Accelerating Minimap2 for Accurate Long Read Mapping on Graphics Cards

Robert Thomson, Julia Anderson and Kurez Oroy

September 21, 2023

Accelerating Minimap2 for Accurate Long Read Mapping on Graphics Cards

Robert Thomson, Julia Anderson, Kurez Oroy

Abstract

Minimap2, a versatile tool for long read alignment, has gained prominence for its speed and precision. In this article, we explore how Graphics Processing Units (GPUs) can be harnessed to accelerate Minimap2, enhancing the efficiency and accuracy of long read mapping for genomics research. Long-read sequencing technologies have transformed genomics research, enabling the study of complex genomes and structural variations in unprecedented detail

I. Introduction:

Long-read sequencing technologies have transformed genomics research, enabling the study of complex genomes and structural variations in unprecedented detail.[1] Minimap2 has emerged as a powerful tool for aligning long reads to reference genomes. However, as genomic datasets continue to expand in size and complexity, the demand for faster and more efficient alignment methods has grown.[2]

The Significance of Accelerating Minimap2 on GPUs

The acceleration of Minimap2 on GPUs offers several key advantages:

Speed: GPUs, with their parallel processing capabilities, significantly reduce the time required for long read alignment.[3]

Accuracy: Maintaining alignment precision ensures the reliability of downstream genomic analyses.[4]

Efficiency: Faster alignment contributes to quicker genome assembly and structural variant detection.[5]

Strategies for Accelerating Minimap2 on GPUs

Efficient GPU acceleration of Minimap2 relies on specific strategies:

Parallel Processing: GPUs excel at parallelism, enabling the concurrent processing of multiple alignment tasks.[6]

GPU Optimization: Adapting Minimap2's algorithms to effectively utilize GPU architecture.[7]

Data Management: Efficient data transfer between the CPU and GPU to minimize overhead.[8]

Applications of GPU-Accelerated Minimap2

GPU-accelerated Minimap2 has broad applications:

Genome Assembly: Speeding up long read alignment enhances the efficiency of genome assembly projects.[9]

Structural Variant Detection: Accelerated alignment improves the detection of structural variations in genomes, critical for understanding genetic diseases and cancer.

Functional Genomics: Rapid and precise alignment supports research on gene expression and regulatory regions.[10]

Experimental Validation and Results

To validate the performance of GPU-accelerated Minimap2, researchers conducted experiments using real sequencing data. These experiments compared execution times and alignment accuracy between GPU-accelerated Minimap2 and traditional CPU-based implementations.

The results demonstrated substantial speed enhancements with GPU-accelerated Minimap2, even for extensive long read mapping tasks. Alignment times were significantly reduced, enabling the analysis of large genomic datasets with greater efficiency. Importantly, alignment accuracy remained consistently high, ensuring the reliability of genomic analyses.

II. Factors of acceleration

Accelerating Minimap2 for accurate long read mapping on graphics cards (GPUs) is a promising approach to speed up the alignment of long sequencing reads, such as those generated by third-generation sequencing technologies like PacBio and Oxford Nanopore. Minimap2 is a popular tool for aligning long reads to reference genomes, and leveraging GPU acceleration can significantly enhance its performance. Below are the key steps and considerations for achieving this acceleration:

1. **GPU Selection**: Choose a suitable GPU for your task. The choice of GPU depends on factors such as the size of your dataset, budget, and available hardware resources. GPUs from NVIDIA, such as the GeForce and Tesla series, are commonly used for bioinformatics tasks.
2. **CUDA Programming**: Minimap2 can be accelerated using CUDA, a parallel computing platform and API developed by NVIDIA for GPU programming. You will need to have CUDA installed on your system.
3. **CUDA-Aware Libraries**: Ensure that you have CUDA-aware libraries installed. These libraries are optimized to work seamlessly with CUDA-enabled GPUs. For Minimap2, you may need CUDA-aware versions of libraries like zlib, libdeflate, and others used by the tool.
4. **GPU-Enabled Minimap2**: Obtain or build a GPU-enabled version of Minimap2. Some versions of Minimap2 come with GPU support built-in, while others may require custom compilation. Check the Minimap2 documentation or GitHub repository for the latest information on GPU support.
5. **Data Preprocessing**: Prepare your sequencing data in a format that is compatible with the GPU-accelerated Minimap2. This may involve converting your data to the appropriate format or indexing reference genomes for faster access.
6. **Batch Processing**: When using GPUs, it's often more efficient to process multiple sequences in parallel (batch processing) to fully utilize the GPU's parallel computing capabilities. Minimize CPU-GPU data transfer overhead by batching sequences for alignment.
7. **Tuning Parameters**: Adjust Minimap2 parameters for GPU acceleration. Some parameters may need to be fine-tuned for optimal performance on a GPU. Experiment with different settings to find the best trade-off between speed and accuracy for your specific dataset.

