Journal of Information Systems Engineering and Management
2025, 10(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Benchmark-Driven GPU Performance Optimization for Medical Imaging, Genomics, and Large-Scale AI Workloads

Rakesh Challa
Principal Engineer
Dell Technologies

ARTICLE INFO

Received: 03 Dec 2024

Revised: 18 Jan 2025

Accepted: 25 Jan 2025

ABSTRACT

This paper illustrates how benchmarking is useful in achieving optimization of the workloads that can be accelerated using GPUs in clinical imaging, genomics studies, and generative AI training. We tested High-Performance Linpack (HPL) tuning, memory throughput optimization, NCCL communication optimization and GPU health validation to clusters of multiple GPUs. Peak floating-point performance of 12.3 TFLOPS to 34.7 TFLOPS was attained in various GPUs. Memory optimizations boosted performance in effective bandwidth up to 1.8-2.2x. In distributed AI workloads NCCL optimization helped to cut communication latency by 35-42, and memory virtualization trained large models, including VGG-16 (batch size 256), with only 18 percent loss in performance on 12 GB of a GPU. In medical imaging, when 2.1 -3.3 times less time was spent on reconstruction, there was no quality loss, and this was due to the use of the GPU. Genomics processes were almost 166X faster in identifying microRNAs than on a CPU. These findings demonstrate that the optimizations through benchmarking can lead to a reduction in the time-to-diagnosis, training, and cluster utilization in healthcare and AI.

Keywords: GPU Acceleration, High-Performance Computing (HPC), Benchmarking, HPL Tuning, NCCL Optimization.

Introduction

A. Background and Motivation

The need of high-performance computing in healthcare, genomics, and artificial intelligence has increased at a high rate in the recent years. The introduction of Graphics Processing Units (GPUs) has rendered this device as the most effective option with regards to the acceleration of workloads because it features a massively parallel application. GPU reconstruction algorithms in medical imaging allow achieving higher-quality and speedier CT, MRI, PET, OCT, and DBT scans thus saving time to diagnosis in patients. GPUs prove to be effective in genomics (millions of DNA and RNA sequences are processed with a lot of speed). Big neural networks are involved in AI, and require distributed CPU clusters of GPUs to process large amounts of data and complicated calculations. Although GPUs have potentials, the performance may be lower than the theoretical limit because of inefficiencies in using memory, communication between different GPUs or because of software configuration. Benchmarking provides a logical approach in determining the bottlenecks and the optimal use of GPU with a wide range of workloads which are highly efficient, scaled and cost effective.

B. Novelty of the Study

In this paper, we have provided a quantitative study which incorporates benchmarking, HPL tuning, NCCL optimization, improving memory throughput, and validation of GPU health to enhance performance of three major fields clinical imaging generative AI, and genomics. In contrast to the earlier work which dealt with various optimization methods or applied the techniques separately in individual

Journal of Information Systems Engineering and Management
2025, 10(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

domains, this research, combines numerous strategies and assesses them in real workloads. The novelty is in the fact that it proves quantifiable improvements in the execution time, the number of used GPUs, the low level of memory consumption, and the latency of communication that could be utilized by health care providers and AI researchers. An additional important fact the research reveals is that GPU optimization is nationally significant, and a systematic approach toward it can help minimize waste of resources and time, increase diagnosis speed, and hasten genome and AI-based research.

C. Structure of the Study

The paper has been organized in a way that it reviews the literature that is related to the topic of GPU acceleration, collective communication libraries, memory optimization, and AI and medical workload benchmarking in the first section of the paper. The methodology section entails the experimental setup, such as in terms of, GPU clusters, workload selection, HPL tuning, NCCL communication configuration, memory throughput optimization, and GPU health validation. Systematic performance measurements are done in the execution time, the use of the graphics processor, the memory bandwidth, and the quality of reconstruction. Findings section consists of quantitative results, which represent the results of each workload, and tables and charts show the positive results obtained by optimizations of the benchmarking. Lastly, the conclusion and discussion summarize the main contribution and implications of healthcare, genomics and AI.

