ISSN: 2754-6705

Journal of Mathematical & Computer Applications

SCIENTIFIC Research and Community

Review Article

Open Access

Optimizing GPU-Intensive Workloads with Converged Infrastructure

Sriramaraju Sagi

[1]NetApp, USA

ABSTRACT

This research paper delves into the utilization of convergent infrastructure platforms to enhance the performance of applications that heavily depend on GPUs. The analysis examines the advantages, challenges and strategies involved in leveraging convergent infrastructure for scaling and hosting GPU workloads. Furthermore, a detailed case study is provided to illustrate the benefits and valuable insights gained through the implementation of convergent infrastructure platforms for GPU applications. Converged infrastructure platforms bring together computing, storage and networking resources into a platform simplifying the deployment and management of IT infrastructure. These platforms offer advantages such as optimized utilization of GPUs reduced ownership costs, streamlined hardware and software management improved scalability for applications enhanced security features, superior performance in cloud environments, flexibility, and stability. Successful implementations of convergent infrastructure have been seen in industries, like institutions, scientific research organizations and video rendering companies. Hence it is crucial to select and deploy a converged infrastructure platform that can effectively manage the needs of GPU intensive applications while ensuring optimal performance.

***Corresponding author**

Sriramaraju Sagi, NetApp, USA.

Received: December 01, 2023; **Accepted:** December 12, 2023, **Published:** December 20, 2023

Keywords: Converged Infrastructure, GPU, Instance, Cluster, System Architecture, Management, Optimization, Case Study

Introduction

Many organizations and individuals face challenges when it comes to scaling and hosting applications that require a lot of GPU processing power. These applications often need GPUs to work which can be expensive and difficult to manage. Additionally scaling these applications can be a process as it involves distributing the workload across different GPUs. It's important to have expertise in distributed computing and parallel processing for this task. As a result companies and individuals may need to invest in hardware, software and dedicated personnel in order to effectively expand and manage GPU applications. This issue is significant because these types of applications are becoming increasingly prevalent, in fields such as research, financial modeling and video rendering. The ability to scale and host these applications efficiently is crucial, for staying competitive.

Converged infrastructure is a system that brings together computing, storage and networking capabilities into one platform. Its main goal is to simplify the implementation and management of IT infrastructure. This type of infrastructure can be really helpful when dealing with the challenges of scaling and hosting applications that heavily rely on GPUs. It offers a preconfigured and optimized platform that can handle the increased demands of these applications. Moreover converged infrastructure plays a role, in managing the complexity associated with deploying and configuring components. This simplifies the process of scaling resources based on requirements.

In industries there is a growing need for scaled and hosted GPU applications. However effectively managing, improving performance and optimizing these applications can be quite challenging. That's why it's important for enterprises to carefully select and implement a converged infrastructure platform. The purpose of this paper is to provide an overview of the software components that need to be integrated into the software stack of a multi instance computing cluster. By examining considerations and specified requirements, in detail organizations will be able to make informed decisions when transitioning their infrastructure into a multi instance cluster setup.

The use of a instance computing cluster is a cost effective and efficient option for hosting applications that heavily rely on GPU processing. By running program instances on a cluster organization can optimize their GPU resources and make the most out of each one. This approach also offers flexibility in workload management as different instances can be assigned to projects or departments based on their needs. Furthermore, employing instance clusters can improve overall performance and reduce latency by distributing workloads across multiple nodes enabling parallel processing and faster execution times. To achieve performance, scalability and efficiency when hosting GPU applications it is crucial to carefully select and integrate appropriate software components in the software stack of the multi-instance computing cluster.

Implementing a converged infrastructure platform simplifies the deployment and management process for applications that require GPU usage. This solution provides enterprises with an effective approach.

