TECHBYTE
Insights Into Research Computing

26 March 2024

SECURING ARTIFICIAL INTELLIGENCE INFRASTRUCTURE

AUTHOR:
RAKESH SABHARWAL
Founder ON DEMAND SYSTEMS PTE LTD

ON DEMAND
systems

26 March 2024

SECURING ARTIFICIAL INTELLIGENCE INFRASTRUCTURE (AI)

To discuss the topic of AI infra security, first, let's understand what constitutes Artificial Intelligence(AI) Infrastructure. There are multiple views on AI infrastructure. To some, it is the same as HPC infra, but to others, it could be very different. This can be seen at multiple levels.

1. Provisioning, De-provisioning, and System Management

Most large deployments would need tools like HPC to deploy, provision, de-provision, and manage the GPU/Compute nodes, e.g. OpenHPC, Bright Cluster Manager (now Base Command Manager) and Moab Cluster Manager. This layer would be very similar to HPC infrastructure, and we need a security component similar to HPC Infrastructure Security. Please take a look at the previous article about this.

https://www.odspte.com/wp-content/uploads/2024/03/Securing-HPC-Infrastructure.pdf

2. Overlay Workload and Data Management Platform

Traditionally, HPC workload managers such as PBS Pro and Slurm have been used for standard compute and container platforms. As more AI implementations move towards Kubernetes as a standard environment, new-age workload managers focusing on GPU scheduling have emerged, making the overall AI infrastructure very complex. There are other tools that provide workflow management, data management, experimentation management, etc.

SECURING ARTIFICIAL INTELLIGENCE INFRASTRUCTURE (AI)

3. Cluster Interconnect and User Networks

As AI advances at an unprecedented pace, the networks need to adapt securely to the colossal growth in traffic, which is transiting hundreds and thousands of GPU, CPU, and TPUs with trillions of transactions and Teraabits of throughput.

From a security perspective, both the user and cluster interconnect need protection.

4. Cluster Storage (Parallel Filesystem/Enterprise Storage)

Modern AI data platforms are capable of addressing needs across data capture, data preparation, model training, and model serving, thus making it of utmost importance to have enterprise security features embedded into the platforms.

Given this brief background, we will need a multilayered and multimodal approach to security.

All the aspects discussed in the previous article on securing HPC infrastructure are still relevant, but Kubernetes introduces an additional set of security challenges. Kubernetes security is a collection of best practices designed to keep the Kubernetes environment secure from internal and external threats and vulnerabilities.

SECURING ARTIFICIAL INTELLIGENCE INFRASTRUCTURE (AI)

- **K8Cluster and Control Plane Security**

The Kubernetes control plane manages the cluster, including scheduling, scaling, and monitoring. Securing the cluster includes securing the control plane components, such as the API server, etcd, and the Kubernetes controller manager, by enabling authentication, authorization, and encryption.

K8 Nodes are the worker machines in a Kubernetes cluster that run the containers. Nodes can be secured through the host operating system by configuring network security (nodes should be on a separate private network) and by securing the Kubernetes runtime environment. Removing unnecessary user accounts and ensuring that nothing runs as root are all best practices to consider when securing K8s nodes.

Another critical aspect is to use SecurityContexts to prevent pods and clusters from accessing the rest of the Kubernetes system.

- **Securing Private Repository**

Access to the repository shall be secured by integrating with the corporate directory and MFA. It is essential to use static analysis tools to find vulnerabilities in images across the registry and all artefacts.

SECURING ARTIFICIAL INTELLIGENCE INFRASTRUCTURE (AI)

- **Container/POD security**

A pod is a container used to run an application. Securing these applications means securing the pod. Kubernetes provides several security features to help secure applications. These features can limit resource access, enforce network policies, and enable secure communication between containers/PODs.

By enforcing network policies, you can segment your application's network traffic and add an additional layer of security.

- **Container Image Integrity and Security**

At various phases of the development lifecycle supply chain, containers take on many forms, each presenting its unique security challenge. This is because a single container may rely on hundreds of external, third-party components, diluting the level of trust at each phase.

- Image Integrity
- Image composition
- Image scanning for Known vulnerabilities

SECURING ARTIFICIAL INTELLIGENCE INFRASTRUCTURE (AI)

- **Secrets Management**

A “secret” is an object in Kubernetes that contains sensitive data such as passwords, certificates, and API keys. It is essential to integrate Kubernetes with Secure Vault. Secrets Vault provides secure storage and instant availability of secrets, SSH keys, certificates, API keys, and tokens.

- **Audit and Logging**

Setting up auditing in Kubernetes enables tracking changes to the Kubernetes API server and other Kubernetes components, helps identify unauthorized system changes and ensures compliance with security policies.

- **Compliance and Governance**

Addresses regulatory and organizational requirements by defining and enforcing compliance policies, protecting data, establishing incident response plans, and ensuring continuous compliance monitoring.

In conclusion, understanding AI Infrastructure security is fundamental to building robust and scalable AI platforms. By implementing robust security measures, organizations can ensure their AI workloads' confidentiality, integrity, and availability.

P.S. There is also a need for a more elaborate discussion on security in the context of GPU Cloud Service Providers.