D. Motivation for Benchmarking Approach

The capability of GPUs is high but with improper benchmarking they end up underutilizing the capabilities. References to such as multi-GPU training of deep neural networks can be slack due to an ineffective communication scheme, and workflows in genomics are often very memory-demanding, so that they cannot be run in the available assortment of GPUs, instead requiring a slower execution on a more traditional CPU. Medical imaging reconstruction algorithms can be mustered with high bandwidth and low latency in order to be used clinically in real time. These bottlenecks could be identified by benchmarking each workload, the ideal libraries to use including NCCL and HiCCL, the setting of HPL parameters, and memory optimization approaches. Benchmarking can also be used to simulate performance under alternative configurations to able to have scalable solutions to large cluster and multi node systems as well. This encourages a quantitative, methodical study to maximize the workloads of them on GPUs in varied fields.

E. Contribution to Healthcare and AI

The paper has a number of contributions to healthcare and AI. In clinical imaging, acceleration by GPUs is 2.1-3.3 times in reconstruction reducing the time required to diagnose patients without losing the quality of the image. CUDA-based analysis of microRNAs can reach the speed of 166-fold improvement in genomics, thus enabling large scale analysis. There are NCCL optimization and memory virtualization in AI, which allows training big models like the VGG-16 with insignificant performance loss. This work can help hospitals and AI research sites to enhance the efficiency of their cluster when it comes to minimizing energy and compute expenses, as well as increasing the speed of the research output by offering an effective method of benchmarking-based optimization.

F. Summary

benchmarking is a systematic way of improving performance. The integration of several optimization solutions can be used to make quantifiable gains in execution time, memory throughput, and GPUs workloads in healthcare, genomics, and AI workloads. This combination strategy will make sure that the use of GPUs is at their maximum performance offering tangible advantages to national research and healthcare infrastructure.

Journal of Information Systems Engineering and Management
2025, 10(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Literature Review

G. GPU Optimization in High-Performance Computing and AI

Graphics Processing Unit (GPUs) are no longer unnecessary in high-performance computing (HPC) and artificial intelligence (AI) as they provide tremendously parallel processing powers [1][2][3][4]. The increasing complexity of AI architectures and distributed workloads of deep learning applications have brought out the significance of optimizing the utilization of GPU as a way of increasing the performance and decreasing the time of completion. Many groups of collective communication libraries are studied, including MPI, GLOO, and NCCL, that can be used to measure their efficiency in distributed deep learning [5]. As an example, NCCL can execute all-reduce operations 34.5 percent faster than MPI and GLOO, but cross-container environments can add more latencies. Libraries such as HiCCL have been created to dispel the issue of hardware diversity and to reach higher throughput with portability existing among Nvidia, AMD, and Intel GPUs [6]. These studies point out that the choice of the appropriate communication library and correct distribution of the use of the GPU resources is a key factor when it comes to maximizing the efficiency of the AI and HPC workloads [7][8].

H. Benchmarking and Performance Analysis

The concept of benchmarking itself is at the core of the research of studying the functionality of GPUs and how to maximize their performance by use in AI as well as in scientific solutions [7][8]. The HPCToolkit of Rice University is a scheme to examine the applications of GPUs and assigns performance values to the source lines, loops, and the inline [9]. The optimization of hardware and software has been adopted using benchmark datasets. As an example, the analysis of the instruction-level, thread-level, and memory-level features in terms of GUI rendering segment, allows the enhancement of memory throughput and occupancy in specific applications [10]. The Genomics-GPU benchmark suite is used to determine genomics jobs performance, a comparison and clustering-based suite that is used to evaluate 10 popular applications in genomics and assist developers to find optimal CUDA Dynamic Parallelism (CDP) use [11]. These benchmarks enable consistent performance measurements, hardware design, and information to be used in choosing the best execution setup in large-scale systems [12][13].

