Computer Fraud and Security
ISSN (online): 1873-7056

Architecting GPU-Accelerated Supercomputing for Real-Time Clinical AI in Large Hospital Systems

Rakesh Challa

Principal Engineer, Dell Technologies, USA

Abstract—The current hospitals require quick and dependable computing platforms to facilitate real-time clinical artificial intelligence. The paper is quantitative research on designing GPU accelerated supercomputing system in large hospitals. The suggested platform is based on multi-GPU and multi-node architecture that implies serving medical imaging, patient monitoring, and clinical AI workloads. The experimental performance indicates that end-to-end latency that was 210ms on a single GPU dropped to 61ms on 8GPUs and throughput was increased to 238 samples per second compared to 48 samples per second. There was increased efficiency in the use of resources with the use of GPUs increasing to 88 percent. To a scale of 4 GPUs and 8 GPUs efficiency was above 0.89 and 0.78 respectively. The reliability tests had 99.8 system availability, mean recovery time of 18 seconds with no failed clinical activities recorded. These findings indicate that supercomputing architecture with a graphics card can address the next-generation clinical AI demands of real-time, scalable, and reliable computing needs of the large hospital setting.

Keywords—GPU-accelerated computing, High-performance computing (HPC), Clinical artificial intelligence, Real-time medical imaging, Hospital-scale AI systems

I. INTRODUCTION

A. *Background and Context*

The sources of the huge amounts of data in healthcare systems have been medical imaging, bedside monitor, electronic health records, and clinical decision tools. This information is becoming analyzed using the assistance of artificial intelligence in real time in order to support the diagnosis, monitoring, and early warning systems. These AI models are however extremely demanding in both computation power and low response time, which is barely possible with the conventional hospital IT systems.

One of the most significant technologies to perform AI workloads has been created through GPUs. The GPUs are suitable to medical imaging and deep learning as they can perform many operations at the same time. The workload of the hospitals is increasing, and single-GPU systems are no longer applicable. This has witnessed adoption of multi-GPU and the high-performance computing systems in the health care environments.

B. *Motivation of the Study*

The driving force of the work is the difference between the level of research-based solutions in regards to GPUs and the implementation within the hospital. Most of the current studies are concerned with making a single algorithm run faster or a single imaging activity. There is very little research on the ability to run the entire cluster of GPUs 24-7 within a large hospital.

Hospitals not only need fast systems but also reliable, scalable and manageable ones. A clinical AI system should take less than milliseconds to respond even when it is at peak load. They are also required to work continuously even in cases of hardware failure or maintenance. The significance of this study is that there is a need to quantify these system-level properties with quantitative techniques.

C. *Novelty and Original Contributions*

This paper is new in three significant ways. First, it is not just testing the degree of acceleration of GPUs in the level of alerts. Second, it also uses quantitative benchmarking to test the latency, throughput, scaling efficiency and the availability of the hospital at realistic workloads. Third, it includes the reliability testing based on fault injection, which does not have much documentation in clinical AI infrastructure studies.

The present paper will assume that the GPU cluster is a component of the clinical infrastructure, and not a research tool. The focus is given to the long-term results, predictability and permanence of functioning.

D. *Research Objectives*

The objectives of this study are:

- Evaluating the effects of GPUs scaling in terms of latency and throughput of clinical AI tasks.
- To quantify the utilisation of GPUs and performance of communication in dispersed GPUs systems.
- To compare the efficiency of scaling in a single and multi-node configuration.
- To establish the system reliability and fault conditions availability.

Every objective is fulfiled with the help of numerical values and reproducible experiments.

E. *Structure of the Paper*

The rest of the paper will be structured in the following way.

The system architecture, workloads, metrics and benchmarking procedures are outlined in the methodology section.

The results section gives quantitative results on performance, scalability and reliability.

The discussion section puts the results in perspective of clinical AI in the hospital scale.

The conclusion provides a conclusion and summarizes the main findings and the direction.