When it comes to tasks that require handling amounts of data the integration of AI technologies, with GPU applications in a single architecture brings about better scalability, performance and versatility [1]. This combination allows businesses to utilize the power of GPUs for processing enhancing the speed of calculations. It empowers them to efficiently handle datasets and run AI algorithms at speeds. The integration of GPU applications, High Performance Computing (HPC) and Artificial Intelligence (AI) technologies in an architecture leads to improved performance, scalability and flexibility for data heavy workloads. With this integration researchers and data scientists can develop, train and deploy AI models in an HPC environment using AI development frameworks like PyTorch and Tensor Flow. Models can be optimized through solutions, hyperparameter optimization and cross validation. This enables the study of datasets and exploration of scenarios in domains such, as healthcare, finance, materials science and climate research. The convergence of these technologies has resulted in groundbreaking advancements in applications including climate modeling weather prediction energy exploration material science finance (risk analysis and fraud detection) well as healthcare (drug discovery and patient care).

Integrating high performance computing (HPC), with Artificial intelligence (AI) has the potential to bring advantages. However there are challenges that need to be tackled. To overcome these obstacles, strategic measures like nurturing talent implementing practices ensuring data governance and staying informed about new technologies and legislation can be employed. By adopting this approach, across sectors and domains of research we can enhance decision making, facilitate scientific breakthroughs and foster innovation.

Literature Review

The sources provided emphasize the significance of converged infrastructure in enhancing the performance of workloads that heavily rely on graphics processing units (GPUs). The authors delve into the difficulties and potential advantages associated with integrating resources, such, as GPUs, CPUs and AI technologies to support tasks related to machine learning and deep learning. When it comes to applications that heavily utilize GPUs there are benefits to utilizing a converged infrastructure that combines GPUs, CPUs and AI technologies.

Companies and research institutions engaged in data science initiatives that require computing power often face the challenge of managing infrastructure. This can lead to difficulties in expanding their operations. The key findings of the study highlight the challenges encountered by companies and research facilities as they strive to grow their infrastructure [2]. It also explores options, between outsourcing IT services or developing a software suite. Moreover it outlines the steps involved in establishing a localized infrastructure and transitioning the software stack into a GPU cluster system.

The sources provided emphasize the significance of converged infrastructure in enhancing the performance of workloads that heavily rely on graphics processing units (GPUs). The authors explore the obstacles and potential advantages of integrating resources, like GPUs, CPUs and AI technologies to support machine learning and deep learning tasks. When it comes to applications that heavily utilize GPUs there are benefits to utilizing a converged infrastructure that integrates GPUs, CPUs and AI technologies. Businesses and research institutions engaged in data science initiatives face the challenge of managing infrastructure which hampers their ability to expand operations efficiently. A recent study sheds light on these challenges faced by businesses and research facilities as they aim to grow their infrastructure [2]. The study evaluates options, between outsourcing IT services or developing a software suite while outlining the steps involved in establishing a local infrastructure and transitioning the software stack into a scalable GPU cluster system.

In a research study they explored the benefits of using a programming model [3]. This model offers programmers flexibility. Empowers them with strong capabilities, for heterogeneous computing. With the advancements in GPU technology and its programmability GPU computing has emerged as a solution for various application domains. To achieve performance in GPU computing it is crucial to employ a parallel programming model. The suggested hybrid model combines GPU acceleration with the PGAS programming paradigm and has been proven to deliver performance improvements in application benchmarks. By utilizing this combined approach, the hybrid parallel programming model maximizes the potential of both CPUs and GPUs resulting in enhanced performance for GPU workloads.

In a research study the effectiveness of ScaleGPU, in running GPU applications with GPU memory sizes and improving performance was explored [4]. The study aimed to address the challenge of managing GPU memory by programmers and introduced ScaleGPU as a solution for high performance GPU programming that doesn't depend on memory sizes. Furthermore it highlighted the performance improvements achieved by ScaleGPU, which utilizes GPU memory as a cache for CPU memory providing programmers with a programming space to CPU memory.

The review of existing literature emphasizes the importance of an integrated infrastructure that can effectively handle GPU workloads by incorporating both CPU and GPU components. This integrated infrastructure should support processing, efficient task offloading, optimal performance, energy conservation, fast data pipelines, minimal communication delays, scalability and resource efficiency. The paper discusses the significance of each component in such a converged infrastructure for enhancing GPU workloads. In conclusion leveraging a converged infrastructure that integrates both CPU and GPU components is crucial, for optimizing tasks that heavily rely on GPUs.