8. **Benchmarking and Validation**: Before deploying GPU-accelerated Minimap2 in a production environment, thoroughly benchmark the tool to ensure that the accelerated version maintains the same level of accuracy as the CPU version. Validate the results against known reference data.
9. **Scaling**: If you have access to multiple GPUs, consider parallelizing the alignment process across multiple GPUs to further improve performance. Tools like NVIDIA NCCL (NVIDIA Collective Communications Library) can help with GPU-to-GPU communication.
10. **Monitoring and Optimization**: Monitor GPU usage and performance during alignment. Tools like NVIDIA's nvprof can help you identify bottlenecks and areas for optimization in your GPU-accelerated workflow.
11. **Documentation and Reporting**: Document the entire process, including software versions, parameters used, and benchmarking results. This documentation will be valuable for reproducibility and troubleshooting.
12. **Deployment**: Once you are satisfied with the performance and accuracy of your GPU-accelerated Minimap2, integrate it into your bioinformatics workflow for routine use.

Accelerating Minimap2 for long read mapping on GPUs can provide significant speedup, enabling faster analysis of large sequencing datasets. However, it requires careful setup, optimization, and validation to ensure accurate results. Additionally, keep in mind that the availability of GPU-accelerated versions of bioinformatics tools may evolve over time, so regularly check for updates and improvements in the software ecosystem.

Regenerate

III. Conclusion:

GPU-accelerated Minimap2 is a transformative development in long read mapping, offering an efficient solution to the computational demands of genomics research. By harnessing the parallel processing capabilities of GPUs, researchers can achieve significant performance improvements while preserving alignment precision. This technology has the potential to accelerate genome assembly, structural variant detection, and functional genomics research, advancing our understanding of complex genomes and their roles in health and disease.

IV. References:

[1] H. Sadasivan, D. Stiffler, A. Tirumala, J. Israeli, and S. Narayanasamy, "Accelerated Dynamic Time Warping on GPU for Selective Nanopore Sequencing," *bioRxiv,* p. 2023.03. 05.531225, 2023.

[2] T. Dunn *et al.*, "Squigglefilter: An accelerator for portable virus detection," in *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture*, 2021, pp. 535-549.

[3] H. Sadasivan, "Accelerated Systems for Portable DNA Sequencing," University of Michigan, 2023.

[4] H. Sadasivan, M. Maric, E. Dawson, V. Iyer, J. Israeli, and S. Narayanasamy, "Accelerating Minimap2 for accurate long read alignment on GPUs," *Journal of biotechnology and biomedicine,* vol. 6, no. 1, p. 13, 2023.

[5] H. Sadasivan *et al.*, "Rapid real-time squiggle classification for read until using rawmap," *Archives of clinical and biomedical research,* vol. 7, no. 1, p. 45, 2023.

[6] P. Teengam *et al.*, "NFC-enabling smartphone-based portable amperometric immunosensor for hepatitis B virus detection," *Sensors and Actuators B: Chemical,* vol. 326, p. 128825, 2021.

[7] K. Wu, T. Klein, V. D. Krishna, D. Su, A. M. Perez, and J.-P. Wang, "Portable GMR handheld platform for the detection of influenza A virus," *ACS sensors,* vol. 2, no. 11, pp. 1594-1601, 2017.

[8] S. Mittal, A. Joshi, and T. Finin, "Cyber-all-intel: An ai for security related threat intelligence," *arXiv preprint arXiv:1905.02895,* 2019.

[9] A. Dutta and S. Kant, "An overview of cyber threat intelligence platform and role of artificial intelligence and machine learning," in *Information Systems Security: 16th International Conference, ICISS 2020, Jammu, India, December 16–20, 2020, Proceedings 16*, 2020: Springer, pp. 81-86.

[10] A. Ibrahim, D. Thiruvady, J.-G. Schneider, and M. Abdelrazek, "The challenges of leveraging threat intelligence to stop data breaches," *Frontiers in Computer Science,* vol. 2, p. 36, 2020.