Concept of a Preservation System for Scientific Experiments in HPC

Kyryll Udod
Ulm University
Germany
kyryll.udod@uni-ulm.de
https://orcid.org/0000-0002-3506-7980

Volodymyr Kushnarenko
Ulm University
Germany
volodymyr.kushnarenko@uni-ulm.de
https://orcid.org/0000-0001-7427-2410

Stefan Wesner
Ulm University
Germany
stefan.wesner@uni-ulm.de
https://orcid.org/0000-0002-7270-7959

Abstract – this poster presents a concept of a preservation system for computations on High Performance Computing (HPC) resources. It covers some important challenges and possible solutions related to the preservation of scientific experiments on HPC systems for their further reproduction. Storage of the experiment as only a code with some related data is not completely enough for its future reproduction, especially in the long term. Preservation of the whole experiment's environment (operating system, used libraries, environment variables, input data, etc.) using containerization technology (e.g. Docker, Singularity) is proposed as a suitable solution for that. This approach allows to preserve an entire environment, but leaves a problem, how to deal with the commercial software that was used within the experiment. As a solution authors propose to replace during the preservation procedure all commercial software with their open source analogues, what should allow future reproduction of the experiment without any legal issues. The prototype of such a system was developed, the poster provides a scheme of the system and the first experimental results.

Keywords – HPC, reproducible research, containerization, research experiments preservation

Conference Topics – What is emerging practice in software preservation and in emulation/ virtualization?

I. Introduction

High Performance Computing plays an important role in almost every research area providing to the users always suitable hardware and software resources to solve complex scientific problems. Because of the continuously growing community of HPC users and amount of research experiments, a question of research data management on HPC starts to play a significant role [1] including reproducibility of the research results as a major aspect for the scientists [2]. Containerization technology (e.g. Docker, Singularity) could be used to preserve a complete environment of the scientific experiment. This approach works well, especially when within the experiment only open source software is used. But in the case of commercial libraries some legal issues can come by the future reproduction procedure. To make a preservation of the commercial software easy and free of any legal problems, some special solution is needed.

This poster begins with generally available preservation options and related to them problems, potential solutions are discussed. Authors touch a question, what type and scope of the preserved information is needed to allow further reproduction of the experiment and how this information could be taken. In the final part the poster presents a scheme of the system prototype that was developed to solve the mentioned above problems related to the preservation and further reproduction of scientific experiments, where also commercial software was involved. The first experimental results and further investigation steps are discussed as well.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

II. Mechanisms for Research Preservation on HPC Systems

To make a scientific experiment reproducible, also software and hardware information should be preserved [3] - information about the operating system and all used within the experiment software libraries, as well as information about the hardware components and configuration aspects of the current HPC system.

Related to the experiment information usually can be extended from the job-script, what can include the used in the experiment software libraries with specific versions, environment variables, etc. Reproduction of the experiment in this case could be possible, but only on the same machine with the same (not changed) configuration of the system, what is practically not possible for the long term because of continuous system updates.

That's why for the long term preservation not only information about the experiment, but also about the whole system components (in a specific for the experiment state) should be stored. In this case containerization technology (e.g. Docker, Singularity) could be used. It allows to create a full copy of the system with all related and used within the experiment software components

III. Problems with the Containerization Approach

With the containerization approach often unlimited access to the preserved components is needed, what can be difficult in the case of HPC systems, where not all software components can be copied or even accessed by the user.

Even when software components are accessible, they could be not open source and some license could be needed for their further reuse. These two issues stay as a central point of the poster's topic

IV. Proposed Approach for Preservation of Software Components with Limited Access

To preserve not fully accessible for the user (because of the account rights) or some commercial software components, authors propose an approach, that all such components should be replaced with their open source alternatives, which are always accessible, free to use and can be stored and reused later without any legal issues.

To collect all needed for the preservation information about the experiment, some available for the user mechanisms should be used, e.g. status request for the current computational job, which represents the experiment (e.g. via "checkjob" or similar command that is traditionally available on HPC systems as a part of the job scheduler).

A proposed preservation system consists of two components. One component is a special script that requests the job status and collects all needed information about the experiment locally on the computing machine (HPC cluster). The second component represents an external server. The server provides a REST-API that can be used for the communication with the client part - the first component on the side of the HPC cluster. Information from the cluster is sent via the POST request. The server is responsible for the containerization procedure and replacement of the commercial libraries with their open source alternatives. The process works automatically, but the user can also steer it via the web-interface (e.g. to replace some software or choose the most suitable version of it). Further running procedure for the created containers can be performed via the EaaS system (Emulation as a Service) [4]. Publication and referencing of the preserved containers are foreseen.

For the current moment the proposed preservation system is in an early prototypical phase, as test experiments some molecular dynamic simulations with "SIESTA" [5] are used.

V. Acknowledgements

The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK).

References

[1] B. Schembera and T. Bönisch, “Challenges of Research Data Management for High Performance Computing,” in Research and Advanced Technology for Digital Libraries, 2017, pp. 140–151.

[2] S. Hunold, “A Survey on Reproducibility in Parallel Computing,” arXiv:1511.04217 [cs], Nov. 2015.

[3] P. F. Klaus Rechert and Tom Ensom, “Towards a Risk Model for Emulation-based Preservation Strategies: A Case Study from the Software-based Art Domain”, 13th International Conference on Digital Preservation (iPRES2016), Bern, Switzerland, 3-6 October 2016.

[4] http://citar.eaas.uni-freiburg.de/

[5] https://departments.icmab.es/leem/siesta/

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.