GPU Offloading in MLOps: Navigating the Multicloud Ecosystem for Flexible AI/ML Deployments

DOI: 10.26314/GARR-Conf24-proceedings-13

GPU Offloading in MLOps: Navigating the Multicloud Ecosystem for Flexible AI/ML Deployments

Fulvio Risso[1], Giuseppe Zangari[2]

[1]Politecnico di Torino, [2]ArubaKube

Abstract. The demand for dedicated computational resources for AI tasks is surging. However, the substantial investment required for high-performance GPUs, coupled with in-creasing market scarcity and extended lead times, poses significant challenges. Simul-taneously, underutilization of these resources is a frequent issue. This paper proposes a solution leveraging cloud-native approaches to utilize unused resources available elsewhere, imposing minimal overhead on the originating cluster. We will present how an "origin" cluster can seamlessly peer with a "donating" cluster, offloading tasks to remote hardware with complete transparency and superior efficiency com-pared to current technologies

Keywords. GPU, Liqo, AI, Efficiency, Offloading

Introduction

In the ever-evolving landscape of artificial intelligence (AI), the demand for computational resources continues to surge, particularly with the advent of generative AI models. These models, which have gained considerable traction across various fields, introduce new challenges and requirements for efficient deployment. Understanding the computational needs inherent in AI and generative AI is crucial for optimizing performance and ensuring seamless integration into real-world applications.

At the core of AI development lies model training, a process that consumes significant computational power. However, it is essential to recognize that training models is just one facet of the multifaceted activities involved in deploying and managing machine learning operations (MLOps). Beyond training, tasks such as data preprocessing, model validation, and deployment orchestration demand substantial computational resources and operational expertise. Neglecting these aspects can lead to inefficiencies and hinder the scalability of AI systems.

Moreover, the emergence of generative AI, which enables machines to produce novel content such as images, text, and music, presents unique computational challenges. Generative models, such as GANs (Generative Adversarial Networks) and VAEs (Variational Autoencoders), require extensive computational resources during both training and inference stages. The complexity of these models, coupled with the need for large datasets and sophisticated optimization techniques, underscores the importance of robust computational infrastructure.

Conferenza GARR 2024 - Navigare la complessità - Selected Papers
Brescia 29-31 maggio 2024 - ISBN: 978-88-946629-3-1

Managing computational demands in inference tasks can be addressed through various strategies, including optimizing model architectures, implementing hardware acceleration techniques, and utilizing cloud-based solutions for scalable computing. However, the initial investment required for high-performance GPUs or dedicated hardware can be substantial, representing a significant upfront cost. This expenditure is often justified by the promise of enhanced computational power and accelerated AI model training.

Despite the allure of high-performance GPUs, underutilization issues frequently arise. GPUs may not operate at full capacity around the clock, leading to financial inefficiencies. Businesses must grapple with the challenge of maximizing GPU utilization to optimize their return on investment. Conversely, peaks in demand can overwhelm current GPU offerings, particularly in on-premises clusters, resulting in substantial backlogs for users such as scientists and students.

The solution proposed in this paper leverages unused resources available elsewhere, with minimal overhead on the originating cluster. This approach aims to address underutilization and demand peaks, thereby enhancing efficiency and scalability in AI operations.

1. Dynamic resource onboarding and cloud bursting

Dynamic allocation is a pivotal strategy for optimizing efficiency and cost-effectiveness in resource management. By dynamically allocating GPU resources in response to real-time demand, organizations can maximize utilization and minimize underutilization, thereby enhancing overall performance while controlling expenses.

However, implementing dynamic resource allocation presents its own set of challenges. One significant challenge is ensuring that Kubernetes clusters can be dynamically extended without necessitating a full reinstallation of the existing infrastructure (Marino et al. 2023). Creating a new cluster (e.g., with Terraform) dedicated to current AI/ML processing on different infrastructure (e.g., another cloud provider) to absorb demand peaks is often impractical. This approach requires replicating all applications and data from the originating cluster, leading to wasted time, resources, and increased costs and operational burdens.

