EO SCIENCE FROM BIG EO DATA ON THE JASMIN-CEMS INFRASTRUCTURE

Victoria Bennett[1,3], Philip Kershaw[1,3],Matt Pritchard[1], Jonathan Churchill[2], Cristina Del Cano Novales[2], Martin Juckes[1,4], Stephen Pascoe[1,4], Sam Pepler[1,4], Ag Stephens[1,4], Bryan Lawrence[1,4,6], Jan-Peter Muller[3,7], Said Kharbouche[3,7], Barry Latter[3,5], Jon Styles[8]

1. Centre for Environmental Data Archival, RAL Space, STFC Rutherford Appleton Laboratory, UK; 2. Scientific Computing Department, STFC Rutherford Appleton Laboratory, UK; 3. National Centre for Earth Observation, UK; 4. National Centre for Atmospheric Science, UK; 5. Remote Sensing Group, RAL Space, STFC Rutherford Appleton Laboratory, UK; 6. University of Reading, UK; 7. Mullard Space Science Laboratory, University College London, UK; 8. Assimila Ltd, Reading, UK

ABSTRACT

We report on a major expansion to JASMIN, a big data infrastructure upon which the UK Centre for Environmental Data Archival (CEDA) operates data centres and a major scientific analysis environment. The academic component of the facility for Environmental Monitoring from Space (CEMS) is hosted on JASMIN and continues to grow significantly, both in capability and its usage by the Earth Observation science community.

We describe recent infrastructure upgrades to storage, compute and networking, and the deployment of a full private cloud. Dedicated data transfer nodes allow high-performance data flows to meet the needs of the climate and earth observation science communities.

Over 20 EO science projects are currently active on JASMIN-CEMS, examples include near-real time atmospheric composition processing, global land surface products, and a collaborative research environment for land data assimilation (Optirad).

Index Terms— e-Infrastructure, data compute, Earth Observation data, data processing, data archive

1. INTRODUCTION

The JASMIN facility (jasmin.ac.uk) at Harwell, UK, is a "super-data-cluster" which delivers infrastructure for data analysis. It is half super-computer and half data-centre and as such provides a globally unique computational environment [1]. JASMIN Phase 1 (with funding from the UK Natural Environment Research Council, NERC, and the UK Space Agency, UKSA) was delivered in early 2012. Two further phases of NERC funded expansion are underway: Phase 2 in early 2014 and Phase 3 in late 2014.

JASMIN provides four basic services to the community: storage (including disk and tape), batch computing (on "Lotus"), hosted computing, and cloud computing.

A range of NERC science community services are run in the JASMIN infrastructure, one of which is the academic component of the facility for Climate, Environment and Monitoring from Space (CEMS), hosting data and services specifically for the Earth Observation science community.

2. RECENT INFRASTRUCTURE UPGRADES

JASMIN Phase 2 and 3 upgrades are now underway, with many components already in operation.

2.1. Storage and Compute

The Phase 2 and 3 upgrades include major expansion to over 3,500 compute cores with 13 PB of fast parallel disk storage, and equivalent capacity in near-line tape (Figure 1).

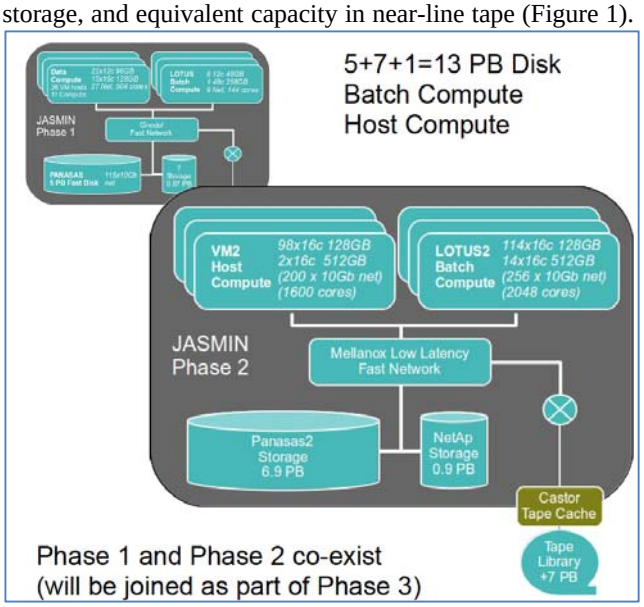


Figure 1: JASMIN Infrastructure upgrades

2.2. Network

Expansion and re-implementation of the internal network has massively increased capacity for large-scale parallel data processing. This gives file system access to archive and shared workspace environments enabling users to process high volume EO and climate datasets at much higher speeds than previously possible.

