Accelerating Genome Sequencing With HPC

Shambhawi Sharma

Introduction

- High-throughput sequencing has revolutionized genomic research, enabling large-scale studies.
- Decreasing costs and increasing data volumes create computational challenges in alignment, variant calling, and data storage.
- HPC is essential, incorporating GPUs, FPGAs, and TPUs to accelerate genomic analysis.
- Innovations like ERT, mixed-precision computing, and SeqBench enhance efficiency and scalability.

Problem Statement

- Current frameworks face inefficiencies in data indexing, seeding, variant calling, and compression.
- Alignment bottlenecks persist due to memory bandwidth constraints of compressed index structures.
- GWAS struggles with the computational overhead of nonlinear models.
- Variant calling scalability is limited by CPU-dominated workflows.
- The report explores integrating HPC principles for a unified genomic pipeline.

Accelerating Read Alignment

- Read alignment is computationally intensive, with seeding being a bottleneck (40% of alignment time).
- Enumerated Radix Trees (ERT) improve data access speed by reducing memory retrievals.
- FPGA-based accelerator boosts seeding (3.3x) and overall alignment performance (2.1x).

Mixed-Precision Computing in GWAS

- GWAS identifies genetic variations (SNPs) influencing traits and diseases.
- Multivariate kernel methods like KRR model nonlinear relationships but are computationally demanding.
- A tile-centric framework with mixed-precision arithmetic (INT8 to FP64) accelerates GWAS (1.805 ExaOps/s, 5x faster than REGENIE).

HPC-Enabled Benchmarking for Genomic Data Compression

- Genomic datasets (FASTQ format) are massive, creating storage and accessibility challenges.
- SeqBench is an open-source benchmark suite for evaluating lossless and lossy compression methods.
- HPC enhances compression and downstream analysis, with some compressors achieving 8-54x space savings.

Accelerating Variant Calling Pipelines

- Variant calling identifies genetic variations and is computationally intensive.
- AVAH* is a GPU-aware system that optimizes variant calling by dynamically assigning tasks to CPUs/GPUs.
- AVAH* achieves 3.6x–5x speedup over CPU-only approaches.

Impact of Sequencing Technologies and Hardware-Accelerated Computing

- Advances in sequencing and hardware-accelerated computing have transformed genomic research.
- HPC, with GPUs, FPGAs, and multi-core CPUs, improves genomic data analysis.
- HPC enables pangenomic studies, capturing genetic diversity across populations.

Emerging Trends and Future Directions

- Emerging trends include AI/machine learning for enhanced data processing.
- Edge-to-core computing for distributed genomic analysis.
- Adaptive precision computing for dynamic resource allocation.

References

- [1] T. Yang, Y. Liu, C. Li, X. M. Shi, and S. Jin, “SeqBench: A Benchmark Suite for Lossless and Lossy Compression of Sequence Data,” in Proc. 15th ACM Int. Conf. Bioinformatics, Computational Biology and Health Informatics (BCB ’24), New York, NY, USA: ACM, 2024, Art. no. 58, pp. 1–7. [Online]. Available: https://doi.org/10.1145/3698587.3701386
- [2] H. Ltaief, R. Alomairy, Q. Cao, J. Ren, L. Slim, T. Kurth, B. Dorschner, S. Bougouffa, R. Abdelkhalak, and D. E. Keyes, “Toward Capturing Genetic Epistasis From Multivariate Genome-Wide Association Studies Using Mixed-Precision Kernel Ridge Regression,” in Proc. Int. Conf. High Performance Computing, Networking, Storage and Analysis (SC ’24), 2024, Art. no. 6, pp. 1–12. [Online]. Available:
- https://doi.org/10.1109/SC41406.2024.00012
- [3] A. Subramaniyan et al., “Accelerated Seeding for Genome Sequence Alignment with Enumerated Radix Trees,” in Proc. 48th Annu. Int. Symp. Computer Architecture (ISCA), Valencia, Spain, 2021, pp. 388–401. doi:
- 10.1109/ISCA52012.2021.00038
- [4] M. Olbrich, L. Bartels, and I. Wohlers, “Sequencing technologies and hardware-accelerated parallel computing transform computational genomics research,” Frontiers in Bioinformatics, vol. 4, 2024. [Online].
- Available: https://doi.org/10.3389/fbinf.2024.1384497
- [5] M. J. Das, K. Shehzad, and P. Rao, “Efficient Variant Calling on Human Genome Sequences Using a GPU-Enabled Commodity Cluster,” in Proc. 32nd ACM Int. Conf. Information and Knowledge Management (CIKM ’23), New York, NY, USA: ACM, 2023, pp. 3843–3848. [Online]. Available: https://doi.org/10.1145/3583780.3615268
- [6] Y. Wang, Y. Zhao, A. Bollas, et al., “Nanopore sequencing technology, bioinformatics and applications,” Nature Biotechnology, vol. 39, pp. 1348–1365, Nov. 2021. [Online]. Available: https://doi.org/10.1038/s41587-021-01108-x
- [7] A. Rhoads and K. F. Au, “PacBio Sequencing and Its Applications,” Genomics Proteomics Bioinformatics, vol. 13, no. 5, pp. 278–289, Oct. 2015. doi: 10.1016/j.gpb.2015.08.002. PMID: 26542840; PMCID: PMC4678779.
- [8] M. Hanussek, F. Bartusch, and J. Kr¨uger, “Performance and scaling behavior of bioinformatic applications in virtualization environments to create awareness for the efficient use of compute resources,” PLoS Computational Biology, vol. 17, no. 7, p. e1009244, Jul. 2021. Doi: 10.1371/journal.pcbi.1009244. PMID: 34283824; PMCID: PMC8323933.
- [9] S. Quazi, “Artificial intelligence and machine learning in precision and genomic medicine,” Medical Oncology, vol. 39, no. 8, p. 120, Jun. 2022. Doi: 10.1007/s12032-022-01711-1. PMID: 35704152; PMCID: PMC9198206.
- [10] R. Poplin, P. C. Chang, D. Alexander, et al., “A universal SNP and small-indel variant caller using deep neural networks,” Nature Biotechnology, vol. 36, pp. 983–987, Nov. 2018. [Online]. Available: https://doi.org/10.1038/nbt.4235