Journal of
Applied
Crystallography

ISSN 0021-8898

Editor: **Gernot Kostorz**

ECSES – examining crystal structures using 'e-science': a demonstrator employing web and grid services to enhance user participation in crystallographic experiments

Simon J. Coles, Jeremy G. Frey, Michael B. Hursthouse, Mark E. Light, Ken E. Meacham, Darren J. Marvin and Mike Surridge

Journal of
Applied
Crystallography
ISSN 0021-8898

Received 26 February 2005
Accepted 5 August 2005

ECSES – examining crystal structures using 'e-science': a demonstrator employing web and grid services to enhance user participation in crystallographic experiments

Simon J. Coles,[a]* Jeremy G. Frey,[a] Michael B. Hursthouse,[a] Mark E. Light,[a] Ken E. Meacham,[b] Darren J. Marvin[b] and Mike Surridge[b]

[a]School of Chemistry, University of Southampton, Southampton, Hampshire SO17 1BJ, UK, and [b]IT Innovation Centre, 2 Venture Road, Chilworth Science Park, Southampton, Hampshire SO16 7NP, UK. Correspondence e-mail: s.j.coles@soton.ac.uk

An application of e-science methodology and grid networking technology is presented that opens up new possibilities to enhance the operation of large high-throughput service-crystallography facilities, exemplified by the UK National Crystallography Service (NCS). A seamless distributed computing approach is used to provide remote secure visualization, monitoring and interaction with the laboratory and the diffraction experiment, supervision and input to the data workup and analysis processes, and to enable dissemination and further use of the resulting structural data. The architecture of the system is based on web and grid services (in particular the use of *Globus*, v1.1.4), which provide a secure environment for two-way information flow and communication between the service users and operators. This capability will enhance operations of instrument and software automation by providing more efficient use of the resources, increasing the throughput of samples and enabling interactions with distributed chemistry information databases, computational services and networks. The viability of these interactions is assessed and directions for future crystallography services suggested. The setup would be equally applicable to protein or powder crystallography services.

1. Introduction

The UK Engineering and Physical Sciences Research Council (EPSRC, http://www.epsrc.ac.uk) National Crystallography Service (NCS, http://www.ncs.chem.soton.ac.uk/) is a facility that provides X-ray crystallographic data collection facilities or full structure solution services to the UK academic chemistry community. The NCS has a throughput in excess of 1000 samples a year in a laboratory environment that processes more than 2000 data sets per annum. As part of a UK National e-Science Development Program (http://www.rcuk.ac.uk/escience/) the NCS gained foundation funding from the Department of Trade and Industry (DTI, http://www.rcuk.ac.uk/escience/documents/ka-middleware.pdf) to develop a 'proof of concept' demonstration outlining how the grid could enable an e-science enhancement for structural chemistry. This activity formed the initial demonstrator product for an EPSRC funded pilot project, 'Comb*e*Chem' (http://www.combechem.org; Frey *et al.*, 2003), which aimed to integrate existing experimental equipment and data sources in the chemical community and augment them within a grid-based information and knowledge environment, providing the desired end-to-end connectivity from the laboratory to the literature (and back again), and thus the basis for automation and tracking of knowledge discovery and modelling.

The grid (Berman *et al.*, 2003) is sometimes referred to as 'the next generation internet', and provides a middleware infrastructure for delivering global access to distributed resources across different administrative domains. Middleware is therefore a set of services and software libraries for handling security, information infrastructure, resource management, data management, communication and fault-detection processes. e-Science is considered as the enabled science running over the grid. The e-science grid (Foster, 2002, 2003) is intended to be a flexible, secure and coordinated infrastructure, comprising dynamic computational facilities, laboratory instruments, data storage and retrieval, and wide-area networks. Ultimately this will enable seamless interaction between research institutions, forming a virtual organization as well as providing access to expertise and distributed resources of all kinds. Currently the grid is a pilot project for use by academics, but is also expected to become central to commerce and industry (*e.g.* http://www.gria.org or http://www.informit.com/articles/article.asp?p=102223) as security and seamless access are developed (Surridge, 2002).

doi:10.1107/S0021889805025197

Compared with a conventional networking solution, the implementation of a grid infrastructure has a number of advantages that would be of considerable worth to service crystallography (von Laszewski *et al.*, 2000). However, secure operations akin to the levels used to run 'e-commerce' over grid/web services would be required to protect data confidentiality, assert intellectual property rights and operate a facility that would be attractive to the commercial sector. The primary benefits would be in situations where: the users and the service are geographically separate; specific details crucial to the success of the data collection could only be supplied by the sample provider at the time of the experiment; a remote user could steer an experiment, thus allowing service staff to concentrate on other matters or maximize diffractometer usage time.