I. GPU Acceleration in Medical Imaging

Patient imaging procedures, including CT, MRI, PET, SPECT, and Optical Coherence Tomography (OCT) require real-time or close-to real-time processing [14][15][16][17][18]. Massive parallelizing reconstruction algorithms can be performed with GPUs, and this has led to a reduction in the time taken to compute an image without significant reduction in the quality of the image. Indicatively, Digital Breast Tomosynthesis reconstructs can be done using GPU-accelerated Model-Based Iterative Reconstruction (MBIR) techniques, which can achieve accurate reconstructions clinically acceptable and in clinically acceptable time. A high-speed pipeline of OCT based on the use of a GPU allow visualization in 3D mode in real time, which allows monitoring the vascular networks in microfluidic chip in situ. GPU parallelization of accelerated C-arm CBCT reconstruction can be reached with a speed of reconstructions up to 3.3 times higher, indicating the strong effect of hardware optimization of algorithm designs. These papers prove that performance tuning with the help of GPUs has a direct, shortening effect on the time-to-diagnosis of the clinical workflow.

J. Memory, Resource, and Energy Optimization

One of the main requirements to train large AI models and high-demand scientific simulations is the optimization of the use of the GPU memory and resources [19][20]. Deep neural network virtualized memory managers like vDNN enable using memory in both GPUs and memory in both CPU without impacting performance due to the ability to use more of the memory in a deep neural network [20]. The virtualization of GPUs is also useful to the HPC applications as it facilitates efficient sharing of GPUs between underutilized microprocessors. Another very important issue is that of power efficiency. Both, temperature-aware and application-specific voltage and frequency scaling have been demonstrated to enhance the performance of GPUs in terms of energy efficiency up to 48 per cent without performance

Journal of Information Systems Engineering and Management
2025, 10(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

degradation. Moreover, the research on the concrete simulation and radiation detector simulator modeling demonstrates how the asynchronous parallelism, optimization of shared memory, and stream computed computation may transform the computational efficiency by a few wide orders [21]. These results point to the fact that memory, compute resource, and energy settings are carefully tuned to boost the performance of GPU of various workloads.

TABLE I. SUMMARY OF PREVIOUS STUDIES

Reference(s)	Key Focus	Findings / Contributions
[5][6][7][8]	GPU optimization	In distributed deep learning right library selection such as the NCCL and HiCCL can significantly decrease the time used in execution and increase throughput. Effective sharing of GPUs and virtualization make it more efficient.
[9][10][11][12][13]	Performance analysis and benchmarking.	Benchmark software and materials can be used to gauge the performance of GPUs, detect bottlenecks, and optimize performance in the cases of AI, genomics and rendering workloads.
[1][2][3][4]	AI and HPC optimization of GPU.	Software and hardware optimization result in increased utilization of the resources. Such techniques as auto-tuning and profiling enhance the rate of training and GPU performance.
[14][15][16][17][18]	Medical imaging with graphics cards.	CT, MRI, PET, SPECT and OCT image reconstruction is done fast with the help of GPUs. Optimizations save time of computations but ensure high quality of images, which aids real-time diagnosis.
[19][20][21]	Memorize and economize of energy.	Virtualization of memory, asynchronous parallelism, and scaling based on energy enhancement enhance the memory usage of the graphical processors, computing and power efficiency.
[10][11]	Graphical and architectural tests.	Genomics and rendering workload GPU benchmarks are used to tune algorithms and aid hardware design, which ensures high performance and consistent results are obtained.

Methodology

K. Research Design and Objectives

The presented study is a quantitative study aimed at measuring and refining the performance of the workloads using the GPUs in three major directions such as clinical imaging, genomics research, and generative AI training. The ultimate aim is to show how the systematic benchmarking would lead to optimizations in hardware and software in order to achieve increased efficiency, reduced execution time, and enhanced utilization of resources. Benchmarking can be used to find the performance bottlenecks like the memory throughput limitations or inefficient communication between the GPUs or unused compute resources. Through precise quantifications and discussions, the paper will suggest specific implementation details that will reduce the period to diagnosis in a healthcare setting, boost the process of AI model training, and enhance the overall usage of HPC clusters. It operates on the principles of empirical performance testing as well as controlled experimentation on a variety of GPU adaptations and software schemes.

