

**Copyright © 2021 IEEE**

**© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.**

# Container Orchestration on HPC Systems

Naweiluo Zhou<sup>1</sup>, Yiannis Georgiou<sup>2</sup>, Li Zhong<sup>1</sup>, Huan Zhou<sup>1</sup>, Marcin Pospieszny<sup>3</sup>

High Performance Computing Center Stuttgart (HLRS), Germany<sup>1</sup>

Ryax Technologies, Lyon, France<sup>2</sup>

Institute of Bioorganic Chemistry of the Polish Academy of Sciences,

Poznan Supercomputing and Networking Center, Poznan, Poland<sup>3</sup>

Email: naweiluo.zhou@hlrs.de, yiannis.georgiou@ryax.org

li.zhong@hlrs.de, huan.zhou@hlrs.de, marcin.pospieszny@man.poznan.pl

**Abstract**—Containerisation demonstrates its efficiency in application deployment in cloud computing. Containers can encapsulate complex programs with their dependencies in isolated environments, hence are being adopted in HPC clusters. HPC workload managers lack micro-services support and deeply integrated container management, as opposed to container orchestrators (*e.g.* Kubernetes). We introduce Torque-Operator (a plugin) which serves as a bridge between HPC workload managers and container Orchestrators.

**Index Terms**—HPC Workload Manager; Orchestration; Containerisation; Torque; Slurm; Kubernetes; Singularity; Cloud Computing

## I. INTRODUCTION

Cloud computing demands high-portability. Containerisation ensures compatibility of applications and their environment by encapsulating applications with their libraries and configuration files [1], thus enables users to move and deploy programs easily among clusters. Containerisation is a virtualisation technology [2]. Rather than starting a holistically simulated OS on top of the host kernel as in a Virtual Machine (VM), a container only shares the host kernel. This feature makes containers more lightweight than VM. Containers are dedicated to run micro-services and one container mostly hosts one application. Nevertheless, containerised applications can become complex, *e.g.* thousands of separate containers may be required in production. Production can benefit from container orchestrators that can provide efficient environment provisioning and auto-scaling.

High Performance Computing (HPC) systems are traditionally applied to perform large-scale financial and engineering simulation, which demands low-latency and high-throughput. The typical HPC jobs are large workloads that are often host-specific and hardware-specific. HPC systems are typically equipped with workload managers. A *workload manager* is composed of a *resource manager* and a *job scheduler*. A resource manager [3] allocates resources (*e.g.* CPU, memory), schedules jobs and guarantees no interference from other user processes. A job scheduler determines the job priorities, enforces resource limits and dispatch jobs to available nodes [4]. Two main-stream workload managers are TORQUE [5] and Slurm [6]. Slurm includes both resource managers and job schedulers, while originally Torque only incorporates resource managers and later extends with job schedulers. Overall, HPC

workload managers lack micro-service supports and deeply-integrated container management capabilities in which container orchestrators manifest their efficiency.

We herein describe a plugin named *Torque-Operator*. It serves as a bridge between the HPC workload manager *Torque* and the container orchestrator *Kubernetes* [7]. Kubernetes has been widely adopted, as it has a rapidly growing community and ecosystem with plenty of platforms being developed upon it. Furthermore, we propose a testbed architecture composed of an HPC cluster and a big data cluster where *Torque-Operator* enables scheduling container jobs from the big data cluster to the HPC cluster. The rest of the paper is organised as follows. Firstly, Section II briefly views the related work. Next, we describe the proposed architecture of our testbed and *Torque-Operator* in Section III. Followed, some preliminary results are given in Section IV. Lastly, Section V concludes this paper and proposes future work.

## II. RELATED WORK

*Torque-Operator* extends WLM-Operator [8] with *Torque* support. Both operators share similar mechanisms, *i.e.* schedule container jobs from cloud clusters to HPC clusters, nevertheless, their implementation varies significantly as *Torque* and *Slurm* have different structures and parameters.