Methodology

This document provides an overview of the 2021 benchmark applications. These applications were specifically designed to simulate real world models, in domains. The main goal of the study was to assess the performance improvements and processing times achieved by running sizes on an HPC cluster comprising eight Cisco UCS C240 M7 Rack Servers with NVIDIA A100 80G GPUs. In this architecture we also utilized NetApps all flash storage. The benchmark tests showcased enhancements in performance. Demonstrated linear scalability resulting from the integration of HPC and AI technologies. These applications span a range of fields, including weather modeling, radiation transport, in engineering and high performance geometric multigrid methods. They provide an evaluation of the performance achieved through the fusion of HPC AI technologies.

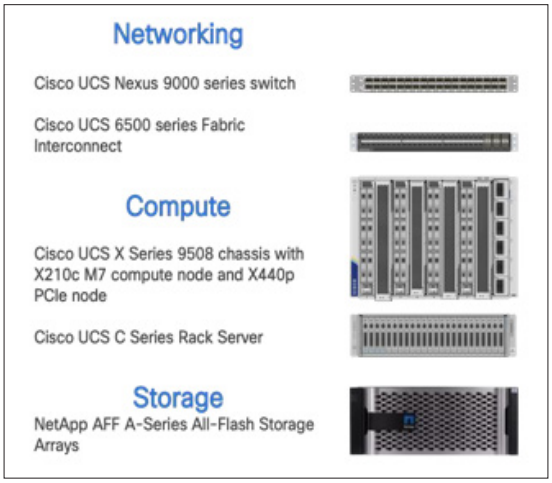

Figure 1: FlexPod Converged Infrastructure

In our research we utilized the FlexPod validated converged infrastructure platform, specifically designed to cater to AI workloads. This platform offers advantages, including deployment and management of AI workloads seamless integration, into AI ecosystems, efficient and straightforward operations faster implementation of AI with significant value gains and safeguarding the integrity of our AI infrastructure by ensuring system safety, data protection and application management. Moreover it demonstrates performance across various dataset sizes through benchmark tests that prove its linear scalability. Additionally Cisco Intersight provides management and automation capabilities for deployment time. The platform optimizes resource utilization while minimizing energy consumption. It streamlines operations by utilizing the NVIDIA HPC X software toolkit for validation purposes. NetApp tools such as DataOps Toolkit enable developers, data scientists and data engineers to efficiently manage data related tasks. Thorough testing of real world workloads is conducted alongside scalability assessments across sizes and application areas. Performance comparisons are made between CPU systems and those equipped with GPUs to gain insights, into solution capabilities.

In a nutshell, when high performance computing (HPC) and intelligence (AI) technologies are brought together with an infrastructure design that utilizes Cisco UCS servers, NVIDIA GPUs and NetApp all flash storage it brings significant improvements, in performance, scalability and flexibility for workloads that heavily depend on GPUs.

Technology Overview

Our approach combines the use of high performance computing and artificial intelligence technologies. This integration involves utilizing Cisco UCS servers, NVIDIA GPUs and NetApp all flash storage, within an infrastructure architecture.

FlexPod Datacenter

FlexPod is a solution for data centers that brings together computing, storage and networking components in an integrated system. This simplifies the implementation process. Provides the ability to handle increased demands. It ensures compatibility. Can handle workloads making it a flexible and efficient choice for modern data centers.

Cisco Unified Computing Systems

The Cisco Unified Computing System is a data center platform that integrates computing, networking, storage and virtualization components into an architecture. The goal of this integration is to optimize data center management improve efficiency reduce ownership costs and enhance organizational agility. The Cisco UCS servers play a role in the converged infrastructure by providing the computational power for high performance computing (HPC) and artificial intelligence (AI) workloads.

Cisco Intersight

Cisco Intersight serves as a management platform, for your infrastructure regardless of its location.It effectively addresses the management challenges presented by corporate sites and leverages the insights acquired through the utilization of Cisco Unified Computing System using Software as a Service.