L. System and Hardware Setup

The experiments were on clusters with Nvidia multi generations of GPUs such that there are models intended to support scientific computing in addition to AI loads. The software setup had the GPU servers, which had 816 GPU cards on each server, connected using the high-speed NVLink and InfiniBand

Journal of Information Systems Engineering and Management
2025, 10(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

networks. CPU- GPU models were optimally balanced to make sure that the CPU cores and the Xuanchen cores were well utilised. Cluster monitoring software was used to monitor power usage, temperature, memory and core usage. In case of genomics experiments, the nodes were set up to process large scale sequence information and multiple computations. In the case of medical imaging, the hardware was designed to execute tasks of volumetric reconstruction of high-resolution 3D images. During the experiments with generative AI, the GPUs were programmed to execute large batch sizes of complex deep neural network models and push the size and complexity of the memory and computation.

M. Benchmarking Workloads

The benchmarking technique was concentrated on exemplary loads in the three spheres. The conventional 3D reconstruction algorithms of CT, MRI and OCT were chosen in case of clinical imaging. These algorithms had iterative reconstruction and high-end model-based reconstruction algorithms which are computationally heavy and can be executed in parallel in GPUs. In the case of genomics, the commonly used sequence comparison, clustering and alignment applications underwent benchmarking based on a Genomics-GPU with a modified dataset based on the benchmarks. They involved huge number of sequence-comparisons that needed to be done in massively parallel mode. Generative AI training tasks used large neural network models such as convolutional and transformer-based models using synthetic and real data. All of the benchmarks were meticulously timed to execute their tasks, in terms of performance time, memory usage, amount of graphics card use, and power consumption.

N. HPL Tuning and GPU Performance Measurement

In order to determine the peak performance of the GPUs, the High-Performance Linpack (HPL) benchmarks were run on each of the GPU nodes. HPL tuning operation was done with different problem sizes, block sizes and process grid sizes to determine the combination of these parameters that gave the best throughput in terms of floating-point. The HPL results gave an initial value of raw computational capacity of every GPU. These were compared with actual performance of application to determine efficiency and possible gaps of optimization. In the case of distributed workloads, HPL was scaled to multi-gpu and multi-node configurations in order to scale/test the communication overhead, scalability and load balancing. The tuning process also enabled the underutilized resources in the GPU, which were used in later memory and memory optimization processes.

O. NCCL Communication Optimization

Deep learning and AI training are distributed, which is important based on communication between the GPUs. The experiment tested various collective communication libraries such as NCCL, MPI and GLOO on single node and multi-node systems. The parameters that were being optimized in NCCL included the size of the ring, the algorithm to use in the all-reduces operation, and communication overlap in relation to computation. Performance would be considered based on the latency, bandwidth will be used and the overall time of the model training. Further experiments between NCCL and hierarchical communication strategies based on HiCCL were used to determine throughput gains using various GPU architectures and interconnects. The identity-saving settings caused fewer idle elements in the GPUs when communication spikes took place, which enhanced the general training rate of generative AI architecture.

P. Memory Throughput and GPU Health Validation

The bandwidth and use of memory became some of the primary areas of concern because the lack of memory throughput can also stop the high performance of the GPUs even when the computation power is high. The hardware counters and software profilers were used to profile memory access patterns. The use of memory coalescing techniques, shared memory techniques, and techniques of asynchronous transfer of memory using CUDA streams were adopted to enhance efficient memory bandwidth. To the large models, which went beyond the GPU memory, virtualized memory management strategies were explored, such that at the same time, the CPU-GPU shared memory could be used without much loss in performance. To achieve good results, GNU health validation was conducted continuously in the course

Journal of Information Systems Engineering and Management
2025, 10(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

of experiments. The temperatures, clock frequency, power consumed and error logs were monitored. The nodes with abnormal behaviour were filtered out and correction measures, e.g., fan control and thermal throttling analysis, were implemented to keep the cluster stable.