*WLM-Operator* only allows submission of *Slurm* batch jobs wrapped in a *Kubernetes yaml* file from a cluster managed by *Kubernetes*. It invokes *Slurm* binaries *i.e.* *sbatch*, *scancel*, *sacct* and *scontrol* to transfer and manage *Slurm* jobs to a *Slurm* cluster. The operator creates *virtual nodes* which correspond to each *Slurm* partition, *e.g.* one *virtual node* corresponds to one *Slurm* partition and contains the information of its corresponding partition. *Virtual node* is a concept in *Kubernetes*. It is not a real worker node, however, it enables users to connect *Kubernetes* to other APIs and allows developers to deploy *pods* (a *Kubernetes* term) and containers with their own APIs. Jobs on the *virtual node* can be scheduled to the worker nodes. *WLM-Operator* creates a *dummy pod* on the *virtual node* in order to transfer the *Slurm* batch job to a specific *Slurm* partition. When the batch job completes, another *dummy pod* is generated to transfer the results to the directory specified in the submitted *yaml* file.

In *Kubernetes* terminology, *WLM-Operator* creates a new *object kind* *i.e.* *Slurmjob*. The operator includes a service

program *red-box* that builds a gRPC proxy between Slurm and Kubernetes. gRPC proxy defines a service and implements a server and clients. The service defines the methods and their message types of responds and requests in a *.proto* format file. The server implements: 1) the interfaces 2) and runs a gRPC server which listens to the requests from clients and dispatches them to the right services. The client defines the identical methods as the server.

### III. TORQUE-OPERATOR AND PLATFORM DESCRIPTION

We firstly illustrate the design of our platform architecture, then describe the structure of Torque-Operator. Torque-Operator is written in Golang programming language. *Singularity* [9] is the runtime container of our choice. Singularity is starting to be applied in many HPC centres [10], as it provides a secure means to capture and distribute software and computer environment. For example, execution of a Singularity container only demands a user privilege, while a Docker container [11], which is a container runtime widely adopted in cloud systems, requires root permission. Kubernetes supports Docker by default, though it can be adjusted to perform services for Singularity by adding Singularity-CRI [12]. Table I manifests the list of core applications that construct the testbed.

Orchestrator	Kubernetes, Torque
Container runtime & its support	Singularity, Singularity-CRI
Operator	Torque-Operator
Compiler	Golang compiler

TABLE I  
THE LIST OF CORE APPLICATIONS FOR THE TESTBED.

#### A. Platform Architecture

The architecture of our platform is designed to serve as the testbeds for the EU research project CYBELE<sup>1</sup>. The platform is composed of an HPC cluster with Torque as its workload manager and a big data cluster with Kubernetes as its orchestrator. Its architecture is illustrated in Fig. 1. Noting that Fig. 1 is for illustration purpose, the number of nodes and the queues can vary in the testbeds.

In Torque, nodes are grouped into queues. Each queue is associated with resources limits such as walltime, job size. One node can be included in multiple queues. The HPC cluster is composed of a head node which controls the whole cluster nodes and compute nodes which perform computation. The Torque login node in Fig. 1 also serves as one of the worker nodes in the Kubernetes cluster. The Kubernetes cluster incorporates a master node which schedules the jobs and worker nodes which execute the jobs. A virtual node indicated in Fig. 2 transfers Torque jobs to the Torque cluster. The Torque job submitted from the Kubernetes login node is scheduled by Kubernetes master node to the virtual node. The virtual

<sup>1</sup>CYBELE: Fostering Precision Agriculture and Livestock Farming through Secure Access to Large-Scale HPC-Enabled Virtual Industrial Experimentation Environment Empowering Scalable Big Data Analytics <https://www.cybele-project.eu/>

