EMS Annual Meeting Abstracts
Vol. 14, EMS2017-332, 2017

JASMIN and the role of Cloud Computing in realising a Big Data facility for the Environmental Sciences

Philip Kershaw (1), Jonathan Churchill (2), Victoria Bennett (1), Matt Pritchard (3), Matt Pryor (3), and Bryan Lawrence (4)

(1) STFC Rutherford Appleton Laboratory, NCEO/Centre for Environmental Data Analysis, Didcot, United Kingdom (philip.kershaw@stfc.ac.uk), (2) STFC Rutherford Appleton Laboratory, Scientific Computing Department, Didcot, United Kingdom, (3) STFC Rutherford Appleton Laboratory, NCAS/Centre for Environmental Data Analysis, Didcot, United Kingdom (philip.kershaw@stfc.ac.uk), (4) NCAS / University of Reading, Reading, United Kingdom

JASMIN is a NERC facility providing a large centralised computing resource to serve the Big Data needs of the UK environmental sciences community and its work with international collaborators. We describe the role of cloud computing in the context of the overall development and evolution of JASMIN from the beginning of operations in 2012 to the present.

JASMIN is architected around the provision of a large volumes of storage (currently approximately 16PB disk) co-located with services to access, analyse and process the data hosted. The latter is split between a curated archive hosting NERC virtual data centres covering atmospheric science and Earth observation domains and user-managed group workspaces. Services include a community cloud and Lotus, a bare-metal batch compute system. These services and the storage are underpinned by an infrastructure architected around a high-performance non-blocking network and parallel file system. The network architecture has been fundamental not only for ensuring good i/o performance internally for data processing but also for ingress and egress of data between the JASMIN environment and external sources, facilitated via a dedicated Data Transfer Zone (DTZ).

Lotus, the batch compute environment has had a high-level of utilisation and has been successively expanded over the course of JASMIN's existence. The cloud service has developed over a number of evolutionary steps and today hosts around twenty separate tenancies for a range of groups and collaborative projects. A number of usage patterns have emerged, for example, the hosting of portals and web services to front data from group workspaces or the archive to the provision of full virtual research environments such as ESA Thematic Exploitation Platforms or the NERC DataLab currently under development. It has also been employed as a means to deliver teaching courses via Desktop-as-a-Service and Jupyter Notebook systems.

Looking to the future, there are challenges to be addressed around scaling: in terms of the volumes of data, the size of the infrastructure, the numbers of users and the impact of all of these factors on the overall management and operation of the system. Cloud technologies have a key role in addressing these. Multi-tenancy and IaaS (Infrastructure-as-a-Service) have already demonstrated how a cloud model can benefit, delegating management of infrastructure to users and segregating the access and allocation of resources. Equally however, the POSIX interface and single global user id space associated with the current parallel file system is fundamentally incompatible with this model and have provided a barrier to realising the full potential of the cloud. Consequently work is underway to evaluate object stores as an alternative solution with a view to full migration. A staged approach will be taken, considering i/o performance requirements in the choice of technology adopted and for its rollout, the necessary paradigm shift for the user community to migrate applications from POSIX to RESTful APIs associated with object storage.