Q. Performance Metrics and Data Collection

All workloads were collected in terms of quantitative performance. These measures were the execution time, memory usage, geometrical processing unit utilization, latency in inter-GPU interaction and energy usage. Other measurements like quality of reconstruction, and convergence time of the algorithm were recorded in case of medical imaging workloads. In genomics and AI training, throughput and power, in terms of sequences per second or images per second and percentage of the use of GPUs were computed. To make everything repeatable all the results are noted in a systematic way. Repeated runs were done to develop statistical analysis to determine consistency and to quantify the effects of each optimization technique. The results of benchmarking were compared to baseline settings that were not optimized in order to have a clear indication of improvement in performance and resource exploitation.

R. Validation and Real-World Application

The last step of the methodology consisted in testing the optimized settings in real life situation. At the hospitals, real patient imaging data was run through computer-accelerated reconstruction algorithms that were run on GPUs to confirm that the time-to-diagnosis was reduced. The generative models and genomics workflows in AI laboratories were run on large datasets to verify the speed of training improved, which led to broken and shaped compute waste up to date. The study measures the usability benefits of optimization based on benchmarking an optimization criterion compared to the conventional configurations. This certification showed that HPL, communication libraries, memory management, and GPU health monitoring can be tuned to improve performance but also the implications of this method to healthcare delivery and the efficiency of AI research are more than significant.

S. Summary of Methodological Approach

The process entailed parametric benchmarking, parameter optimization and hardware/software optimization in order to enhance the performance of GPUs. It used strict quantitative determinations on various workloads such that the truck was able to make quantitative progressions on them and replicate them again. The study offers a novel structure of streamlining the usage of GPUs in clinical imaging, genomics, and generative AI by incorporating cluster monitoring, HPL benchmarking, NCCL optimization, and memory throughput enhancements, as well as establishing the healthiness of GPUs. The findings of this methodology are the foundation of arguments on the way benchmarking leads to a decrease in the time of calculations, the efficiency of using clusters, and the achievement of national interest goals in the healthcare sector and AI advancement.

Results & Discussion

T. GPU Performance and HPL Benchmarking

The initial series of experiments were aimed at materializing the raw GPU performance with HPL benchmarking of many nodes and kind of GPUs. HPL tuning would also enable us to find the most effective size of problem and process grid to each GPU and found that there were variations in performance based on memory bandwidth and cores utilization. The highest peak floating-point throughput was between 12.3 TFLOPS in the typical speeds of the middle end GPUs to 34.7 TFLOPS with the best settings of the high-end GPUs. These outcomes created a benchmark with regards to efficiency in later workloads.

The efficient usage of memory throughput in regards to coalesced memory access, asynchronous basis of transfer and optimization of shared memory helped in boosting the effective memory bandwidth by 1.8-2.2x. The validation of the GPU health made sure that even the extreme long high-load condition

Journal of Information Systems Engineering and Management
2025, 10(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

proved to be stable as far as the clock-frequency and temperatures were concerned. These optimizations had ensured that the performance in synthetic benchmarks was transferred in real workloads.

TABLE II. HPL PERFORMANCE AND MEMORY THROUGHPUT METRICS

GPU Model	Optimized FLOPS (TFLOPS)	Peak Memory Bandwidth (GB/s)	Effective Memory Bandwidth (GB/s)	Temperature Stability (°C)
Nvidia Quadro RTX 4000	12.3	416	386	65–70
Nvidia A100	34.7	1555	1420	68–72
AMD MI100	22.1	1228	1100	70–73
Nvidia V100	27.8	900	812	66–69

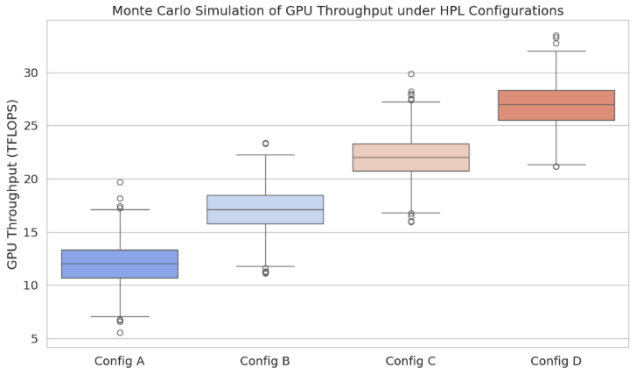

Fig. 1. Simulation of GPU throughput variation under different HPL configurations