A more effective solution involves adopting robust systems and processes capable of seamlessly integrating new, external GPU resources into the existing cluster. This approach allows the current cluster to extend geographically and integrate with other underutilized clusters that have free GPUs available at any given moment. Adding elasticity to the originating cluster enables it to dynamically offload GPU workloads across the infrastructure, scaling resources up or down in response to changing demands by leveraging available free GPU resources elsewhere.

Such a mechanism allows research infrastructures to optimize GPU utilization proactively, resulting in significant cost savings by avoiding unnecessary hardware investments and minimizing operational overhead. By aligning GPU resources with workload requirements in real-time, research institutions can achieve enhanced agility, resilience, and cost efficiency. This positions them as frontrunners in the competitive landscape of GPU-driven environments, and ultimately in innovation.

2. Kubernetes and Liqo

In this context, Liqo (Iorio et al. 2023), emerges as a cutting-edge technology that empowers organizations to harness the full potential of their Kubernetes clusters. Designed to facilitate seamless interconnection between Kubernetes clusters, Liqo plays a pivotal role in enabling efficient GPU offloading and resource optimization across distributed environments.

At its core, Liqo serves as a bridge, connecting Kubernetes clusters regardless of their location or underlying infrastructure. By abstracting cluster boundaries, Liqo creates a highly dynamic, unified, federated environment where resources, including GPUs, can be shared and utilized dynamically.

Liqo benefits from sophisticated orchestration capabilities that enable the efficient distribution of container workloads across disparate clusters. Leveraging advanced scheduling algorithms and network overlays, Liqo orchestrates the deployment and execution of GPU-accelerated applications, ensuring optimal resource utilization and performance.

One of Liqo's key strengths is its ability to operate seamlessly in multicloud environments, where Kubernetes clusters span multiple cloud providers as well as on-premises clusters. By abstracting cloud-specific complexities and providing a unified management interface, Liqo empowers organizations to leverage GPU resources across diverse cloud environments, driving agility and flexibility in their operations.

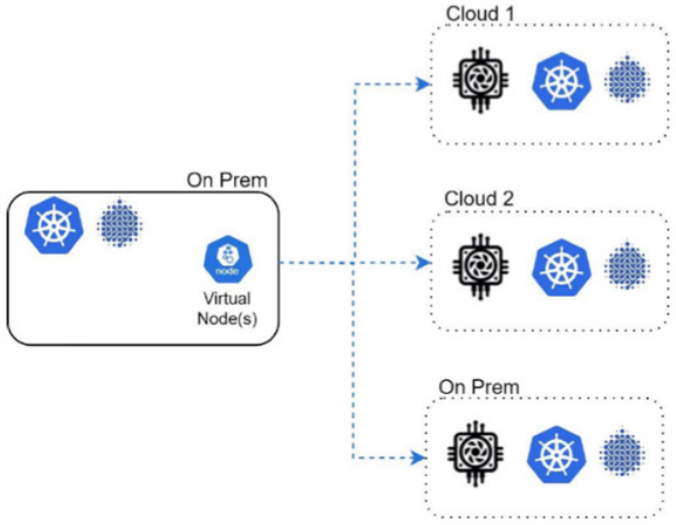

Fig. 1
An On Prem cluster is offloading AI tasks to GPU-equipped clusters located in different environments

3. An Example of Full Stack Solution

A comprehensive solution to the above problem is emerging that amalgamates Kubernetes, Liqo, Jupyter Notebooks, and MetaFlow[1]. Each component brings unique capabilities to the architecture, collectively forming a potent framework for advanced data science and AI endeavours.

Kubernetes serves as the cornerstone of this solution, offering powerful orchestration capabilities for containerized workloads. With features like automated scaling, self-hea-

ling, and resource management, Kubernetes streamlines deployment and management processes, ensuring optimal performance and efficiency across diverse computing environments. Liqo complements this by providing seamless interconnection between Kubernetes clusters, facilitating the creation of easily managed, highly dynamic federated computing infrastructures.

Jupyter Notebooks, renowned for their interactive and collaborative nature, provide a versatile platform for data exploration, visualization, and analysis. Supporting various programming languages and libraries, Jupyter Notebooks empower researchers to experiment, iterate, and share insights in real-time, fostering a culture of collaboration and innovation.