Cisco UCS C240 M7 Rack Server

The Cisco UCS C240 M7 Rack Server improves upon the capabilities of the Cisco UCS rack server series by integrating, up to two 4th Generation Intel Xeon Scalable CPUs, with each CPU offering a maximum of 60 cores. With DDR5 technology this system can support a memory capacity of 8 terabytes for 2 processing units (CPUs) by utilizing 32 memory modules, each with a size of 256 gigabytes and a speed of 4800 mega transfers per second. It has a compact form factor of 2 Rack Unit. Is designed to accommodate either eight PCIe 4.0 slots or four PCIe 5.0 slots along with a LAN on motherboard (mLOM) slot. Additionally it can house up to five GPUs. Delivers performance and efficiency improvements that greatly enhance your applications overall performance.

Cisco Nexus 93600CD-GX

The Cisco Nexus 9300 GX switches leverage Cisco Cloud Scale technology. Represent the iteration in the fixed Cisco Nexus 9000 Series Switches lineup. These switches are capable of supporting high speed connections up to 400 Gigabit Ethernet. This platform fulfills the demand for performance energy space saving switches, in networking infrastructure as applications utilizing Artificial Intelligence and Machine Learning continue to grow. These switches cater to industries such, as network edge 5G, Internet of Things Professional Media Networking platform and Network Functions Virtualization. They are designed to support 100G and 400G fabrics in configurations used by mobile service providers.

NetApp AFF A-Series Storage

The NetApp AFF a Series Storage offers performance while maintaining a range of features. These systems provide data management and security capabilities that're well suited for enterprise use. They can handle NVMe technologies, including NVMe attached SSDs. Frontend NVMe over Fibre Channel (NVMe/FC) host communication. With their outstanding performance characteristics they are a choice for managing the most demanding workloads and applications. By upgrading to a NVMe/FC SAN infrastructure these systems can effectively handle applications with improved response times without any interruptions or the need for data migration. Furthermore as enterprises increasingly adopt a "approach there is a growing demand for data services at both, on premises data centers and cloud environments. Therefore it is essential, for all flash storage systems to offer data services, integrated data protection techniques, effortless scalability options and enhanced performance levels while seamlessly integrating with applications and cloud platforms. The earlier flash systems were not equipped with the performance capabilities to handle these emerging workloads.

Software Components

Component	Software version
Cisco Intersight	SaaS platform
Cisco UCS Server Firmware	4.3(2.230207)
Host OS	Ubuntu 22.04 LTS
Mellanox ConnectX-6 NIC	22.36.1010
	MLNX_OFED 5.8-1.1.2.1
NVIDIA Tesla A100-80GB GPU	NVIDIA CUDA 12.2.2
	NVIDIA Driver 535.104.05
NetApp AFF A400	ONTAP 9.14.1
Cisco Nexus 93600CD-GX	NX-OS 10.3.3