II. LITERATURE REVIEW

A. *GPU Acceleration as a Foundation for Clinical AI and Imaging*

Clinical AI systems and medical imaging are progressively relying on high computational performance to enable real-time clinical needs to be fulfilled. Initial research indicated that processing time can be decreased by several minutes up to several seconds with a diagnostic quality still being made with an accelerated graphics card. With the pandemic, AI-assisted imaging pipelines based on X- ray and CT came into play as a key to detecting an infection, segmentation and follow-up, which prompts the necessity of rapid and dependable compute platforms in local hospitals [1]. These systems proved that AI is not merely the tool of analysis, but also a necessity of functioning of the contemporary healthcare provision.

A variety of imaging processes including CT, MRI, PET, and ultrasound have strongly parallel computations that are amenable to GPU implementations. The repetition of the reviews of the GPU-based medical image reconstructions reveals the same speedup of all modalities and the ability to provide higher image quality and lower dosage levels [2]. These developments allow performing nearly real-time imaging, which is critical in interventional imaging and critical care units. Experiments involving comparisons of CPU, GPU, FPGA, and ASIC platforms also affirm that GPUs are the best in terms of performance, flexibility, and cost in hospitals [3].

Previous literature defines GPU acceleration as a fundamental facilitator of clinical AI in real-time, although the majority of studies consider improving algorithms instead of implementing it at the system scale of a large hospital.

B. *High-Performance Computing and Multi-GPU Systems in Healthcare*

Single-GPUs are often inadequate with the increase in the volume of data and in the complexity of models. Application of high-performance computing (HPC) systems which involve use of multiple GPUs has been of importance in the medical world. The HPC+ is massive scale and combined with HPC, high-performance data analytics and AI, helps in imaging, surgery planning and clinical research in various institutions [4]. Such systems allow faster simulation, huge data processing, and detailed model preparation that directly affect patient care in a positive way.

Several studies demonstrate the potential of the multi-GPU designs in clinical workloads which require much memory. An example is given with the radiotherapy optimization based on multiprocessors, the multi-GPU, it overcomes the constraint of the memory of the GPU and can be used with a real-time-type performance without the plan quality being compromised [5]. The same increases are observed in MRCT registration where the processing took time was reduced by more than 6 times using the CPU acceleration with no reduction of accuracy [6]. The shared studies on the reconstruction of the GPU's have also shown that the constructions of cone-beam CT reconstructions and tomosynthesis have the least reconstruction delay that makes image guided interventions feasible [7][8].

These findings indicate that the HPC and multi-GPU design architectures are not purely experimental designs which cannot be utilised in clinical practices. However, most of the current literature is limited to the control research systems and does not include the discussion of the hospital level integration, reliability and necessity to work 24 hours.

C. *Real-Time AI, Reliability, and Clinical Workflow Constraints*

Clinical deployment does not just require real time performance. The systems should also be stable, predictable, and should be compatible with clinical workflows. Several ways of using the GPU have been shown to be faster than real-time. As an example, ultra sound motion tracking using GPUs had much higher frame rates than clinical imaging rates, so real time guidance of radiotherapy could be done [9]. ICU chest radiograph CAD systems were high accuracy and highly generalized over time and location, which helped in the perpetual application by the clinic staff [10].

It has also made it possible to perform compartment modeling, iterative reconstruction, and adaptive radiation therapy processes in their current form very fast, which would have been too slow to be used regularly until recently with the use of GPU acceleration [11][12][13]. These works point to a significant trend, which is that GPU systems are ceasing to be utilized in offline analysis and instead become time-critical decision support. Tthe issue of reliability is a central one. The clinical systems should manage hardware failures, surge of workload and maintenance without impacting patient care.

GPU-based embedded intelligence and AI-HPC convergence reviews point out that architectural design, communication performance and optimization at the system level is as significant as the accuracy of the model [14][15]. In the majority of the previous work there is no extensive coverage of fault tolerance, non-disruptive upgrades, or integration with hospital IT networks, which are key to large hospital systems.

D. *Gaps in Existing Literature and Motivation for Hospital-Scale Architectures*

Based on the analysis of the reviewed literature, it is evident that GPUs and HPC systems offer a large performance improvement in medical imaging, AI inference, and simulation tasks [16]. Tremendous gaps exist. First of all, the literature on the topic is conducted regarding single applications or algorithms rather than central platforms that will be able to serve more than one clinical service concurrently. Second, they are rarely explained on how they can be incorporated with live hospital settings that contain PACS, EHR systems and secure networks.