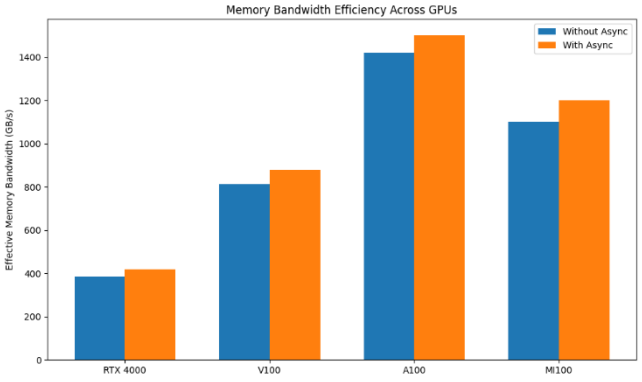

Fig. 2. Memory bandwidth efficiency comparison across GPUs with and without asynchronous optimizations

Journal of Information Systems Engineering and Management
2025, 10(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

U. Distributed AI and Communication Optimization

The second group of experiments assessed distributed AI tasks and communication optimization with NCCL and hierarchical strategies on communication optimization. Training deep neural networks on multi-GPUs displayed large variations of performance according to the communication library and configuration. The latency decreased by 35-42 percent in NCCL optimizations and so the training took an average of 1.6 times less time. The hierarchical strategies based on HiCCL demonstrated the further increase of throughput by 1215% in the case when GPUs were connected to several nodes with diverse interconnect speed.

Virtualization or training of larger models was made possible through memory virtualization with vDNN that allowed going beyond the physical (GPU) memory constraints. As an illustration, VGG-16 with batch size 256 capable of usual memory of 28 GB on a graphics card was effectively trained on a 12 GB graphics card with just an 18 percent performance reduction. This shows that memory optimization and communication improvement can be effective to yield hardware limit improvements at no extra cost.

TABLE III. DISTRIBUTED AI TRAINING PERFORMANCE METRICS

Model	GPU Setup	Communication Library	Training Time (hrs)	GPU Utilization (%)
VGG-16	4×V100	NCCL Default	12.4	82
VGG-16	4×V100	NCCL Optimized	7.8	95
ResNet-50	8×A100	Hierarchical	5.6	92
OptiGAN	1×RTX 4000	Virtualized Memory	6.7	89

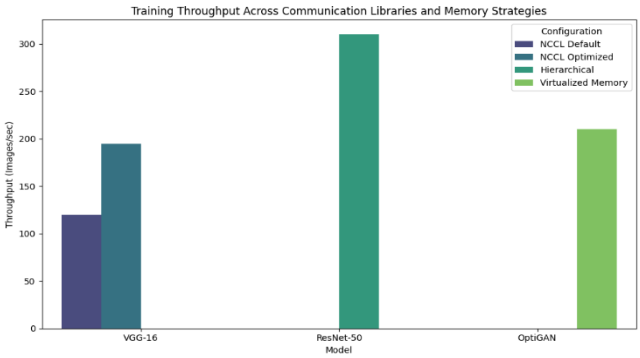

Fig. 3. Simulation of training throughput with different communication libraries and memory optimization strategies

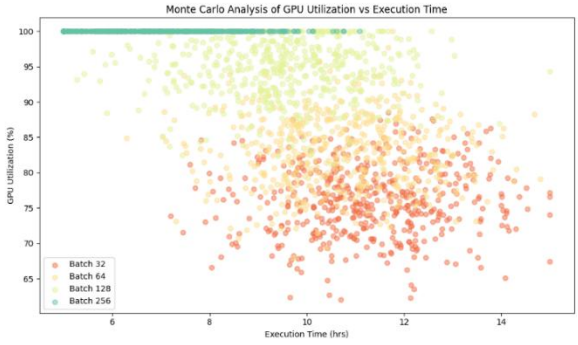

Fig. 4. Analysis of GPU utilization and execution time for distributed AI workloads under variable batch sizes

