eResearch AUSTRALASIA2019

# Tools for reproducible and portable research workflows at Pawsey Centre

*Marco De La Pierre*[1]

**Brian Skjerven[2], Mark Cheeseman[3], William Davey[4], Mark Gray[5]**

Pawsey Supercomputing Centre, 6151 Kensington WA, Australia

[1]marco.delapierre@pawsey.org.au
[2]brian.skjerven@pawsey.org.au
[3]mark.cheeseman@pawsey.org.au
[4]william.davey@pawsey.org.au
[5]mark.gray@pawsey.org.au

## ABSTRACT

An increasing number of research workflows for HPC and Cloud involve a moderately large number of tasks to be executed over large datasets (hundreds to thousands of samples) and each requiring distinct application packages or modules. These requirements pose a number of challenges, including installation and maintenance of large software stacks, enforcement of reproducibility and portability of the analyses, workflow throughput and scalability.
We will present an overview of ongoing efforts in investigating and deploying software technologies that allow researchers to tackle these issues. In particular, we will discuss container technology and workflow management tools, both for HPC and Cloud environments. Practical examples from user cases will be provided to document advantages and limitations of these technologies.

## CONTAINER TECHNOLOGIES

Following an investigation over available technologies, container engines have been deployed at Pawsey Centre for the past year. Multiple engines have been adopted to allow users to run containers both on HPC and Cloud. Interfaces are in place to leverage the different types of available hardware resources. Our user services provide support to build and run containers, in addition to developing a dedicated body of documentation and tutorials.

## WORKFLOW MANAGEMENT TOOLS

Several workflow management systems have been assessed over the past few months at Pawsey Centre, aiming to identify a subset for which to provide in-depth user support. A number of aspects are to be taken into account in this evaluation, including portability, ease of use, documentation. Researcher pipelines have been implemented to probe real use cases, too. Taken all together, these factors have permitted to identify some promising candidate technologies.

## EXAMPLES

In the past year Pawsey staff has been engaging with researchers to assist them in containerizing their workflows, and then automate them through workflow tools. At the moment, examples come largely from bioinformatics, although there is a perspective to expand these tools to other domains, for instance radio-astronomy.