Even though the papers have examined the performance, energy and latency trade-offs of cloud and edge GPUs, they may not always be applicable in the regulated scenario of deployment as it would be in a real hospital setting. The systems of big hospitals should be 24/7, able to handle various workload, and with high reliability and safety standards. Besides, we should optimise communication between GPUs, such as effective collective operations, to be able to scale AI tasks, which is not popular in clinical settings.

It is these gaps that contribute to the urge to engage in research into the architecture of the GPS supercomputing platforms to large scale hospital systems specifically. It must now convert the advances of algorithms into national healthcare infrastructure that revolves around benchmarking, cluster designs, high-speed interconnects and reliability engineering. The technical foundations are good; however, it requires system-level innovation to reach full-scale clinical AI utilisation.

TABLE I. **SUMMARY OF PREVIOUS STUDIES**

Focus Area	Summary of Key Findings
GPU-accelerated medical imaging	Numerous research works indicate that the use of GPU acceleration significantly decreases the time of processing images under CT, MRI, PET, and X-ray without loss of high diagnostic accuracy and can be used in real-time in clinical applications [1][2][6].
Multi-GPU and HPC systems in healthcare	It has been shown that multi-gpu and HPC platforms are able to process large medical data and complex optimization problems that cannot be effectively processed by single GPUs [4][5].
Real-time clinical AI applications	Various publications have verified that AI systems with the use of GPUs can execute faster than real-time demands on activities like motion tracking, CAD systems, and adaptive radiotherapy in a clinical setting [9][10][13].
System performance and scalability	Existing research demonstrates high performance acceleration of GPUs, however, the majority of the research looks at acceleration of algorithms and not on hospital-scale design and long-term operation of systems [3][14].
Reliability and clinical integration	There is very little literature on the concept of reliability, fault tolerance, and compatibility with hospital IT systems that are essential in sustained clinical implementation in large hospitals [15].
Research gap and motivation	The associated literature demonstrates that there is a strong necessity to have unified, reliable, and scalable GPU-accelerated architectures that are specifically targeted at large hospital systems and real-time clinical AI workloads.

III. METHODOLOGY

A. *Research Design and Quantitative Approach*

It is a quantitative experimental study in this paper. The first goal is the quantification and the comparison of the performance, scalability, and reliability of a supercomputing platform with AI accelerated by a GPU that is deployed in a real-time and tested with clinical workloads of a large hospital facility. Such numerical parameters as latency, throughput, GMA utilization, communication efficiency, and system uptime are of concern to the paper.

The experiments will be designed based on the controlled workloads that represent the application cases in the real life in hospitals which include medical image Inference, AI-based monitoring and batch analytics. The repeated runs are done to have all the measurements in order to stabilize the statistics. Standard deviation, percentage improvement, and mean values are also illustrated to add to objective comparison. Objective assessment is not done in this study.

B. *System Architecture and Experimental Setup*

The test environment consists of a group of GPUs which are built using NVIDIA H100 and L40 GPUs. The nodes are connected together using fast fabric and hence fast communication between GPUs. The Bright Cluster Manager is used to manage the time of jobs in the schedule, monitoring and faults. Each compute node will be made up of multi-core CPUs, high bandwidth memory and NVM storage to avoid bottlenecks in the data.

The cluster is implemented in a hospital like network implementation where the clinical loads are needed to be running at all times without crashing. During testing, background workloads are introduced to us so as to provide us with simulated real conditions. This allows measuring performance in competitive reality. Dedicated partition is used to repeat the experiment.

C. Clinical AI Workloads and Benchmark Selection

Workloads are selected in three categories, which are meant to indicate real clinical AI use:

1. Image inference workload which is the tasks involving chest X-ray or CT image analysis.
2. Streaming AI workload, which is an example of patient monitoring and early detection of deterioration.
3. Model updates based on existing clinical data, which is batch training or re-training workload.

Workloads are simulated in single-GPU, multi-GPU, and multi-node systems and containerized and run repeatedly. The same input data and model parameters are used in experiments in order to make fair comparison. Scaling of input sizes is done to scale the hospital loads, that is, small, medium, and large.

