e computação intensiva. Além disso, o RISC-V estabeleceu-se como a base padrão para novos aceleradores de IA. As linhas de placas PCI Express Blackhole e Wormhole, da fabricante Tenstorrent, exemplificam essa tendência.

Este artigo explora o panorama atual e as aplicações do RISC-V em HPC, analisando como a maturidade da arquitetura, o surgimento de hardware competitivo e um ecossistema de software maduro o posicionam como uma possível alternativa para a supercomputação.

2. Trabalhos Relacionados

Em [Brown et al. 2023], os autores realizaram as primeiras avaliações abrangentes do processador Sophon SG2042 utilizando a suite de benchmarks RAJAPerf. Eles concluíram que, apesar de um salto de desempenho em relação a gerações anteriores, o hardware sofria com gargalos de memória e um ecossistema de software imaturo, destacando a necessidade de otimizações manuais em bibliotecas e o uso de compiladores não padronizados. O trabalho subsequente em [Brown 2025] demonstrou um avanço significativo com a CPU Sophon SG2044, que, ao corrigir os gargalos de memória e adotar a última ratificação das extensões vetoriais RVV 1.0, reduziu drasticamente a lacuna de desempenho em relação a processadores x86 e Arm nos algoritmos presentes no conjunto de benchmarks NPB.

Em [Diehl et al. 2024], a aplicação de astrofísica Octo-Tiger foi portada para o SG2042, onde o nó RISC-V se mostrou competitivo em desempenho e superior em eficiência energética em comparação com o processador Arm A64FX, embora o estudo tenha exposto limitações no ecossistema de software.

Paralelamente, a exploração de aceleradores RISC-V tem sido outra área de foco. Os trabalhos em [Brown and Barton 2024] e [Brown et al. 2025] investigam o desempenho dos aceleradores Tenstorrent Grayskull e Wormhole para kernels de HPC, como algoritmos de stencil e FFT. Uma conclusão recorrente é a alta eficiência energética desses dispositivos, que chegam a consumir até oito vezes menos energia que CPUs de servidor, mas o desempenho é criticamente dependente da otimização do movimento de dados entre a memória do host e a SRAM dos núcleos.

3. Aplicações Risc-V HPC

Cluster Monte Cimone

Desenvolvido na Universidade de Bolonha em parceria com o CINECA, o projeto Monte Cimone [Venieri et al. 2025] destaca-se como um ambiente experimental para avaliação do uso de processadores RISC-V comercialmente disponíveis em computação de alto desempenho. Em sua primeira versão (MCv1), o sistema era composto por oito nós baseados no SBC HiFive Unmatched da SiFive, equipado com o SoC Freedom U740. Já a segunda versão (MCv2) passou a utilizar três nós Milk-V Pioneer Box, cada um com o SoC Sophgo SG2042 de 64 núcleos Xuantie C920, além de um nó dual-socket com dois processadores SG2042. Os novos nós foram acoplados ao sistema Monte Cimone usando a rede de 1 Gb/s existente como uma partição adicional do SLURM e foram configurados via Spack e integrados ao sistema de monitoramento ExaMon.

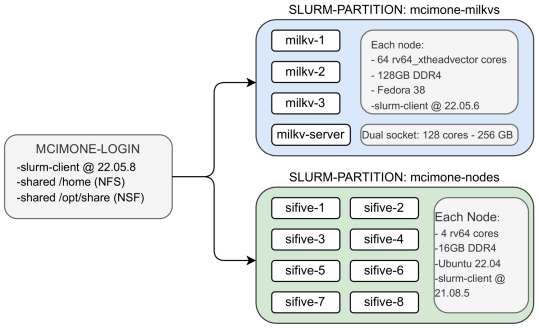

Figura 1. Monte Cimone V1 + V2

Os nós do MCv2 utilizam o Fedora 38 como sistema operacional, junto à toolchain mainline do GCC 13. Além disso, os autores do experimento disponibilizaram como shared libraries as toolchains Xuantie GNU e GCC 14 para fins de compatibilidade com as unidades vetoriais dos núcleos Xuantie C920 presentes nos processadores SG2042.

Para avaliar a nova versão do cluster, foram utilizados os benchmarks STREAM para largura de banda e HPL para avaliação das capacidades computacionais. Os resultados obtidos no MCv2 evidenciam ganhos expressivos com a nova linha deprocessadores, cerca de 127× em desempenho computacional, medido em GFlops/s, e aproximadamente 69× em largura de banda de memória quando comparado ao MCv1.