2.3. Private cloud

A further development this year is the provision of a full private cloud. This is divided up into Managed and Unmanaged zones. The Managed provides a PaaS (Platform as a Service) offering in which users can deploy a virtual machine from a restricted catalogue of templates. VMs have direct access to the archive and high performance processing and storage. The Unmanaged service provides full IaaS (Infrastructure as a Service), enabling users to host their own virtual infrastructures of applications and services.

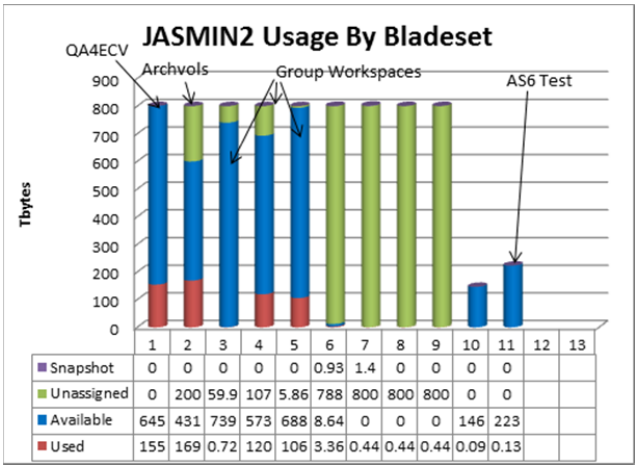
2.4. Data transfer

A dedicated "Science DMZ" has also been deployed with high-performance data transfer nodes and monitoring tools for a variety of large bidirectional science data flows.
Collaboration with other sites as part of the International Climate Network Working Group has set ambitious throughput targets representative of climate community needs, but these are likely to be at least matched, if not exceeded, by the data transfer requirements of the Earth Observation community in the Sentinel era.

3. JASMIN-CEMS OPERATIONS

The JASMIN-CEMS infrastructure now supports nearly 500 users (409 JASMIN-login users, 88 CEMS-login users) and more than 80 different projects. Previously, these users mostly purchased their own, or used hardware in their own institutions, e.g. university departments, and transferred large volumes of data across the network for local processing and analysis. The JASMIN-CEMS community shared infrastructure allows users to share access to the same copy of datasets, to process and analyse results *in situ*, and efficiently make their results available to other researchers.
To date (October 2014) 4.9 PB has been allocated to science projects as Group Workspace data storage (allocations of storage managed by users/project teams) and 2.8 PB is in use for the Data Centre archives (datasets curated by CEDA, the Centre for Enviromental Data Archival) . In the first two years of operation, over 1.3 million processing jobs were executed.

Figure 2 shows the current usage of the JASMIN2 storage, which was commissioned in April this year. JASMIN1 storage is already fully allocated. JASMIN3 storage purchase this year will likely add about 4-5 more bladesets.



	1	2	3	4	5	6	7	8	9	10	11	12	13
Snapshot	0	0	0	0	0	0.93	1.4	0	0	0	0		
Unassigned	0	200	59.9	107	5.86	788	800	800	800	0	0		
Available	645	431	739	573	688	8.64	0	0	0	146	223		
Used	155	169	0.72	120	106	3.36	0.44	0.44	0.44	0.09	0.13		

Figure 2: JASMIN2 usage October 2014. Blue is allocated but not yet used. Red is used. Green is unallocated.

The largest datasets hosted on the infrastructure at present are climate model simulations (CMIP5, Coupled Model Intercomparison Project, Phase 5) data, 1.2 PB total volume and still growing. However, with the advent of data from the Sentinel satellites, and CMIP6 (PB's/year) in the coming years we plan to expand both the online storage capacity, as well as making increased use of nearline tape storage at much lower cost.

4. EO SCIENCE APPLICATIONS

More than 20 science projects are currently actively using JASMIN-CEMS for EO data processing and analysis. We focus on three different usage examples in this paper: NRT-Ozone, QA4ECV and Optirad. The first two of these projects use a combination of hosted compute and batch compute, whereas the third, Optirad, uses the new cloud compute environment.