Figure 2: Software Components

Tested Solution Reference Architecture

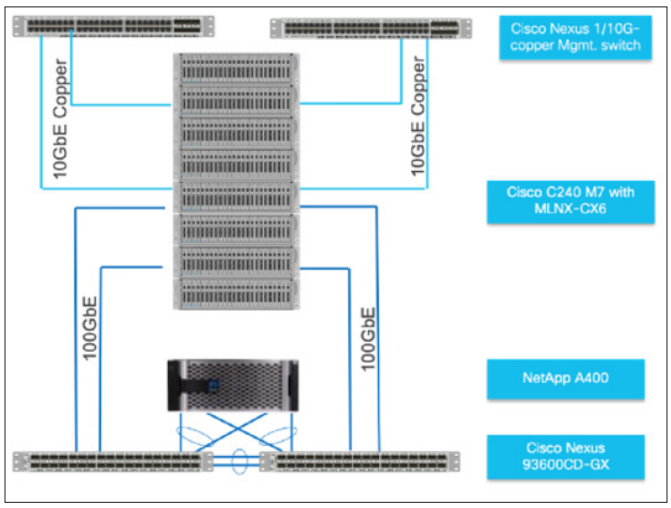

Figure 3: Infrastructure diagram

Results

In this section we present the results of tests conducted on datasets of sizes and application domains. We wanted to assess how well clusters can handle scaling so we measured the time it took to complete each dataset on a CPU system and then repeated the process on a system, with one or two GPUs. To conduct the CPU tests we utilized all cores on each node by using the argument " np 128*x" where x represents the number of nodes used as a multiplier. We collected performance metrics for the system including CPU utilization (measured using vmstat) GPU utilization (measured using nvidia smi and NVIDIA GPU exporter Grafana WebUI console) and network port utilization, on a Cisco Nexus switch connected to a Cisco UCS server and NetApp A400 Storage. These measurements were taken while running the test in a configuration. We performed tests with configurations comprising one, two four and eight nodes to demonstrate how well the system scales linearly. The test results show the duration in seconds required to complete each test scenario regardless of whether it's executed on the CPU or GPU. The workload dataset used in this approach has system requirements determined by both application needs and dataset size.

The goal is to evaluate how well a cluster performs by studying the impact of the number of nodes, cores and GPUs, on cluster sizes when executing a benchmark application. The dataset used is quite significant covering a range of units from 2048 to 32,768 ranks. The benchmark requires up to 14.5 terabytes of RAM. In this approach we. Organized the datasets in a NFS volume on NetApp A400 storage. To mount them on each node within the HPC cluster test we made use of NetApp FlexClone volumes. We performed tests on combinations to assess scalability, in HPC cluster workloads.

Mini Weather

The MiniWeather Mini App aims to replicate the dynamics observed in weather and climate. These dynamics involve compressibility, stratification and non hydrostatic behavior. Buoyant forces play a role causing disturbances, within a hydrostatic background state. The equations used in this code serve as the framework for fluid dynamics codes. This specific variant forms the foundation for weather and climate modeling.

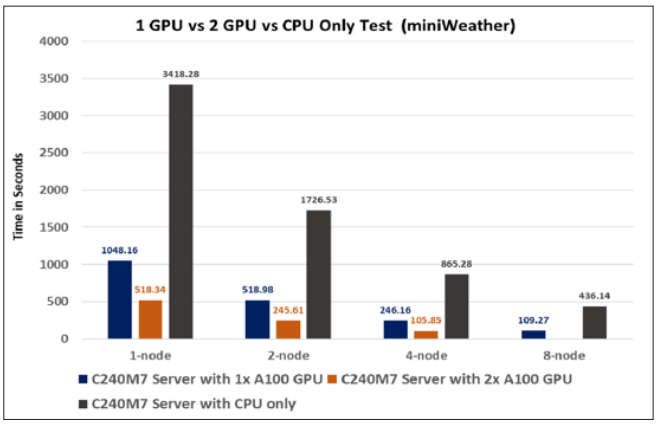

Figure 4: Cluster Scalability when Running Large Mini Weather Dataset

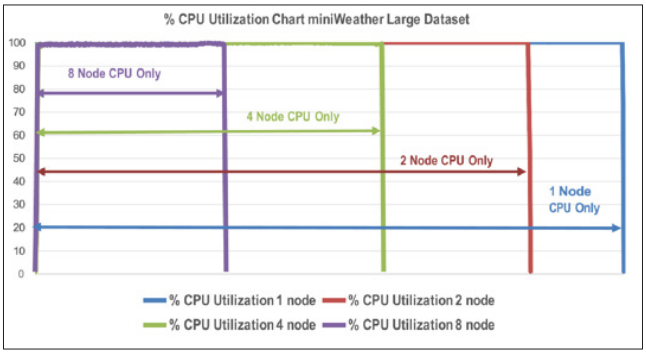

Figure 5: % Avg CPU utilization when running Large miniWeather Dataset only on CPU

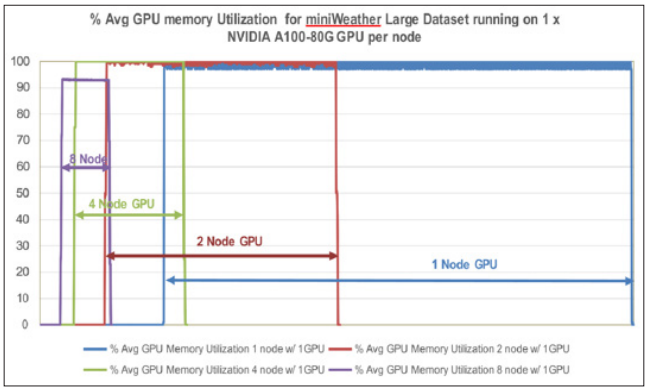

Figure 6: % Avg GPU memory utilization when running Large miniWeather Dataset on nodes w/ 1 NVIDIA A100- 80G GPU

