Seamless Access to Multiple Datasets

K. J. Cole, M. A. S. Jones, S. M. Pickles, M. Riding, K. Roy, C. Russell, M.Sensier, W. T. Hewitt,
Manchester Computing, The University of Manchester

The objective of SAMD is to integrate and streamline the access to data repositories and analysis engines in a social science context. We demonstrate this using Grid-based modifications to the current National Statistics (NS) Time Series Data interface operated by MIMAS.

Current access to time series in the National Statistics Time Series Data held by the MIMAS service at the University of Manchester is through a web interface (http://www.mimas.ac.uk/macro_econ/ons/), with username/password authentication. Once authenticated the user may search for and select various data series. The process involves several transfers of any selected datasets back and forward between server and client (the web browser) before analysis can even begin. The user is left with one or more files on her local disk, which is not necessarily the system on which subsequent analysis is to be performed. Although adequate for *ad hoc* enquiries, the present situation is not conducive to systematic cross-analysis of multiple datasets, nor to automation of queries that must be rerun periodically (for example, whenever a particular dataset changes) nor even to retrieval of very large datasets (there are faster protocols than HTTP).

SAMD offers a simple solution. It allows databanks to be searched and resulting series to be assembled into and used where required for files ready for analysis. Data files are transferred directly to the HPC engine of choice using GridFTP. The user's proxy credential is picked up by the application authentication in a single sign-on environment.

The demonstrator application transfers multiple time series from MIMAS (a JISC supported national data centre based at Manchester Computing) to an HPC engine for computationally intensive analysis, and then returns the results to the user. The entire operation requires only a single sign-on (grid-proxy-init) on the user's workstation.

With reference to Figure 1, in step 1, the application on the user's workstation searches for and requests one or more datasets via HTTPS with GSI authentication. In steps 2-3, CGI programs running on the web server verify that the user is privileged to access the requested dataset. In steps 4-7, the requested data is extracted from the MIMAS data repository and copied to a short-lived file, and an XML "ticket" is returned to the application.

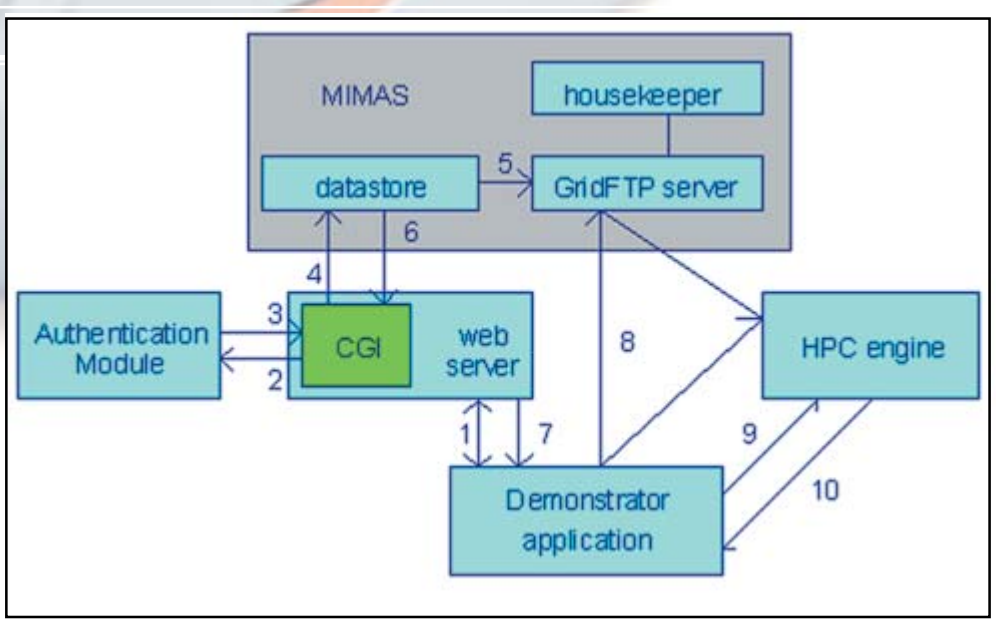


Figure 1: The SAMD Architecture

Actions performed in step 8 are based on information supplied in the ticket, the application uses GridFTP to initiate a third-party file transfer from MIMAS to the HPC engine. Steps 1-7 are repeated for each group of datasets. In steps 9 and 10, Globus mechanisms are used to launch an analysis run on the HPC engine and to retrieve the results. A housekeeping task cleans up temporary files.

User Interface

The user interface, shown in figure 2, has a number of features: "Certification" creates a proxy credential."Data Acquisition" functionally recreates the current web-based interface to the NS datasets from within the demonstrator application."Data Analysis" allows the user to search for and select a target HPC machine for the analysis stage, to locate an executable to perform the analysis, to launch the remote job, and to retrieve job status or results. The three list boxes at the bottom of the Data Acquisition panel (left-to-right) are used to show the time series that match the search criteria, select a subset of these, and transfer them to the HPC Engine.