4.1. NRT-Ozone

In this project, the Remote Sensing Group in RAL Space are generating profiles of trace gases in near real time (NRT) in a pilot study for support to operational atmosphere services. The RAL Remote Sensing Group's algorithm [2,3] has been developed to produce tropospheric ozone data from MetOp's uv-nadir spectrometer GOME-2 which substantially exceeds the quality provided of the standard operational products. Due to the large data volumes and CPU requirements, combined with the time-critical

acquisition of auxiliary data, near real time processing of the MetOp data with the RAL scheme has not previously been viable. With the dedicated allocation of resource within the JASMIN-CEMS infrastructure, the team are now able to deliver this improved data product within the required timeframe to feed into the forecasting of air quality and numerical weather prediction. Initial calculations indicate that the NRT "Gome-2 only" O3 scheme requires ~80 cores 24/7, 360 more will be required for IASI CH4 & Gome-IASI O3, to be implemented soon. Systems have been put in place to collect NRT data from EUMETSAT and ECMWF into a dedicated workspace for rapid processing on the JASMIN-CEMS infrastructure and performance testing is underway. This project is funded by the UK National Centre for Earth Observation and the NERC Big Data programme.

4.2. QA4ECV

As part of the EC FP7-funded QA4ECV [4] project, new essential climate variables are being generated by MSSL/UCL on the JASMIN-CEMS system. This activity involves processing of hundreds of TB to generate 35 years of quality assured global land surface products from European and US missions. A dedicated project workspace (> 700 TB) on JASMIN-CEMS has been funded through the UK NERC Big Data programme. Early work has demonstrated that the reprocessing of MODIS priors using all spectral, and 3 broadbands was completed on CEMS in 3 days, 81 times faster than on an in-house 8-core blade. Other processing tasks show similar benefits, e.g. reprojecting BRDF files from SIN coordinates to lat/lon requires a huge number of polygons to be spatially indexed and processed. This process requires massive RAM and usually takes a very long time, however the JASMIN-CEMS system is ~100 times faster than the previously used 224-core in-house linux cluster [5]. The project has also benefited from a dedicated high-performance transfer server in the "science DMZ", which has allowed input datasets (MODIS, MISR) to be transferred from US data providers onto CEMS very efficiently, at rates up to 28 TB/day.

4.3. Optirad

OPTIRAD (OPTImisation environment for joint retrieval of multi-sensor RADiances) is developing a collaborative research environment for land data assimilation using IPython Notebook on the CEMS Unmanaged cloud. The work is funded by ESA and the project team at UCL aim to provide the scientific community with a dedicated software environment to generate products from raw EO data. Compared to traditional retrieval techniques, the data assimilation algorithms used for OPTIRAD [6] are compute intensive, with high memory demands to store the state. A significant effort in the project will devoted to improving the efficiency of the current implementation by both mathematical and algorithmic improvements. Nevertheless a heavy computing load is still expected. Benchmarking the current implementation shows that a validation experiment over an area the size of a Sentinel-2 scene, assimilating 500 million observations (the equivalent of about 20 Sentinel-2 scenes through the course of a growing season) would cost approximately 15 core-months of CPU time. Use of the JASMIN-CEMS compute resources will significantly speed up the test and validation of the algorithms, as well as making the resulting system more feasible for scientific experiments.

5. CONCLUSION AND FORWARD LOOK

JASMIN (and JASMIN-CEMS) Phase 2 and 3 go beyond the storage and batch compute service offered in Phase 1. The storage and batch compute model produced many excellent results for first users, but there is a "long tail" of the user community who are less expert users of e.g. the Linux command line and high performance computing. The new cloud services, and projects like Optirad, present the opportunity to support this much wider community.

6. REFERENCES

[1] Lawrence, B. N., Bennett, V. L., Churchill, J., Juckes, M., Kershaw, P., Pascoe, S., Pepler, S., Pritchard, M. and Stephens, A. (2013) Storing and manipulating environmental big data with JASMIN. Proceedings of IEEE Big Data 2013, p68-75 doi:10.1109/BigData.2013.6691556.

[2] Miles, G. M., Siddans, R., Kerridge, B. J., Latter, B. G., and Richards, N. A. D.: Tropospheric ozone and ozone profiles retrieved from GOME-2 and their validation, Atmos. Meas. Tech. Discuss., 7, 7923-7962, doi:10.5194/amtd-7-7923-2014, 2014

[4] Munro R; Siddans R; Reburn WJ; Kerridge BJK, 1998, Direct measurement of tropospheric ozone distributions from space, NATURE 392 (6672): 168-171

[4] http://www.qa4ecv.eu/

[5] Boersma, F and J-P Muller, EU-FP7 QA4ECV: a 35 year ECV of albedos, fapar & LAI and their uncertainties, presentation at GlobAlbedo Second User Consultation Meeting (UCM-2), September 2014, ECMWF, Reading, UK, http://www.ecmwf.int/sites/default/files/Muller-QA4ECV-GA-UCM2-140917.pdf

[5] Lewis, P. & et al., 2012. An Earth Observation Land Data Assimilation System (EO-LDAS). Remote Sensing of Environment, 120, pp. 219-235.