Once the raw diffraction data have been acquired, the grid can provide distributed software resources for the workup and analysis of crystal structures, further data mining and 'value added' exercises (*i.e.* computational services following data collection and structure refinement). The grid may similarly facilitate the efficient management of the data and rapid dissemination of results.

2. Objectives

The primary objective of the ECSES project was to build an e-science demonstrator to investigate the potential for building a service for remote users to collaborate in experiments being performed at a large service crystallography facility and seamlessly perform 'value added' work on the result. The grid provides a mechanism whereby either an expert user (*i.e.* one skilled in the art of crystallography) or a non-expert user [*i.e.* a synthetic chemist unskilled in the art, who requires structural information on their compound(s) of interest] may have significant involvement in an experiment remotely. These operations may be performed to some extent independently of the service personnel, or alternatively contribute specific knowledge of the sample to assist the service personnel. In order to demonstrate how a grid environment could enhance this type of collaborative crystallography, a prototype environment for the acquisition of data and the generation, exploration and further studies on crystal structures was constructed. The intention was to investigate current technologies and from the experience gained in design, construction and deployment of such a facility, produce a set of outcomes and guidelines to inform the e-science and crystallographic communities. These technologies included the operation of different high-speed multicasting procedures on a grid operating system. Perhaps more importantly, however, the suitability of current crystallographic software [*i.e.* diffractometer control software and the Cambridge Structural Database, CSD (Allen, 2002) suite of programs] to a grid environment could be assessed.

The ECSES project also provided insight as to how grid technologies might be developed or improved for the forthcoming EPSRC e-science 'testbed' projects. In particular the outputs from the project described here were used as a guide for the construction of a service, as part of the Comb*e*Chem e-science testbed project, to be integrated into the operation of the EPSRC UK National Crystallography Service (NCS). The details of the construction and deployment of the NCS service is to be the focus of a separate paper.

The ECSES demonstrator was deployed from the University of Southampton (*i.e.* inside the university firewall security structure, but outside the NCS firewall) to the UK academic network (NeSC, Edinburgh, April 2002), a commercial site (IBM, June 2002) and outside the UK (Supercomputing, Baltimore, USA, November 2002).

3. The ECSES demonstrator description

3.1. The ECSES scenario and 'experiment' workflow

In order to develop the ECSES system, a suitable scenario, typical of a structural and materials chemistry application, was created that included a typical set of tasks that an end user might wish to carry out, and which an e-science environment could or should facilitate. The details of the workflow for a service interacting with a remote user, to collaborate in performing a crystallographic experiment, are presented in Appendix *A*. The scenario devised to capture and enact this workflow is as follows.

A scientist researching in the field of optically active materials design, applied to sensor technology, requires materials which can function at a high operational temperature. The scientist discovers a candidate material with good optical properties; however, its melting point is found to be too low for operational use and alternative candidate materials must be sought. Knowing that the solid-state physical properties of these materials are related to their crystal structure, the researcher intends to fine-tune aspects of the structure with the aim of generating a related, or derived material that would exhibit similar properties and a higher melting point. The steps required to achieve this are numbered on Fig. 1 and are as follows.

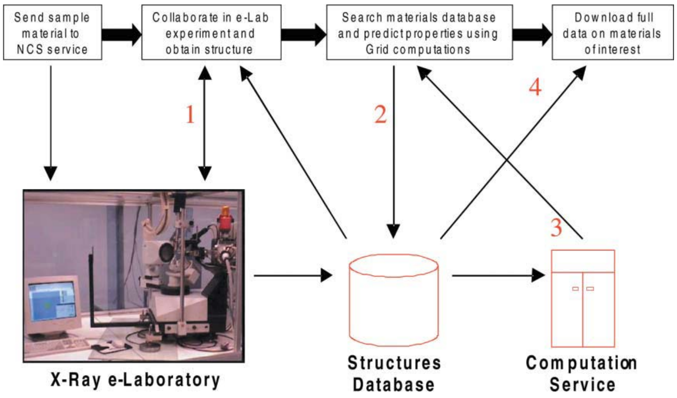
Send sample material to NCS service
Collaborate in e-Lab experiment and obtain structure
Search materials database and predict properties using Grid computations
Download full data on materials of interest
1
2
3
4
X-Ray e-Laboratory
Structures Database
Computation Service

Figure 1
Flowchart depicting the experiment workflow and points of interaction between the user and the service.

(1) Determine the crystal structure of the reference material through collaboration with a service crystallography facility.

(2) Query existing structure databases for materials of similar structure.

(3) Calculate melting points for these alternative candidate materials.

(4) Refine the selection of materials, by sorting on melting points and selecting those above the operational temperature cut-off.