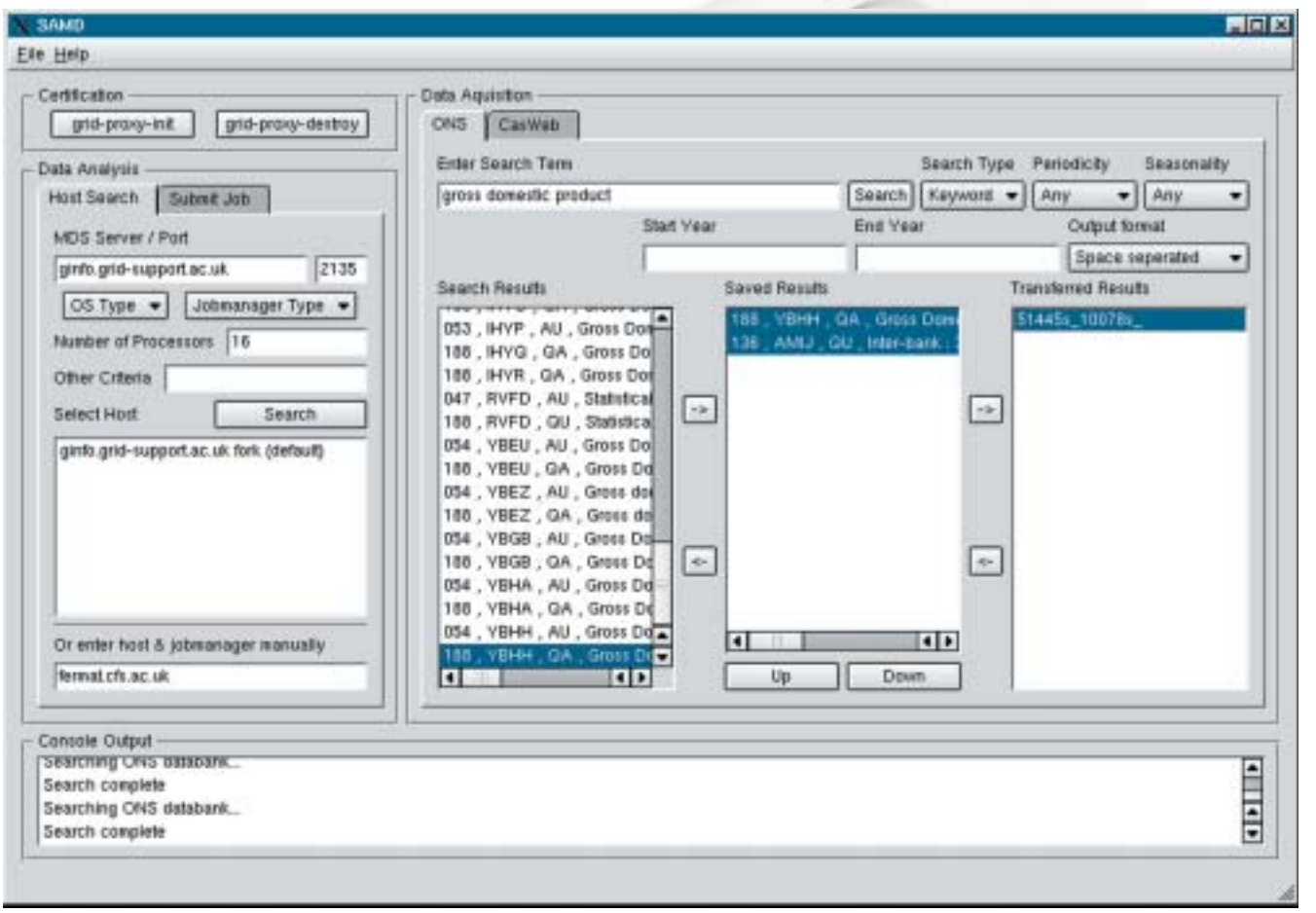
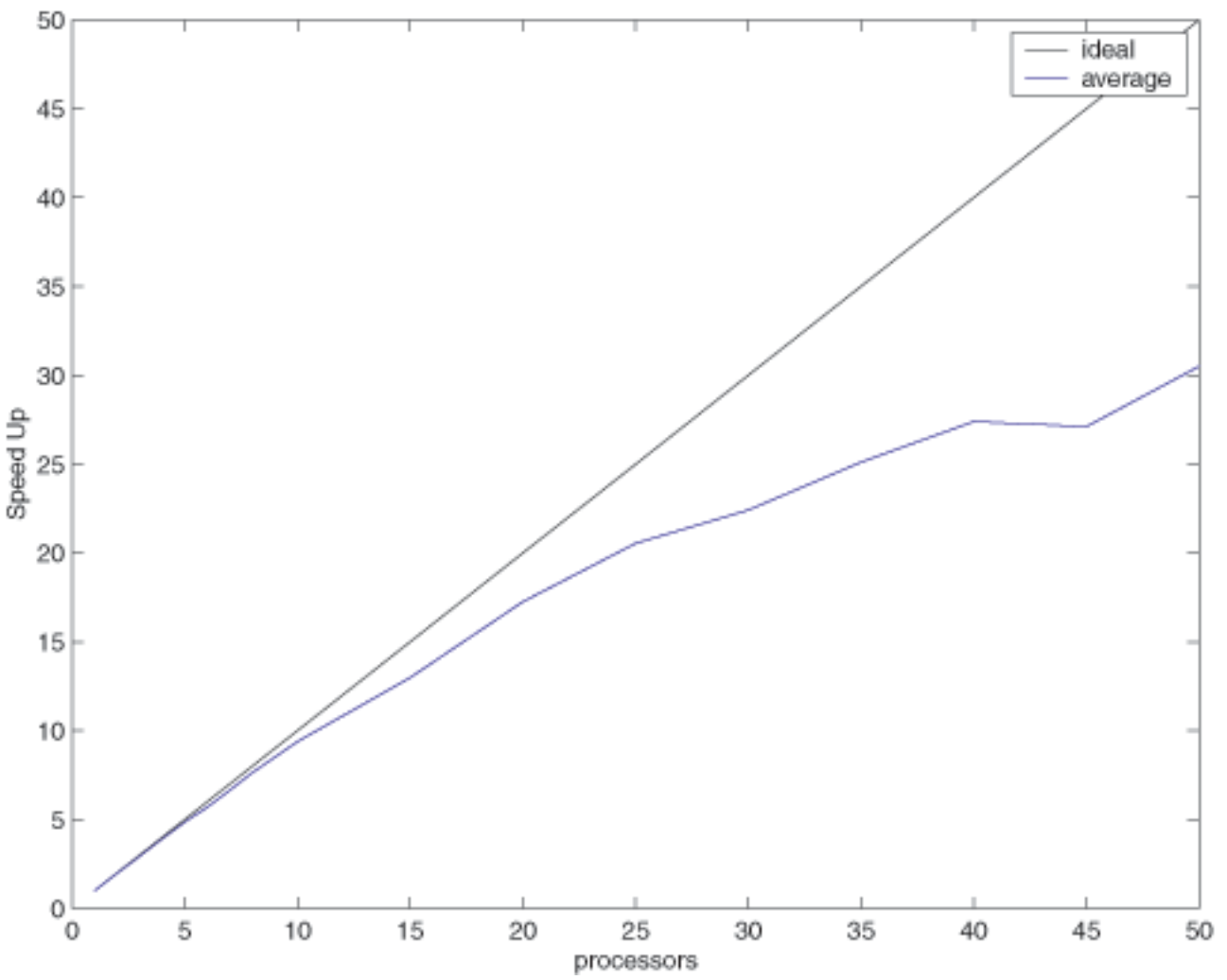


