Jade: using on-demand cloud analysis to give scientists back their flow

Niall Robinson and Rachel Prudden

(rachel.prudden@informaticslab.co.uk)

The UK's Met Office generates ∼400 TB weather and climate data every day by running physical models on its Top 20 supercomputer. As data volumes explode, there is a danger that analysis workflows become dominated by watching progress bars, and not thinking about science. We have been researching how we can use distributed computing to allow analysts to process these large volumes of high velocity data in a way that's easy, effective and cheap.

Our prototype analysis stack, Jade, tries to encapsulate this. Functionality includes:

An under-the-hood Dask engine which parallelises and distributes computations, without the need to re-train analysts

Hybrid compute clusters (AWS, Alibaba, and local compute) comprising many thousands of cores

Clusters which autoscale up/down in response to calculation load using Kubernetes, and balances the cluster across providers based on the current price of compute

Lazy data access from cloud storage via containerised OpenDAP

This technology stack allows us to perform calculations many orders of magnitude faster than is possible on local workstations. It is also possible to outperform dedicated local compute clusters, as cloud compute can, in principle, scale to much larger scales. The use of ephemeral compute resources also makes this implementation cost efficient.