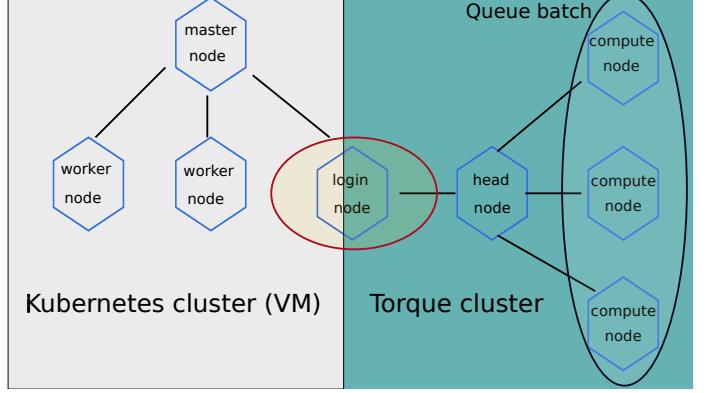


Fig. 1. Architecture of the testbed. The login node belongs to both Kubernetes and Torque clusters. One queue (named batch) is shown in the Torque cluster.

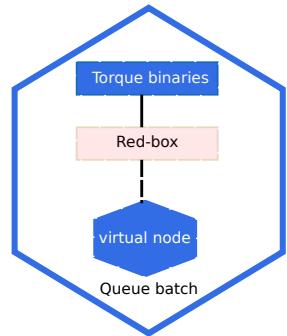


Fig. 2. Architecture of Torque-Operator. This is the internal architecture of the login node as in Fig. 1. The virtual node corresponds to the Torque queue (named batch) in Fig. 1

node transfers the abstracted Torque jobs to the Torque queue through the Torque login node. The merits of this architecture are: 1) it provides users with flexibility to run containerised and non-containerised jobs, 2) the containerised applications can be better scheduled to Torque cluster by taking advantage of the scheduling policies of Kubernetes.

#### B. Structure of Torque-Operator

The Torque job script is encapsulated into a Kubernetes yaml job script. The yaml script is submitted from a Kubernetes login node (in our case, the login node is also the master node). The PBS script part is processed by Torque-Operator. A dummy pod is generated to transfer the Torque job specification to a scheduling queue (*e.g.* waiting queue, test queue, which is a concept in the job scheduler). Torque-Operator invokes the Torque binary *qsub* which submits PBS job to the Torque cluster. When the Torque job completes, Torque-operator creates a Kubernetes pod which redirects the results to the directory that the user specifies in the yaml file.

As in WLM-Operator (Section II), Torque-Operator includes a service program *red-box*. *Red-box* generates a Unix socket which allows data exchange among the Kubernetes and Torque processes. Torque-Operator introduces a new *object kind* *i.e.* *Torquejob* (*Slurmjob* in WLM-Operator) and sets

```

1  apiVersion: wlm.sylabs.io/v1alpha1
2  kind: TorqueJob
3  metadata
4    name: cow
5  spec:
6    batch: |
7      #!/bin/sh
8      #PBS -l walltime=00:30:00
9      #PBS -l nodes=1
10     #PBS -e $HOME/low.err
11     #PBS -o $HOME/low.out
12     export PATH=$PATH:/usr/local/bin
13     singularity run lolcow_latest.sif
14   results:
15     from: $HOME/low.out
16   mount:
17     name: data
18     hostPath:
19       path: $HOME/
20       type: DirectoryOrCreate

```

---

```
$kubectl apply -f $HOME/cow_job.yaml
```

Fig. 3. An example of the yaml script and its submission command. The script encloses a PBS script.

```
$kubectl get torquejob
NAME    AGE    STATUS
cow     2s     running
```

Fig. 4. The command to view the status of the yaml job

it as *Kubernetes deployment*. Torque-Operator builds four Singularity containers which are deployed by Kubernetes on its worker nodes to perform the corresponding services, *e.g.* create dummy pod to transfer the results from Torque to Kubernetes.

#### IV. TEST CASE

Simple experiments have been conducted to validate Torque-Operator. Fig. 3 presents a Kubernetes yaml job script (*cow\_job.yaml*). More specifically, inside the yaml script, the Torque script requests 30-minute walltime and one compute node. The error file and output file are stored in *low.err* and *low.out* which locate in *\$HOME/* directory. The script appends the path */usr/local/bin* where Singularity binary resides. The Singularity container image *lolcow\_latest.sif* is executed. The results are given in Fig. 5. The user can view the status of the job easily from Kubernetes login node as shown in Fig. 4. Additionally, the status of the PBS job can be output using the Torque commands on the Torque login node.