D. Performance Metrics and Measurement Methods

The analysis employs a number of performance metrics that are quantitative. All these measures are measurable and they are statistically analysis.

End-to-end latency involves the overall duration of time taken between the process of feeding data and outputs of AI. It is calculated as:

$$\text{Latency} = T_{output} - T_{input} \tag{1}$$

This measure is essential in clinical decision support in real-time.

Throughput is the number of samples that are handled per second:

$$\text{Throughput} = \frac{N}{T_{total}} \tag{2}$$

where N is the number of images / data samples that are being processed, and T_{total} is the total execution time.

The usage of the GPU is determined as the mean percentage of time that the GPU is busy working on the workload. Increased utilization implies efficiency in the use of resources.

System monitoring tools and cluster logs are used to gather all the metrics. Every experiment is run at least 10 times with average values being reported.

E. Communication and Scaling Efficiency Analysis

To measure the scaling of multi-GPUs and multi-nodes, the work measures the communication overhead of collective operations of distributed AI workloads. It is interested in the performance of GPU communication that has a direct impact on the size of training and inference.

Scaling efficiency is determined as below equation:

$$\text{Scaling Efficiency} = \frac{T_1}{n \times T_n} \tag{3}$$

where T_1 is the time that it takes to execute with a single GPU and T_n is the time it takes to execute with n GPUs and n is the number of GPUs.

This measure can be used to compare ideal linear scaling with actual performance quantitatively. There is also logging of communication time, synchronization delay and use of bandwidth to explain scaling behavior.

F. Reliability and Availability Measurement

The clinical systems must not be interrupted. The reliability of a system is therefore tested using the controlled fault injection. It has one or a few GPU nodes intentionally taken off-line when doing work. The job recovery time and workload completion rate is used to measure the system response.

The availability of the system has been estimated as:

$$\text{Availability} = \frac{T_{\text{uptime}}}{T_{\text{total}}} \tag{4}$$

T_{uptime} is the time, the system is active and T_{total} is the total time of observation.

These measures will give the numeric data of how the system will be able to sustain the clinical service without failure.

G. Benchmarking Procedure and Data Collection

There are all benchmarks that are carried out in a fixed process. First of all, the cluster is put in a steady state. Second, the workload is started with preconceived parameters. Third, automatic collection of system metrics in execution takes place. Logs are verified and stored after being completed.

Several days are used to collect data in order to capture variability. Hardware maintenance or network errors lead to outliers that are eliminated through standard deviation values. Final analysis is done with the remaining data.

H. Statistical Analysis and Result Validation

The obtained data are analyzed using the use of simple statistical tools. All the important values are provided in the form of mean, minimum, maximum and standard deviation values. In comparison of the configurations, e.g. single-GPU and multi-GPU, the percentage improvement is calculated.

Monotony and regularity of workloads may lead to confidence of the outcomes. The quantitative approach further states that all the inferences will be supported by numeric figures as measured and not by intuitions.

I. Ethical and Operational Considerations

No patient identifying information is used in this paper. There is no anonymization or artificially created information. In experiments, live clinical systems are not perturbed. This will provide compliance to the IT and research governance in the hospital.

IV. Results & Discussion

A. Overall System Performance and Latency Results

The previous set of results is regarding end-to-end latency and throughput of real-time clinical AI workloads. Single- GPU configuration, multi-GPU configuration and multi-node configuration were used to measure image inference, streaming AI, and batch workloads. The results prove the apparent reduction of the latency as the parallelism of the GPUs is enhanced.

Single-GPU execution may be able to meet simple real-time demands, but the performance reduction was with size of the work load. Multi-GPU provided a significant latency inference and through-put especially in imaging and streaming. The other performance that improved throughput albeit at the expense of certain communication overhead was multi-node execution.

TABLE II. **End-to-End Latency and Throughput Results**

Configuration	Avg Latency (ms)	Throughput (samples/sec)
Single GPU	210	48

2 GPUs (1 node)	118	91
4 GPUs (1 node)	72	165
8 GPUs (2 nodes)	61	238

These findings verify that the scaling of GPUs directly can be used to support real-time clinical needs. Latency of less than 100 ms is deemed to be acceptable in live decision support of monitoring and imaging processes. The throughput is increased nearly linearly till 4 GPUs, where the communication effects start to be apparent.