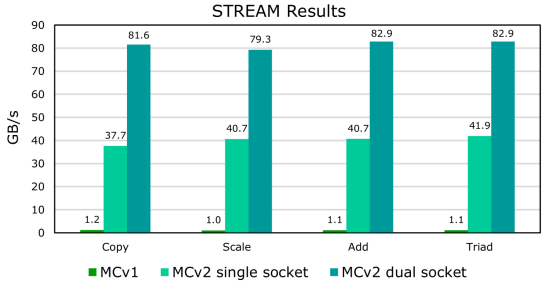

Figura 2. Resultado STREAM

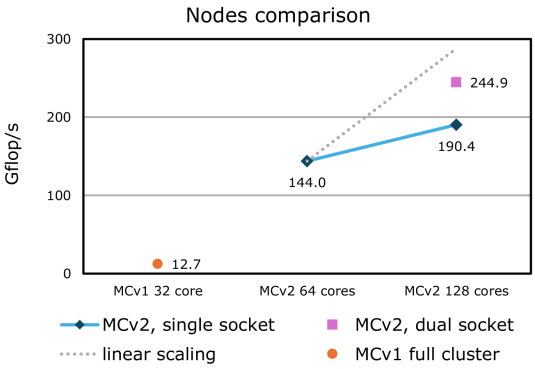

Figura 3. Resultado HPL

Como destacado anteriormente, o experimento exigiu diferentes toolchains por questões de compatibilidade. A BLAS utilizada pelo HPL precisou ser compilada com um fork do GCC fornecido pela Xuantie. O SoC SG2042 implementa a versão 0.7.1 das extensões vetoriais RVV, que não possui suporte no GCC padrão. Conforme descrito em [Brown 2025], a nova geração SG2044 já adota a ratificação final 1.0 do RVV, permitindo o uso de compiladores mainline como GCC e LLVM. Isso demonstra que o ecossistema de software RISC-V está em acelerada expansão e que as demandas da comunidade vêm sendo atendidas pelos fabricantes.

Aceleração para IA

No mercado de aceleradores voltados à inteligência artificial, a empresa americana Tenstorrent emergiu como uma grande competidora, oferecendo uma ampla gama de produtos: placas PCI Express, workstations, servidores, soluções em computação em nuvem e uma pilha de software já bastante madura, voltada ao desenvolvimento de aplicações para seus aceleradores.

A linha de aceleradores da empresa é baseada em uma mesma arquitetura. As placas Grayskull e75 e Grayskull e150, que serão analisadas em maior detalhe, contam com 96 e 120 núcleos Tensix, respectivamente, canais de memória posicionados na parte superior e inferior da malha de processadores, oferecendo uma capacidade total de 8 GB (LPDDR4). Enquanto a e75 apresenta uma largura de banda de 102,4 GB/s, a e150 atinge 118,4 GB/s. Operando a 1 GHz, a e75 alcança um desempenho de pico de 55 TFLOPs em FP16, ao passo que a e150, operando a até 1,2 GHz e oferece 92 TFLOPs em FP16. As gerações mais recentes expandem significativamente essa capacidade, a linha Wormhole

n300 chega a 128 núcleos, enquanto a Blackhole p150 atinge 140 núcleos Tensix, ambas contando com muito mais memória, 24 GB (GDDR6) e 32 GB (GDDR6).

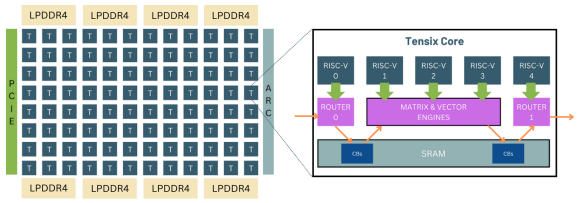

Figura 4. Arquitetura Tensix

Cada núcleo Tensix é formado por cinco núcleos RISC-V, 1 MB de memória SRAM local (L1), uma unidade SIMD de processamento vetorial e matricial e roteadores NoC (Network on Chip). Os núcleos RISC-V 0 e RISC-V 4 executam kernels de movimentação de dados, coordenando leituras e escritas assíncronas entre as SRAMs dos demais núcleos, os bancos de DRAM externos e a SRAM local. Já os três núcleos centrais interagem diretamente com as unidades de processamento SIMD: o primeiro realiza o desempacotamento dos dados, o segundo executa os kernels de computação e o terceiro cuida do empacotamento dos resultados. A comunicação entre os núcleos ocorre por meio de buffers circulares (CB) na SRAM, que armazenam dados provenientes das memórias externas e de resultados intermediários.

No estudo realizado em [Cavagna et al. 2025] sobre multiplicação matricial, operação essencial para cargas de trabalho de IA, o Grayskull demonstrou um equilíbrio competitivo entre desempenho e eficiência energética. Embora GPUs como a NVIDIA A100 ainda dominem em termos de FLOPs brutos, o Grayskull apresentou a maior eficiência energética entre as plataformas avaliadas, atingindo 1,56 TFLOPs/Watt. O uso otimizado da memória L1 (SRAM) de forma distribuída revelou ganhos expressivos, especialmente quando os blocos de matrizes se ajustam à capacidade local dos núcleos.