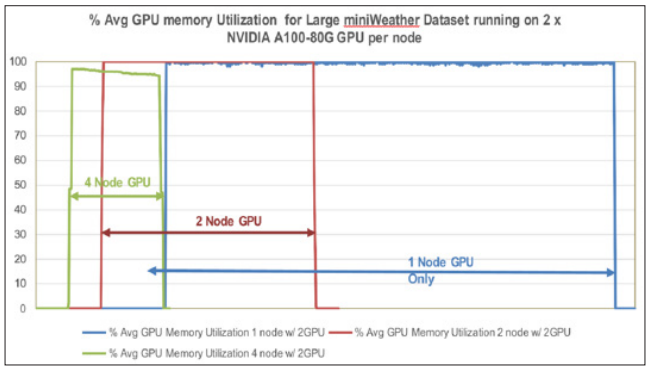

Figure 7: % Avg GPU memory utilization when running Large miniWeather Dataset on nodes w/ 2 NVIDIA A100- 80G GPU

Figure 8: Sample Grafana dashboard showing GPU statistics

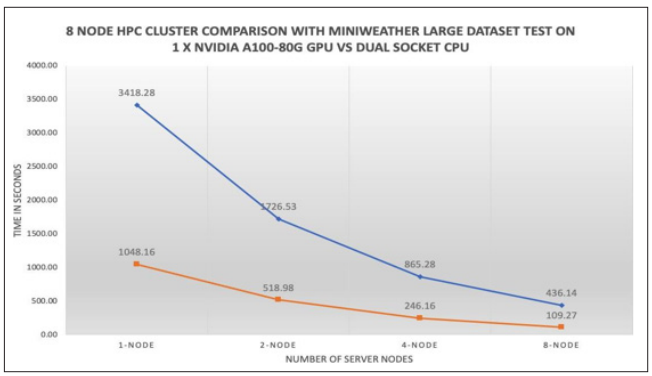

Figure 9: 8 Node HPC Cluster comparison with miniWeather larger dataset test on 1 x NVIDIA A100-80G GPU vs Dual Socket CPU

When executing the miniWeather dataset on an 8 x Cisco UCS C240 M7 node based HPC cluster as described in this research solution we obtained the following observations:

- The FlexPod HPC cluster showed scalability in both CPU only and GPU accelerated tests.
- Each nodes resources (computation, storage and network) were evenly utilized across the entire cluster during testing.
- The GPU achieved utilization with 100% usage along, with its memory.
- The test using the NVIDIA A100 80G GPU did not require any CPU cycles.
- The test can be finished four times faster by using one GPU, per node in an eight node cluster compared to conducting the test on an eight node cluster, without any GPUs.

Conclusion

The combination of High Performance Computing and Artificial Intelligence is a partnership that allows for computational capabilities. This enables management of data intensive challenges, with improved speed, accuracy and efficiency. To achieve performance, scalability and agility it is important to integrate CPUs and GPUs with high speed data access through a 100GbE network. By leveraging converged infrastructure solutions like the AFF A Series storage and Cisco Nexus switches businesses can enhance the efficiency of GPU workloads while ensuring top notch performance, scalability and compatibility. Researchers and data scientists can fully harness the potential of GPU accelerated computing by utilizing high performance hardware, robust data management features and seamless integration with cloud platforms. This empowers them to drive innovation and accelerate advancements. Organizations can meet the demands of GPU workloads by deploying converged infrastructure solutions that offer scalable modular adaptable designs while keeping things simple. To achieve performance in tasks reliant, on the graphics processing unit (GPU) it is crucial to have a well balanced setup that includes properly configured compute, network and storage components [5,6].

Each components importance stems from its capacity to integrate the advantages of Performance Computing (HPC) and Artificial Intelligence (AI) fields.