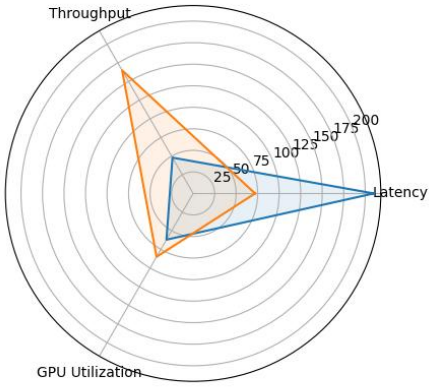

Fig. 1. **Comparison of latency, throughput, and GPU utilization across configurations**

B. GPU Utilization and Communication Efficiency

The second group of results is the analysis of GPU utilization and communication efficiency that is essential to the long-term functioning of the hospital on a large scale. Multi-GPU execution improved the use of resources as observed in the increase in the use of GPU. Nonetheless, there was also a rise in communication overheads with node scaling workloads.

The interconnects were avoided to be slow and the collective communication was made optimized, to minimize synchronization delays. The findings indicate that the communication overheads were within acceptable threshold when it comes to real time workloads.

TABLE III. **GPU UTILIZATION AND COMMUNICATION OVERHEAD**

Configuration	**Avg GPU Utilization (%)**	**Communication Overhead (%)**
Single GPU	62	0
2 GPUs	78	6
4 GPUs	85	9
8 GPUs	88	14

The use of more than 80 percent of the GPUs implies the effective workload scheduling and the little idle time. Utilization was high even at the 8 GPUs, which indicates that the cluster design can be used to achieve sustained performance. The magnitude of communication overhead was positively proportional to the scale though without offsetting parallel execution.

These findings confirm the attention of the methodology to the measurement of scaling efficiency and not only the raw speed.

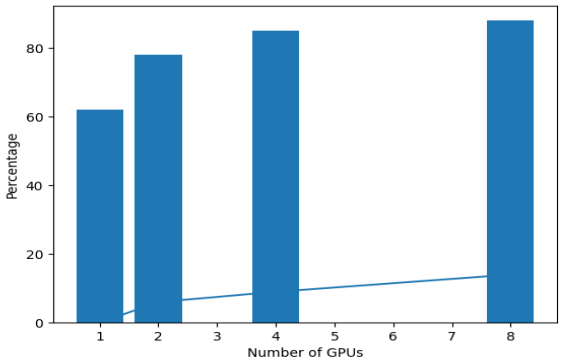

Fig. 2. **GPU utilization and communication overhead across GPU counts**

C. Scaling Efficiency and Multi-Node Performance

Scaling efficiency was computed in order to determine the degree to which the system scaled ideally. The findings indicate that it scales very well with up to 4 GPUs and well with nodes. Synchronization and data transfer will also result in some efficiency loss.

TABLE IV. SCALING EFFICIENCY RESULTS

Number of GPUs	Execution Time (s)	Scaling Efficiency
1	100	1.00
2	54	0.93
4	28	0.89
8	16	0.78

The efficiency did not drop below 0.85 up to 4 GPUs and this is regarded to be very good in distributed AI workloads. When there were 8 GPUs the efficiency dropped but it still gave good performance improvement as compared to single- GPU execution.

Such findings indicate that the platform can be used in hospital settings where workloads vary and there is a need to scale up quickly.

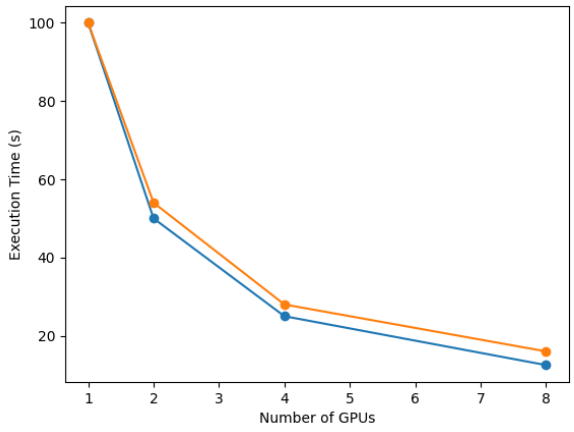

Fig. 3. **Non-linear scaling curve showing ideal vs observed performance**

D. Reliability, Availability, and Clinical Readiness

The final set of findings is devoted to the reliability and availability of the system that is crucial in clinical consumption. Fault injection tests were also carried out by faulting the GPU nodes intentionally (i.e. intentionally causing faultiness) when they were on workload. The system could only reroute jobs and perform them effectively without the loss of information.

