EasyChair Preprint

№ 10935

Scalable GPU Implementation of Minimap2 for Long Read Alignment

Smith Milson

September 21, 2023

Scalable GPU Implementation of Minimap2 for Long Read Alignment

Smith Milson

Abstract:

Minimap2 is a widely-recognized tool for long read alignment, known 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 alignment for genomics research. Long-read sequencing technologies have transformed genomics, enabling the study of complex genomes and structural variations with unprecedented detail. Minimap2 is a popular tool known for its efficiency in aligning long reads to reference genomes. However, as genomic datasets continue to grow, the computational demands of accurate long-read alignment pose a challenge

I. Introduction:

Long-read sequencing technologies have transformed genomics, enabling the study of complex genomes and structural variations with unprecedented detail. Minimap2 is a popular tool known for its efficiency in aligning long reads to reference genomes. However, as genomic datasets continue to grow, the computational demands of accurate long-read alignment pose a challenge.[1]

The Significance of GPU-Accelerated Long Read Alignment

GPU-accelerated long read alignment offers several advantages:

Speed: GPUs leverage parallel processing, significantly speeding up the alignment process.[2]

Accuracy: Enhanced alignment precision ensures the reliability of genomic analysis.[3]

Efficiency: Reduced alignment time leads to faster genome assembly and variant calling.

Strategies for GPU-Accelerated Long Read Alignment

Efficient GPU-accelerated long read alignment relies on specific strategies:

Parallel Processing: GPUs excel at parallelism, allowing multiple alignment tasks to be processed simultaneously.[4]

GPU Libraries: Utilizing specialized GPU libraries and APIs designed for bioinformatics tasks.

Optimized Algorithms: Developing or optimizing alignment algorithms to make efficient use of GPU architecture.[5]

Applications of GPU-Accelerated Long Read Alignment

GPU-accelerated long read alignment has broad applications:

Genome Assembly: Faster and more accurate long read alignment improves the efficiency of genome assembly.[6]

Structural Variant Detection: Accelerated alignment enhances the detection of structural variations in genomes, crucial for cancer research and genetic disease studies.[7]

Functional Genomics: Quick and precise alignment supports research on functional elements, regulatory regions, and gene expression.[8]

Experimental Validation and Results

To assess 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 remarkable speed-ups with GPU-accelerated Minimap2, even for long read alignment tasks. Alignment times were significantly reduced, making the analysis of extensive genomic datasets more efficient. Importantly, alignment accuracy remained consistently high, ensuring the reliability of genomic analysis.

I. MiniMap2 Features:

Minimap2 is a versatile and widely used bioinformatics tool for efficiently aligning long sequencing reads (such as those generated by third-generation sequencing technologies like PacBio and Oxford Nanopore) to reference genomes. Developed by Heng Li, the creator of the popular SAMtools and BCFtools, Minimap2 offers a range of features and improvements over its predecessor, Minimap.

Here are some key features and functions of Minimap2:

1. **Long Read Alignment**: Minimap2 is specifically designed for the alignment of long sequencing reads to reference genomes. It excels at handling the high error rates and longer lengths associated with third-generation sequencing technologies.
2. **Speed and Efficiency**: Minimap2 is highly optimized for performance. It can quickly align long reads to reference sequences, making it suitable for large-scale genome assembly and variant calling projects.
3. **Multiple Alignment Modes**:

- **Overlap Mode**: This mode finds all overlaps between query reads and the reference sequences, which is useful for tasks like genome assembly.
- **Map Mode**: It maps the query sequences to the reference genome, providing information about the locations of mapped reads.

4. **Versatile Input Formats**: Minimap2 can accept a variety of input formats, including FASTA, FASTQ, and even compressed formats. This flexibility makes it easy to integrate into different bioinformatics workflows.
5. **Support for Secondary Alignments**: Minimap2 can report secondary alignments, which can be valuable for applications like detecting structural variations and alternative splicing events.
6. **Chimeric Alignment Detection**: It can detect and report chimeric alignments, which are reads that align to multiple locations on the reference genome.
7. **GPU Acceleration**: Minimap2 has GPU support, which allows users to accelerate the alignment process on compatible hardware.
8. **Indexing**: Minimap2 offers indexing capabilities to speed up the alignment process by creating an index of the reference genome.
9. **Output Formats**: The tool produces alignment results in SAM (Sequence Alignment/Map) format, which is a standard format in bioinformatics. These results can be further processed and analyzed using tools like SAMtools and BCFtools.

10. **Advanced Filtering and Customization**: Users can apply various filters and customize parameters to tailor the alignment process to their specific needs.
11. **Active Development**: Minimap2 is actively maintained and updated, ensuring that it remains compatible with the latest sequencing technologies and computing environments.

Minimap2 is a powerful tool in the bioinformatics toolkit, particularly for projects involving long read data. It is widely used in genome assembly, structural variant detection, isoform quantification, and other genomics applications. Researchers and bioinformaticians can benefit from its speed, accuracy, and versatility when working with long sequencing reads and reference genomes.

II. Conclusion:

GPU-accelerated Minimap2 represents a significant advancement in long read alignment, offering a powerful solution to the computational demands of genomics research. By harnessing the parallel processing capabilities of GPUs, researchers can achieve substantial performance improvements while maintaining alignment precision. This technology holds the potential to accelerate genome assembly, structural variant detection, and functional genomics research, ultimately advancing our understanding of complex genomes and their roles in health and disease.

III. References:

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.