#### V. CONCLUSION AND FUTURE WORK

We described the testbed architecture for the EU research project CYBELE and introduced the structure of Torque-Operator that extends WLM-Operator with Torque support. This testbed architecture creates a connection between HPC and cloud clusters. Moreover, it provides users with flexibility to run containerised and non-containerised jobs and may enhance the capability of container scheduling on HPC.

```
If one cannot enjoy reading a book over \
| and over again, there is no use in
| reading it at all.
|
\ -- Oscar Wilde
```

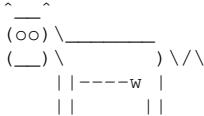


Fig. 5. A result of the Singularity job

The future work will focus on optimization of Torque-Operator that can offer more stable deployments. Performance evaluation will be carried out to compare efficiency of scheduling the container jobs by Kubernetes and Torque. The pilots of CYBELE project will be adopted as the benchmarks.

#### ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement NO.825355.

#### REFERENCES

- [1] M. Khan, T. Becker, P. Kuppuudaiyar, and A. C. Elster, "Container-Based Virtualization for Heterogeneous HPC Clouds: Insights from the EU H2020 CloudLightning Project," in *2018 IEEE International Conference on Cloud Engineering (IC2E)*, pp. 392–397, April 2018.
- [2] M. A. Rodriguez and R. Buyya, "Container-based cluster orchestration systems: A taxonomy and future directions," *Softw., Pract. Exper.*, vol. 49, no. 5, pp. 698–719, 2019.
- [3] M. Hovestadt, O. Kao, A. Keller, and A. Streit, "Scheduling in hpc resource management systems: Queuing vs. planning," in *Job Scheduling Strategies for Parallel Processing* (D. Feitelson, L. Rudolph, and U. Schwiegelohhohn, eds.), (Berlin, Heidelberg), pp. 1–20, Springer Berlin Heidelberg, 2003.
- [4] D. Klusáček, V. Chlumský, and H. Rudová, "Planning and optimization in torque resource manager," in *Proceedings of the 24th International Symposium on High-Performance Parallel and Distributed Computing*, (New York, NY, USA), Association for Computing Machinery, 2015.
- [5] G. Staples, "Torque resource manager," in *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, (New York, NY, USA), p. 8, Association for Computing Machinery, 2006.
- [6] M. A. Jette, A. B. Yoo, and M. Grondona, "SLURM: Simple Linux Utility for Resource Management," in *Lecture Notes in Computer Science: Proceedings of Job Scheduling Strategies for Parallel Processing (JSSPP)* 2003, pp. 44–60, Springer-Verlag, 2002.
- [7] L. Abdollahi Vayghan, M. A. Saeid, M. Toeroe, and F. Khendek, "Deploying Microservice Based Applications with Kubernetes: Experiments and Lessons Learned," in *2018 IEEE 11th International Conference on Cloud Computing (CLOUD)*, pp. 970–973, July 2018.
- [8] V. Pisaruk and S. Yakovtseva, "WLM-operator." Gitlab: <https://github.com/sylabs/wlm-operator> (accessed on 13/02/2020).
- [9] G. M. Kurtzer, V. V. Sochat, and M. Bauer, "Singularity: Scientific containers for mobility of compute," in *PloS one*, 2017.
- [10] G. Hu, Y. Zhang, and W. Chen, "Exploring the Performance of Singularity for High Performance Computing Scenarios," in *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPC-C/SmartCity/DSS)*, pp. 2587–2593, Aug 2019.
- [11] C. Negus, *Docker Containers (Includes Content Update Program): Build and Deploy with Kubernetes, Flannel, Cockpit, and Atomic*. USA: Prentice Hall Press, 1st ed., 2015.
- [12] Sylabs, "Singularity-CRI." <https://sylabs.io/guides/cni/1.0/user-guide/k8s.html> (Accessed on 03/03/2020).