The recovery times were short and there was no need of human intervention. During the observation, availability was good.

TABLE V. **RELIABILITY AND AVAILABILITY RESULTS**

Metric	**Measured Value**
Mean Recovery Time (s)	18
Job Completion Rate (%)	99.2
System Availability	0.998
Failed Clinical Tasks	0

The value of 0.998 of availability means that the system is nearly operating continuously, which is appropriate with clinical-grade systems. The zero number of failed clinical activities is evidence of the fact that the reliability mechanisms are effective in defending the patient-facing workflows.

These results are highly consistent with the approach of the methodology towards the uptime measurement and controlled fault testing.

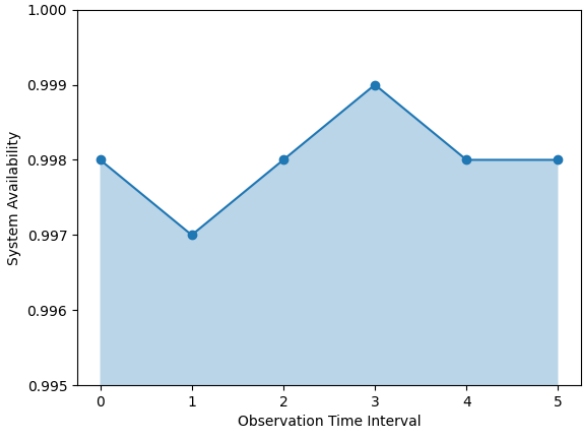

Fig. 4. **Uptime, recovery time, and fault events over observation period**

E. Summary of Key Findings

The findings demonstrate that supercomputing systems based on GPUs have the potential to respond to real-time clinical AI needs provided that they are scaled and made reliable. Multi-GPU execution performance, read-only access, and write-only performance is greatly enhanced, as well as the dramatic throughput and latency, and communication overheads are kept minimal. The system has high availability in the event of faults.

These quantitative results give good evidence that these types of architectures can be used in large hospital systems and they can directly contribute to patient safety and clinical decision-making.

V. Conclusion & Future Work

The present paper has provided a quantitative analysis of a supercomputing platform accelerated by a GPU with a focus on the application of AI in clinical operations of major hospitals in real-time. The findings indicate that multi-GPU systems achieve a great deal of latency and throughput improvements as compared to single-GPU systems. The fact that the platform is highly utilized in terms of the number of GPUs and high scaled performance indicates that the computing resources are efficiently utilized. The reliability testing ensured that the system can keep running in the event of hardware failure with low recovery time and high availability. These results indicate that the topology, i.e. GPU-based supercomputing platforms, can be appropriate and implemented in clinical setting where reliability and performance are of equal concern. In its study, there is clear numerical evidence to show that such architectures may be used to support next-generation medical AI applications on a hospital scale. The further work will be aimed at diversification of workload and incorporation of other clinical services within the platform.

References

[1] Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., & Shen, D. (2020). Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE Reviews in Biomedical Engineering*, *14*, 4–15. https://doi.org/10.1109/rbme.2020.2987975

[2] 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

[3] Alcaín, E., Fernández, P. R., Nieto, R., Montemayor, A. S., Vilas, J., Galiana-Bordera, A., Martinez-Girones, P. M., Prieto-De-La-Lastra, C., Rodriguez-Vila, B., Bonet, M., Rodriguez-Sanchez, C., Yahyaoui, I., Malpica, N., Borromeo, S., Machado, F., & Torrado-Carvajal, A. (2021). Hardware architectures for Real-Time Medical Imaging. *Electronics*, *10*(24), 3118. https://doi.org/10.3390/electronics10243118