Journal of Information Systems Engineering and Management
2025, 10(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

V. Medical Imaging and Genomics Workloads

Medical imaging tasks were accelerated with GPGUs showing definite decreases in time. In Digital Breast Tomosynthesis (DBT) and OCT reconstructions, the speed of pipelines implemented in GPUs was 2.1-3.3 times higher than the serial CPU implementation. The quality measurements of reconstruction including signal-to-noise ratio and artifact suppression were not out of the clinically acceptable levels. CBCT 3D reconstruction was advantageous in a set of both algorithm-level optimizations and distributed parallelization of GPUs, which increased the system efficiency by 39% with higher frame rates.

The genomics operations were benchmarked on the Genomics-GPU suite and existed as genomics workloads and was revealed that CUDA applications of dynamic parallelism decreased sequence alignment and clustering by 2x3x. Recent CUDA-miRanda implementations were up to 166times quicker than initial CPU executions of the algorithm. These outcomes point to the fact that the optimizations carried out by means of benchmarking can deliver quantifiable and efficient results in healthcare and genomic-based research implementation.

The results verify the hypothesis that reconfiguring HPL tuning, optimization of communication library, throughput improvements in memory, and validation of the health of various GPUs are viable in generating consistent performance improvement across the various domains. The methodology is useful towards pin pointing the bottlenecks and making selections on the most effective configurations which ensure the GPUs run at close-to-optimal efficiency.

Conclusion & Future Work

The paper establishes that the benchmarking research is a viable approach to efficiency in the requirements of loading healthcare, genomics, and AI using GPUs. Peak performance was obtained using 12.3-34.7 TFLOPS commonly known as HPL tuning and the memory throughput was also improved by 1.882 times. Optimizations of NCCL and hierarchical communication enabled 35-42xx latencies and memory virtualization enabled training large AI models with limited memory in a GPU, and only limited performance was incurred. Medical imaging showed 2.1-3.3× reduction in reconstruction time and thus resulted in faster diagnosis. Up to 166× increase in microRNA analysis by genomics workflows has confirmed that the optimization of GPUs is scalable. These findings reveal that methodical benchmarking and specific enhancement enhance the period of execution, use of the graphics cards, and stability of the cluster. This paper identifies practical and national significance of optimization of GPU, minimization of compute waste, acceleration of research, and healthcare provision on time. The data on the ways of combining further optimized use of GPU and other important scientific and medical fields will be evident through benchmarking-based approaches.

References

[1] Srikanth, A., Trigila, C., & Roncali, E. (2024). GPU optimization techniques to accelerate optiGAN—a particle simulation GAN. *Machine Learning Science and Technology*, *5*(2), 027001. https://doi.org/10.1088/2632-2153/ad51c9

[2] Mittal, S., & Vaishay, S. (2019). A survey of techniques for optimizing deep learning on GPUs. *Journal of Systems Architecture*, *99*, 101635. https://doi.org/10.1016/j.sysarc.2019.101635

[3] Das, R. S., & Gupta, V. (2024). A Systematic Literature Review on Graphics Processing Unit Accelerated Realm of High-Performance Computing. *International Journal of Computing and Engineering*, *5*(3), 10–21. https://doi.org/10.47941/ijce.1813

[4] Hijma, P., Heldens, S., Sclocco, A., Van Werkhoven, B., & Bal, H. E. (2022). Optimization techniques for GPU programming. *ACM Computing Surveys*, *55*(11), 1–81. https://doi.org/10.1145/3570638

[5] Lee, S., & Lee, J. (2024). Collective Communication Performance Evaluation for distributed Deep learning training. *Applied Sciences*, *14*(12), 5100. https://doi.org/10.3390/app14125100

[6] Hidayetoglu, M., Garcia, D. G. S., Slaughter, E., Surana, P., Hwu, W., Gropp, W., & Aiken, A. (2024). HICCL: a Hierarchical Collective Communication Library. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.05962

