Geophysical Research Abstracts
Vol. 16, EGU2014-10630, 2014
EGU General Assembly 2014

The JASMIN Analysis Platform – bridging the gap between traditional climate data practicies and data-centric analysis paradigms

Stephen Pascoe, Alan Iwi, philip kershaw, Ag Stephens, and Bryan Lawrence
Science and Technology Facilities Council, RALSpace, Didcot, United Kingdom (Stephen.Pascoe@stfc.ac.uk)

The advent of large-scale data and the consequential analysis problems have led to two new challenges for the research community: how to share such data to get the maximum value and how to carry out efficient analysis. Solving both challenges require a form of parallelisation: the first is social parallelisation (involving trust and information sharing), the second data parallelisation (involving new algorithms and tools). The JASMIN infrastructure supports both kinds of parallelism by providing a multi-tennent environment with petabyte-scale storage, VM provisioning and batch cluster facilities.

The JASMIN Analysis Platform (JAP) is an analysis software layer for JASMIN which emphasises ease of transition from a researcher's local environment to JASMIN. JAP brings together tools traditionally used by multiple communities and configures them to work together, enabling users to move analysis from their local environment to JASMIN without rewriting code. JAP also provides facilities to exploit JASMIN's parallel capabilities whilst maintaining their familiar analysis environment where ever possible.

Modern opensource analysis tools typically have multiple dependent packages, increasing the installation burden on system administrators. When you consider a suite of tools, often with both common and conflicting dependencies, analysis pipelines can become locked to a particular installation simply because of the effort required to reconstruct the dependency tree. JAP addresses this problem by providing a consistent suite of RPMs compatible with RedHat Enterprise Linux and CentOS 6.4. Researchers can install JAP locally, either as RPMs or through a pre-built VM image, giving them the confidence to know moving analysis to JASMIN will not disrupt their environment.

Analysis parallelisation is in it's infancy in climate sciences, with few tools capable of exploiting any parallel environment beyond manual scripting of the use of multiple processors. JAP begins to bridge this gap through a veriety of higher-level tools for parallelisation and job scheduling such as IPython-parallel and MPI support for interactive analysis languages. We find that enabling even simple parallelisation of workflows, together with the state of the art I/O performance of JASMIN storage, provides many users with the large increases in efficiency they need to scale their analyses to conteporary data volumes and tackly new, previously inaccessible, problems.