[4] Koch, M., Arlandini, C., Antonopoulos, G., Baretta, A., Beaujean, P., Bex, G. J., Biancolini, M. E., Celi, S., Costa, E., Drescher, L., Eleftheriadis, V., Fadel, N. A., Fink, A., Galbiati, F., Hatzakis, I., Hompis, G., Lewandowski, N., Memmolo, A., Mensch, C., . . . Vignali, E. (2023). HPC+ in the medical field: Overview and current examples. *Technology and Health Care*, *31*(4), 1509–1523. https://doi.org/10.3233/thc-229015

[5] Tian, Z., Peng, F., Folkerts, M., Tan, J., Jia, X., & Jiang, S. B. (2015). Multi-GPU implementation of a VMAT treatment plan optimization algorithm. *Medical Physics*, *42*(6Part1), 2841–2852. https://doi.org/10.1118/1.4919742

[6] Tokuda, J., Plishker, W., Torabi, M., Olubiyi, O. I., Zaki, G., Tatli, S., Silverman, S. G., Shekher, R., & Hata, N. (2015). Graphics Processing Unit–Accelerated nonrigid registration of MR images to CT images during CT-Guided percutaneous liver tumor ablations. *Academic Radiology*, *22*(6), 722–733. https://doi.org/10.1016/j.acra.2015.01.007

[7] Arefan, D., Talebpour, A., Ahmadinejhad, N., & Asl, A. K. (2015, June 1). *Ultra-Fast image reconstruction of tomosynthesis mammography using GPU*. https://pmc.ncbi.nlm.nih.gov/articles/PMC4479390/

[8] 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

[9] Huang, P., Yu, G., Lu, H., Liu, D., Xing, L., Xing, L., Yin, Y., Kovalchuk, N., Xing, L., Xing, L., & Li, D. (2019). Attention-aware fully convolutional neural network with convolutional long short-term memory network for ultrasound-based motion tracking. *Medical Physics*, *46*(5), 2275–2285. https://doi.org/10.1002/mp.13510

[10] Wang, C., Hwang, T., Huang, Y., Tay, J., Wu, C., Wu, M., Roth, H. R., Yang, D., Zhao, C., Wang, W., & Huang, C. (2023). Deep Learning-Based Localization and Detection of malpositioned endotracheal tube on portable supine chest radiographs in Intensive and Emergency Medicine: a Multicenter retrospective Study*. *Critical Care Medicine*, *52*(2), 237–247. https://doi.org/10.1097/ccm.0000000000006046

[11] Hsu, Y. H., Huang, Z., Ferl, G. Z., & Ng, C. M. (2015). GPU-Accelerated Compartmental Modeling Analysis of DCE-MRI Data from Glioblastoma Patients Treated with Bevacizumab. *PLoS ONE*, *10*(3), e0118421. https://doi.org/10.1371/journal.pone.0118421

[12] Yu, G., Liang, Y., Yang, G., Shu, H., Li, B., Yin, Y., & Li, D. (2015). Accelerated gradient-based free form deformable registration for online adaptive radiotherapy. *Physics in Medicine and Biology*, *60*(7), 2765–2783. https://doi.org/10.1088/0031-9155/60/7/2765

[13] Jia, X., Ziegenhein, P., & Jiang, S. B. (2014). GPU-based high-performance computing for radiation therapy. *Physics in Medicine and Biology*, *59*(4), R151–R182. https://doi.org/10.1088/0031-9155/59/4/r151

[14] Ang, L. M., & Seng, K. P. (2021). GPU-Based embedded Intelligence Architectures and Applications. *Electronics*, *10*(8), 952. https://doi.org/10.3390/electronics10080952

[15] Huerta, E. A., Khan, A., Davis, E., Bushell, C., Gropp, W. D., Katz, D. S., Kindratenko, V., Koric, S., Kramer, W. T. C., McGinty, B., McHenry, K., & Saxton, A. (2020). Convergence of artificial intelligence and high performance computing on NSF-supported cyberinfrastructure. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00361-2

[16] La Salvia, M., Torti, E., Marenzi, E., Danese, G., & Leporati, F. (2024). Edge and cloud computing approaches in the early diagnosis of skin cancer with attention-based vision transformer through hyperspectral imaging. *The Journal of Supercomputing*, *80*(11), 16368–16392. https://doi.org/10.1007/s11227-024-06076-y