- CPUs play a role, in supporting a range of operations, including system administration controlling the flow of tasks and handling sequential processes. This makes them essential for both high performance computing (HPC) and intelligence (AI) infrastructure.
- When it comes to tasks like learning and scientific simulations GPUs are extremely valuable due to their remarkable parallel processing capabilities. They greatly enhance the speed of calculations.
- Task offloading involves distributing tasks between CPUs and GPUs. CPUs can efficiently transfer workloads to GPUs resulting in improved efficiency and faster processing.
- The combination of CPUs and GPUs creates an high performing computing environment of effectively handling various types of tasks.
- In terms of energy efficiency CPUs generally exhibit power efficiency for workloads and play a significant role, in overall system administration. On the hand Graphics Processing Units (GPUs) commonly used in computing systems excel at computing throughput. By integrating both elements we can achieve energy efficiency while maintaining performance.
- For high performance computing (HPC) and artificial intelligence (AI) dealing with datasets is sometimes necessary to ensure data pipelines.
- A network, with a speed of 100GbE enables transmission of data ensuring communication between CPUs, GPUs and storage devices. This boosts the performance of the system by eliminating any obstacles. In performance computing (HPC) and intelligence (AI) tasks low latency is crucial as it reduces waiting time for data transfers and accelerates processing speed.
- Scalability: High speed networking allows for expansion of resources to accommodate growing workloads. You can effortlessly add compute nodes, GPUs or storage as needed.
- Efficient Resource Utilization: The CPUs and GPUs efficiently utilize their capabilities by transferring data, between them. This minimizes periods. Maximizes the overall optimization of system performance.

In this study we conducted tests, on applications designed for weather modeling (miniWeather) Nuclear Engineering. Radiation Transport (Minisweep) and Cosmology, Astrophysics, Combustion (HPGMG). We documented the suggested parameters to achieve

a balanced design among the compute, network and storage components. Additionally we demonstrated that as the cluster size increases from 1 to 8 nodes the solution shows scalability. By combining processing units (CPUs) and graphics processing units (GPUs) into an infrastructure we enable effective simultaneous processing, task delegation, optimal execution, energy conservation, rapid data transmission with minimal communication delay. This integrated approach allows us to harness the benefits of both HPC and AI simultaneously. It supports workloads leverages parallel processing power smartly distributes tasks optimizes resource efficiency and offers high performance computing capabilities. Furthermore, integrating high performance computing (HPC) and intelligence (AI) technologies into an architecture enhances efficiency expandability adaptability, in managing data intensive jobs.

References

1. Hardik Patel TP (2023) Scaling FlexPod for GPU Intensive Applications. Retrieved from Cisco.com https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/flexpod_gpu_aiml.html#:~:text=The%20FlexPod%20AI%20(Artificial%20Intelligence,high%2Dspeed%20data%2Dfabric.
2. M Uray, E Hirsch, G Katzinger, M Gadermayr (2021) Beyond Desktop Computation: Challenges in Scaling a GPU Infrastructure. arXiv DOI: https://doi.org/10.48550/arXiv.2110.05156.
3. Lingyuan Wang, Miaoqing Huang, Vikram K Narayana, Tarek El-Ghazawi (2011) Scaling scientific applications on clusters of hybrid multicore/GPU nodes. ACM International Conference on Computing Frontiers https://dl.acm.org/doi/10.1145/2016604.2016612.
4. Youngsok Kim, Jaewon Lee, Donggyu Kim, Jangwoo Kim (2014) ScaleGPU: GPU Architecture for Memory-Unaware GPU Programming. IEEE computer architecture letters https://ieeexplore.ieee.org/document/6559969/authors#authors.
5. Roberto R Expósito, Guillermo L Taboada, Sabela Ramos, Juan Touriño, Ramón Doallo (2012) General-purpose computation on GPUs for high performance cloud computing. https://onlinelibrary.wiley.com/doi/full/10.1002/cpe.2845.
6. Vetter JS, Glassbrook R, Dongarra J, Schwan K, Loftis B, et al. (2011) Keeneland: Bringing Heterogeneous GPU Computing to the Computational Science Community. Computing in Science and Engineering 13: 90-95.