CONCEPT OF 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 – the poster represents a concept of the preservation system for computations on High Performance Computing (HPC) resources. It covers some important challenges related to the preservation of scientific experiments on HPC systems for their further reproduction and provides possible solutions for that. Storage of the experiment as a code with some related data is not only enough for its future reproduction, especially in long term. Preservation of the whole experiment's environment (operation system, used libraries, environment variables, input data, etc.) using the containerization technology (Docker, Singularity) is proposed as a possible solution for that. This approach allows to preserve the entire environment, but leaves a problem, how to deal with the commercial software that was used within the experiment. As a possible solution authors propose to replace all commercial software with the open source analogues, what should allow future reproduction of the experiment without any legal issues. The prototype of such system was developed, the poster provides the 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 nowadays providing to the users always suitable hardware and software resources to solve complex scientific problems. Because of continuously growing community of HPC users and amount of research experiments, the question of research data management on HPC starts to play one of the most important roles [1] including reproducibility of research results as a major aspect for the scientists [2]. A containerization technology (e.g. Docker, Singularity) could be used for preservation of the complete environment of the scientific experiment. This approach works well, especially when within the experiment only open source software is used. In case of commercial libraries some legal issues can come by future reproduction procedure. To make a preservation of the commercial software easy and without legal problems, some special solution is needed.

The poster begins with the question of generally available preservation options and related to them problems. Then possible solutions are discussed. Authors touch the 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 represents the poster a scheme of the system prototype, that was developed to solve the mentioned above problems related to the preservation and further reproduction of the scientific experiments, where also commercial

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

DOI: 10.1145/nnnnnnn.nnnnnnn

iPRES 2019

software components were involved. The first results and further investigation steps are discussed.

II. Mechanisms for Research Preservation on HPC Systems

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

Usually the related to the experiment information can be extended from the job-script - e.g. used in the experiment software libraries with specific versions, environment variables, etc. The 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 long term because of continuous system updates.

That's why for the long term preservation not only the information about the system, but also the whole components of the system (in a specific for the experiment state) should be stored. In this case the 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 an 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 the software components are accessible, they can be not open source and some license will be needed for their further reuse.

These two problems represent the main focus of the poster.

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

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

To collect the 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. "checkjob" command which is available on many HPC systems)

The proposed system consists of two components. One component is a special script that requests the job status and collects the information about the experiment. The second component represents an external server. The server provides a REST-API, what can be used for the communication with the client part - the first component on the side of the HPC cluster. The information from the cluster is sent via 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 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 of the containers is possible via EaaS system (Emulation as a Service) [4]. Publication and referencing of the created containers are foreseen.

For the current moment as a use case the molecular dynamic simulations with "SIESTA" [5] are used. The proposed system is currently in early prototypical phase.

V. Acknowledgements

The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

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/