University of BRISTOL

ACRC Case Study:
Post-processing a Climate Simulation

Overview

The purpose of this document is to describe the data processing required to analyse the output of a climate simulation run on BlueCrystal using the UK Met. Office Unified Model (UM).

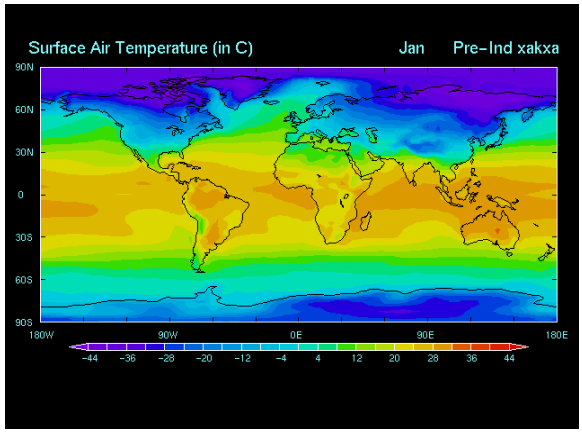

fig1: A plot of surface air temperature created from climate model output data.

Characteristics of the Model

The UM runs on distributed memory machines, such as BlueCrystal using MPI. Lower resolution versions of the model are sensitive to network latency. Data for output is gathered onto the master process and is written to file in 50KB chunks. File writing is sensitive to file-system latency. Using one node of BlueCrystal phase 3 (16 cores), the low-resolution version of the model (called FAMOUS) can simulate 300 years of weather in a (wall clock) day. This corresponds to 600GB of data in files varying between 2 and 35MB in size. A higher-resolution version (HadCM3) will produce data at a slower rate—75GB/day—using the same computational resources. The files in this case vary in size between 3 and 80MB.

The BRIDGE group are one of the largest users of BlueCrystal resources. Prof. Paul Valdes alone has to date used millions of CPU-hours on BlueCrystal phase3.

The sensitivity of the model to file-system latency could possibly be reduced by modifying the code to use asynchronous file writes.

Post-processing the Model Output

A user will typically run several simulations at the same time. We can see that, especially for FAMOUS, a user can quickly exceed their quota (5TB on phase 3, but less than 1TB on phase 2). For this reason, model output must be continually copied off the cluster as the simulations are running. This task is given to a script which is programmed to run periodically (using cron). The script looks at file access times and, if run frequently, can make significant demands on the meta-data component of a file-system. The files are copied off the cluster one-by-one and so the login nodes can also become swamped with SSH connections. *sshfs* has been used to keep the number of SSH sessions down to a minimum. A 1 gigabit ethernet connection can transfer ~30MB/s (~100MB/s if jumbo frames are enabled). Transferring the files one-by-one off the cluster does not typically saturate the network.

Considerable scratch space is required to support the post-processing activities of a large resource group. The BRIDGE group has over 100TB of spinning disk dedicated to the task. The key post-processing task is to calculate climate means. For example, the model may write out files containing data for each month of a simulation. To compute an annual mean, data must be extracted from the monthly files. Files will contain many climate variables and will be revisited when computing the climate means for those variables. These operations and file-system bound and are

again highly sensitive to file access latency. The disks holding the data will be subject to the data processing demands of many users simultaneously.

The sensitivity of the post-processing stage to file-system latency could be reduced by opening each output file only once. However, the computing climate means is fundamentally a file-system bound task, as relatively little computation is involved.

Long-Term Storage

Once the analysis of the module output is complete, a number of key files are archived in case the experiment needs to be re-analysed, repeated or extended. To date this archive extends to well over 100TB and grows by 30-40TB each year. Since the archived is seldom accessed, the data is a good candidate for tape storage.

In addition to storing certain model output files, a searchable web resource is maintained and is populated by plots made from the output of each experiment run. This resource is currently around 10TB in size and will grow by a few terabytes each year. A backup of this resource is highly desirable.