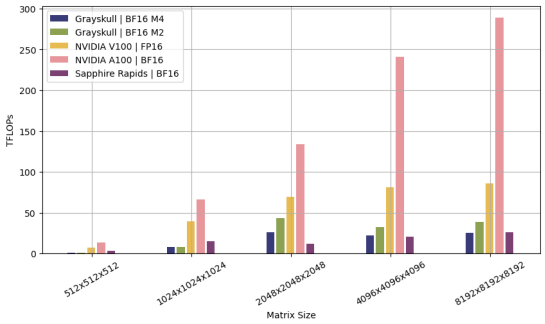

Figura 5. Performance

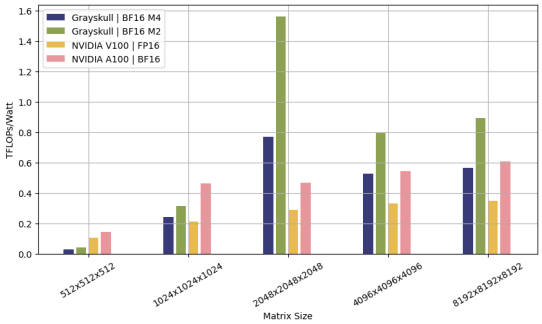

Figura 6. Eficiência Energética

Já em [Thüning 2024], é ressaltado o grande potencial de aceleração de kernels de atenção proporcionado pela ampla quantidade de SRAM disponível na arquitetura Tensix. A Grayskull e150 e oferece aproximadamente 1,5 vez mais SRAM que uma NVIDIA H100 PCIe. Especificamente, a Grayskull e150 integra 120 MB de SRAM, 120 núcleos Tensix com 1 MB cada, enquanto a Nvidia H100 PCIe possui apenas 80 MB no total, 114 Streaming Multiprocessors com 256 KB de L1 cada e 50 MB de L2 compartilhada. Além disso, considerando a largura de banda de 384 GB/s por núcleo, o acesso paralelo dos 120 núcleos Tensix resulta em uma largura de banda agregada de aproximadamente 46 TB/s, em contraste com os cerca de 6 TB/s de leitura em L2 da Nvidia H100 PCIe.

4. Conclusões e Trabalhos Futuros

Os avanços recentes no uso de RISC-V em HPC evidenciam o potencial da arquitetura, mas também apontam desafios que demandam investigações futuras. É fundamental aprofundar os estudos sobre o desempenho e a escalabilidade em clusters distribuídos, como o Monte Cimone, para avaliar a interoperabilidade e a eficiência da comunicação em larga escala. Além disso, é essencial avaliar hardware de última geração, utilizando novos benchmarks para os processadores mais recentes da Sophgo e para aceleradores como o Blackhole, da Tenstorrent, e o ET-SoC-1, da Esperanto.

Em síntese, a consolidação do RISC-V em HPC dependerá do amadurecimento conjunto de hardware, software e interconexões, além da exploração de cenários heterogêneos e distribuídos. Esses esforços têm o potencial de posicionar a arquitetura como uma alternativa competitiva frente a outras arquiteturas já consolidadas no domínio da computação de alto desempenho.

Referências

Brown, N. (2025). Is risc-v ready for high performance computing? an evaluation of the sophon sg2044. *arXiv preprint arXiv:2508.13840*. Preprint of paper submitted to RISC-V for HPC SC25 workshop.

Brown, N. and Barton, R. (2024). Accelerating stencils on the tenstorrent grayskull risc-v accelerator. In *SC-W '24 Workshops on HPC*, pages 1690–1700.

Brown, N., Davies, J., and LeClair, F. (2025). Exploring fast fourier transforms on the tenstorrent wormhole. *arXiv preprint arXiv:2506.15437*. Author accepted version, submitted to RISC-V for HPC ISC workshop 2025.

Brown, N., Jamieson, M., Lee, J., and Wang, P. (2023). Is risc-v ready for hpc prime-time: Evaluating the 64-core sophon sg2042 risc-v cpu. In *SC-W '23: Proceedings of the SC '23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*, pages 1566–1574.

Cavagna, H. P., Cesarini, D., and Bartolini, A. (2025). Assessing tenstorrent's risc-v matmul acceleration capabilities. *arXiv preprint arXiv:2505.06085*. Accepted to the Computational Aspects of Deep Learning Workshop at ISC High Performance 2025.

Diehl, P., Syskakis, P., Daiß, G., Brandt, S. R., Kheirkhahan, A., Singanaboina, S. Y., Marcello, D., Taylor, C., Leidel, J., and Kaiser, H. (2024). Preparing for hpc on risc-v: Examining vectorization and distributed performance of an astrophysics application with hpx and kokkos. In *SC-W '24 Workshops on HPC*, pages 1656–1665. Published Feb. 2025.

Thüning, M. (2024). Attention in sram on tenstorrent grayskull. *arXiv preprint arXiv:2407.13885*. 8 pages, 6 figures, code available online.

Venieri, E., Manoni, S., Madella, G., Ficarelli, F., Gregori, D., Cesarini, D., Benini, L., and Bartolini, A. (2025). Monte cimone v2: Down the road of risc-v high-performance computers. *arXiv preprint arXiv:2503.18543*. Accepted at RISC-V Summit Europe 2025.