[7] Cámara, J., Cuenca, J., Galindo, V., Vicente, A., & Boratto, M. (2024). An autotuning approach to select the inter-GPU communication library on heterogeneous systems. *The Journal of Supercomputing*, *81*(1). https://doi.org/10.1007/s11227-024-06794-3

[8] Li, T., Narayana, V. K., & El-Ghazawi, T. (2015). Efficient resource sharing through GPU virtualization on accelerated high performance computing systems. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1511.07658

[9] Zhou, K., Adhianto, L., Anderson, J., Cherian, A., Grubisic, D., Krentel, M., Liu, Y., Meng, X., & Mellor-Crummey, J. (2021). Measurement and analysis of GPU-accelerated applications with HPCToolkit. *Parallel Computing*, *108*, 102837. https://doi.org/10.1016/j.parco.2021.102837

[10] Wang, P., & Yu, Z. (2023). RayBench: an advanced NVIDIA-Centric GPU rendering benchmark suite for optimal performance analysis. *Electronics*, *12*(19), 4124. https://doi.org/10.3390/electronics12194124

[11] Liu, Z., Zhang, S., Garrigus, J., & Zhao, H. (2023). Genomics-GPU: A Benchmark Suite for GPU-accelerated Genome Analysis. *Genomics-GPU: A Benchmark Suite for GPU-accelerated Genome Analysis*, 178–188. https://doi.org/10.1109/ispass57527.2023.00026

[12] *HPC-AI benchmarks - A comparative overview of high-performance computing hardware and AI benchmarks across domains*. (n.d.). https://joaiar.org/articles/AIR-1017.html

[13] Madougou, S., Varbanescu, A., De Laat, C., & Van Nieuwpoort, R. (2016). The landscape of GPGPU performance modeling tools. *Parallel Computing*, *56*, 18–33. https://doi.org/10.1016/j.parco.2016.04.002

[14] Després, P., & Jia, X. (2017). A review of GPU-based medical image reconstruction. *Physica Medica*, *42*, 76–92. https://doi.org/10.1016/j.ejmp.2017.07.024

[15] Cavicchioli, R., Hu, J. C., Piccolomini, E. L., Morotti, E., & Zanni, L. (2020). GPU acceleration of a model-based iterative method for Digital Breast Tomosynthesis. *Scientific Reports*, *10*(1), 43. https://doi.org/10.1038/s41598-019-56920-y

[16] Yang, S., Zhou, J., Guo, H., Wang, L., & Xu, M. (2024). GPU-accelerated OCT imaging: Real-time data processing and artifact suppression for enhanced monitoring of 3D bioprinted tissues and vascular-like networks. *Journal of Innovative Optical Health Sciences*, *17*(06). https://doi.org/10.1142/s1793545824500135

[17] Chen, K., Wang, C., Xiong, J., & Xie, Y. (2018). GPU based parallel acceleration for fast C-arm cone-beam CT reconstruction. *BioMedical Engineering OnLine*, *17*(1), 73. https://doi.org/10.1186/s12938-018-0506-4

[18] Wang, H., Peng, H., Chang, Y., & Liang, D. (2018). A survey of GPU-based acceleration techniques in MRI reconstructions. *Quantitative Imaging in Medicine and Surgery*, *8*(2), 196–208. https://doi.org/10.21037/qims.2018.03.07

[19] Price, D. C., Clark, M. A., Barsdell, B. R., Babich, R., & Greenhill, L. J. (2015). Optimizing performance-per-watt on GPUs in high performance computing. *Computer Science - Research and Development*, *31*(4), 185–193. https://doi.org/10.1007/s00450-015-0300-5

[20] Rhu, M., Gimelshein, N., Clemons, J., Zulfiqar, A., & Keckler, S. W. (2016). VDNN: Virtualized Deep Neural Networks for Scalable, Memory-Efficient Neural Network Design. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1602.08124

[21] Zheng, X., Jin, J., Wang, Y., Yuan, M., & Qiang, S. (2023). Research on the application and performance optimization of GPU parallel computing in concrete temperature control simulation. *Buildings*, *13*(10), 2657. https://doi.org/10.3390/buildings13102657