MetaFlow, on the other hand, offers a sophisticated workflow management system tailored for machine learning pipelines. With features like versioning, dependency management, and experiment tracking, MetaFlow simplifies the development and deployment of machine learning models, accelerating time-to-insight and facilitating reproducibility.

At the forefront of this transformative initiative is a prototypical implementation currently underway at Politecnico di Torino. Leveraging the combined capabilities of Kubernetes, Liqo, Jupyter Notebooks, and MetaFlow, researchers at Politecnico di Torino are exploring new frontiers in collaborative research and computational science.

However, the potential of this solution extends beyond individual institutions. By embracing a fully federated infrastructure that includes HPC main players and aggregates resources across a wider (e.g., national) research community, additional benefits can be realized. A federated approach enables the pooling of resources, facilitates cross-institutional collaboration, and maximizes the collective impact of research efforts, ultimately driving innovation and advancing knowledge across diverse domains.

4. Conclusions

The advancement of AI and ML technologies necessitates efficient and scalable computational resources, with GPUs playing a pivotal role. However, fluctuating GPU utilization presents challenges, including high ownership costs and resource underutilization. Liqo offers a solution by enabling dynamic GPU offloading in a multi-cloud environment, optimizing resource use and reducing operational expenses. By leveraging remote Kubernetes clusters, Liqo facilitates seamless and efficient ML deployments tailored to fluctuating computational demands, mirroring the “resource sharing” model popularized by services like Airbnb.

This paper addresses the operational inefficiencies in traditional GPU management within MLOps, demonstrating how Liqo provides a scalable, cost-effective, and low-code solution. It highlights the significant business impact of adopting a multi-cloud approach, emphasizing enhanced flexibility and the ability to adapt swiftly to research needs. This exploration showcases the competitive advantages offered by such technological agility, including cost reductions, operational efficiency, and high responsiveness to the needs of the academic community.

Bibliographic References

Iorio M., Risso F., Palesandro A, Camiciotti L., Manzalini A. (2023) Computing without borders: The Way Thowards Liquid Computing, IEEE Transaction on Cloud Computing (vol. 11, no. 3), pp 2820-2838

Marino J., Camiciotti L., Cheinasso F., Olivero A., Risso F. (2023), Enabling Compute and Data Sovereignty with Infrastructure-Level Data Spaces, ESAAM '23: Proceedings of the 3rd Eclipse Security, AI, Architecture and Modelling Conference on Cloud to Edge Continuum (October 2023), pp 77-85

Authors

Fulvio Risso fulvio.risso@polito.it

Fulvio Risso is full professor at Politecnico di Torino. Born in Saluzzo, Italy on November 15, 1971, he shares his birthday with the announcement of the Intel 4004 chip. Fulvio completed his BSc in Computer Engineering from Politecnico di Torino in July 1995 and got his PhD in Computer Engineering from the same institution in January 2000. His academic journey has been marked by significant contributions in the field of cloud computing, edge computing, network functions virtualization, and software-defined networking. He greatly contributed to open-source software, starting many successful project such as WinPcap, the de-facto packet capture library for Windows, and many others. He recently started the ArubaKube spin-off of Politecnico di Torino, where he serves as Chief Innovation Officer.

Giuseppe Zangari giuseppe.zangari@arubakube.cloud

Giuseppe Zangari (born in 1982) graduated from the Politecnico di Torino and holds an EMBA from the Graduate School of Management at Politecnico di Milano. He has held various leadership positions in global software organizations like Nokia and Pirelli, in Italian SMEs and in Politecnico di Torino, leading the development of business effective solutions with technologies ranging from IoT to cloud computing and AI. He is an expert of digital transformation, a startup mentor, and he also served as Innovation Lead. At ArubaKube, he is responsible for maximizing the software project's value, serving concurrently as Product and Business Development Lead.

Notes

[1]The described tools represent a possible example of deployment, with several alternatives (e.g., KubeFlow vs. MetaFlow) available to adapt to diverse application and usage scenarios.