Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240729

Beyond Docker: Enhancing vantage6 with Kubernetes for Federated Learning

Héctor CADAVID[a,1] and Cunliang GENG[a]
[a] *Netherlands eScience Center, The Netherlands*
ORCiD ID: Héctor Cadavid *https://orcid.org/0000-0003-4965-4243,*
Cunliang Geng *https://orcid.org/0000-0002-1409-8358*

Abstract. Vantage6, a powerful platform for privacy-preserving analysis in the life-sciences domain, employs Docker in its data nodes when performing tasks like federated learning. This tight bond to Docker poses challenges on infrastructure security, efficient computing resources usage, and integration of alternative container technologies. To overcome these challenges, we explored the integration of Kubernetes into vantage6 through a Proof-of-Concept (PoC) approach. The PoC designed and implemented a Kubernetes-based architecture for vantage6, which can be easily deployed on a single machine or a cluster. This PoC serves as the reference to accelerate the adoption of Kubernetes in vantage6.

Keywords. Federated learning, vantage6, Kubernetes, Docker, Containers

1. Introduction

Vantage6 or V6 [1] is a platform designed to facilitate privacy-enhancing techniques, such as federated learning (FL), for safeguarding data privacy in medical data analysis. In V6 a server orchestrates the entire process, managing tasks, algorithms, and results. The V6 Nodes, running on the organizations that host the data, autonomously execute Docker image-based algorithms on their local data and communicate intermediate results back through the V6 Server. This tight coupling between the V6 Node and Docker has raised concerns within the research community relying on this platform. First, some organizations are hesitant to enable a Docker daemon on their infrastructure due to security concerns associated with the root privileges it needs. Second, it limits the possibility of using alternative container technologies for the algorithms, some of which are more tailored to scientific applications, such as Singularity [2]. Third, it makes the algorithms run on a V6 Node constrained to the locally available computing resources (e.g., GPU). To overcome these challenges, we have explored the integration of Kubernetes with vantage6.

2. Methods and Results

V6's client-server architecture involves sophisticated mechanisms for task processing, data access, and algorithm execution across distributed environments. Consequently,

[1] Corresponding Author: Hector Cadavid; E-mail: h.cadavid@esciencecenter.nl

making a major refactoring on the platform architecture, in this case, a transition from an ad-hoc container orchestration model to a Kubernetes-based one, poses significant challenges and risks. Given this and considering the principles of experimentation in software engineering [3], a proof-of-concept (PoC) approach was adopted to validate the feasibility of the alternative architecture in a controlled environment. This way, the PoC served as a bridge between the original concept, the requirements elicited from the community, and the eventual implementation as an official project branch. Thus, we designed the Kubernetes-based architecture of V6 Node (Fig. 1) and implemented its PoC (https://github.com/vantage6/v6-on-kubernetes-PoC).

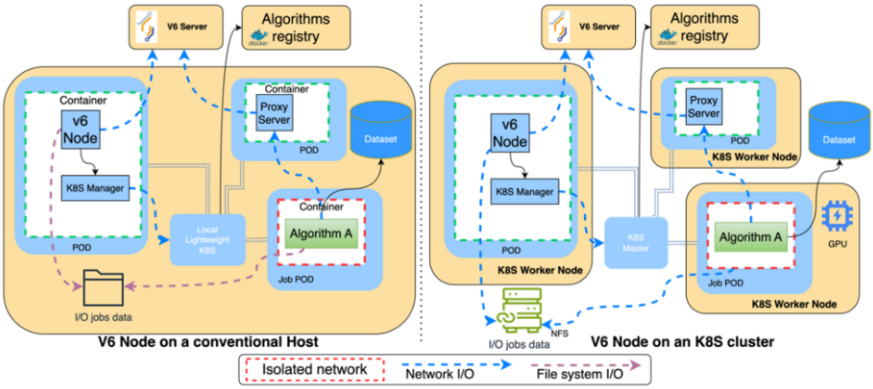

Figure 1. Kubernetes-based architecture of V6 Node and the two ways of deployment: within a conventional host (left) or on a Kubernetes cluster (right).

3. Discussions and Conclusions

The PoC architecture can be deployed either on a single machine or on an existing Kubernetes cluster, which overcomes some of the V6 limitations. First, it enables a more efficient distribution of the high-performance computing resources available in hospitals and medical research institutions. Moreover, it reduces the complexity of node-to-node communication across separate institutions and improves the compatibility with OCI-compliant images when employing V6 on FL-supported research endeavors. Finally, by delegating V6's container orchestration concerns to Kubernetes, the essential complexity of the platform on the node side was reduced significantly, thereby improving modularity and maintainability. The PoC is going to be used as a reference for enhancing vantage6 with Kubernetes.

References

[1] Moncada-Torres A, Martin F, Sieswerda M, Van Soest J, Geleijnse G. VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange. InAMIA annual symposium proceedings 2020 (Vol. 2020, p. 870). American Medical Informatics Association.

[2] Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PloS one. 2017 May 11;12(5):e0177459.

[3] Farley D. Modern Software Engineering: Doing What Works to Build Better Software Faster. Addison-Wesley Professional; 2021 Nov 16.