Figure 2 The Graphical User Interface

The demonstrator problem chosen for the SAMD project examines the asymmetric effects of interest rate changes on the UK's GDP. This is a substantive research question of genuine policy interest, and is based on data drawn from the National Statistics datasets hosted by MIMAS. The National Statistics Time Series Data (formerly the NS Databank) contains some 55,000 time series in 34 major datasets from the Office for National Statistics relating to economics, trade, employment and industry. A subset of time series related to interest rates and GDP is selected via the SAMD interface and then sent to an HPC engine on the Grid for analysis – an array of first difference of the logarithm of real GDP is compared to a series of arrays of earlier changes in interest rates. The results show that changes in interest rates have greater effecton output when past growth has been high than when past growth has been normal or negative, i.e. the effect of interest rate changes on GDP varies over the business cycle. A version of the analysis already existed as a Gauss program, taking 40 minutes to run to completion on a small dataset. We re-implemented the algorithm in Fortran 90, and then parallelised it using OpenMP directives. Converting to Fortran 90 improved the performance by a factor of 8 and we observe very good parallel efficiency.



.

Figure 3 Parallel Speedup on large processor counts

Authentication and Authorisation

The SAMD demonstrator application communicates with the web server using HTTPS, an approach that permits a high level of re-use of existing CGI code in the server. Authorisation is implemented by an addition to the original CGI scripts. We use standard Apache mechanisms for authentication but found it necessary to make modifications to the SSL module to handle proxy credentials.

Conclusion

SAMD demonstrates the successful incorporation of emerging Grid technologies into an existing social science data service. It shows how the integration of access to both data and computational resources within a single sign-on environment enables the automation of complex workflows, facilitating the scaling up of social science research applications. Finally, it shows how adding programmable interfaces (protocols) to existing services facilitates the development of third-party, value-adding client applications

The SAMD project funded by the ESRC and the DTI (DTI Ref: THBB/008/00082C).