To determine the crystal structure, the researcher submits a sample of the reference material to the crystallography service for X-ray diffraction analysis. The researcher then collaborates with the service, *via* the ECSES environment, to perform the experiment. This interaction enables the researcher to do the following.

(*a*) View images being captured from the diffractometer in real time.

(*b*) View information, such as unit-cell parameters, as it is calculated.

(*c*) Discuss the quality of the images and data with the service crystallographer, *via* audio and video conferencing.

(*d*) Be involved in the decision making as to whether to continue with a full data collection, or abort the experiment after the initial prescans or unit-cell calculation.

(*e*) Discuss the optimum data collection parameters with the crystallographer.

On completion of the data collection, the crystal structure is solved, refined and staged in a database local to the service. The researcher may immediately download the structure to view and analyse it using preferred, local structure visualization software.

The structure may then be submitted as a query (*via* the ECSES grid environment) to a remote database *via* the grid, in order to search for compounds with similar crystal structures. The resulting structural data file is then staged to a convenient location and a remote melting-point grid computational service invoked. To avoid lengthy download times for potentially large search results, selected details (*i.e.* melting-point value and pertinent CSD reference code) are returned to the researcher and displayed in a summary table. These are ordered in terms of melting point and those compounds with melting points which lie above the required operational temperature cut-off are selected. Finally, the researcher queries the remote structure database (*via* the ECSES grid environment), to return the full crystal structure data relating to the materials of interest, thus only moving all the relevant data at the final stage of the investigation.

3.2. Features

The platform for the development and enactment of the ECSES scenario was the grid-services-based system *Globus*, v1.1.4 (http://www.globus.org; Foster & Kesselman, 1997), which was emerging as one of the leading grid technologies (Takemiya *et al.*, 2003), and by standardizing on technology it was envisaged that the various grid services being developed would eventually become interoperable. Appendix *B* describes in detail the various aspects of the architecture of the ECSES system as it was constructed on the *Globus* platform and outlines some of the software tools employed.

Features and particular software aspects of the ECSES demonstrator included the following.

(i) Single sign-on security framework using *Globus Security Infrastructure* (*GSI*); secure access, monitoring and basic steering of a diffraction experiment, including video and audio links to the laboratory.

(ii) User involvement in the decision-making process during an experiment, with respect to *e.g.* the experimental strategy and data collection parameters; downloading and visualization of X-ray diffraction images in real time.

(iii) Secure access *via* a web service to the CSD, running in the NCS laboratory; visualization of extracted structures, using the *Mercury* (Bruno *et al.*, 2002) application from the Cambridge Crystallographic Data Centre (CCDC).

(iv) Use of the CCDC *ConQuest* application to launch an e-science structure–property query across the grid; staging of exported structure files using the *Globus* file handler transfer system.

(v) Example of a structure-based calculation, in this case a melting-point calculation based on the exported structural data, launched by *Globus* on a remote job submission.

4. Results, benefits, outcomes and recommendations

The ECSES demonstrator fulfilled its role by discovering and presenting not only the positive and achievable aspects of developing a remote and interactive crystallographic service *via* a distributed network, but also highlighted those aspects that would be difficult or inadvisable. This section outlines recommendations for the construction, deployment and use of such a system.

4.1. Positive outcomes

The following aspects of the ECSES demo easily enhanced the operation of a crystallography service open to remote users and are relatively easily constructed and operated on a grid infrastructure.

(*a*) The ability to view and assess raw data as generated by instruments and participate in decision-making points during an experiment. This is one of the primary benefits of such a service for both a trained crystallographer and the originator of a new compound (*i.e.* the synthetic chemist). For the crystallographer, the quality of the diffraction pattern may easily be gauged, allowing an assessment of whether the experiment should proceed (at decision-making points or during the collection should the crystal denature in the X-ray beam) or if the correct data collection strategy has been applied (*e.g.* has the correct unit cell been chosen, is the exposure time suitable, should higher resolution scans be performed, *etc.*). For the chemist, a better understanding of the issue of crystal quality can be developed and they can supply useful chemical information to the service in the timeframe of the diffraction experiment. For both types of user, it would also be beneficial for an image of the bulk sample and that of the selected crystal to be transmitted to the user so that the crystal quality may be discussed and understood. The points to consider for this aspect of the service are the number and file size of the images which, coupled with the quality and speed of the network connection, determine the speed of download to the users' machine. Raw image files as they are generated by the diffractometer are typically very large in size and in a binary form, which makes viewing images *via* the commonly available and multifunctional software used to access the service, *i.e.* a standard internet browser, impossible. For this reason, the diffractometer images were converted to a common image format (*i.e.* JPG in this case) before being made available for download. The disadvantage here is that the trained crystallographer would not be able to download and manipulate these images with local software (assuming software licenses are compatible, see below), but current bandwidth availability is most likely to dominate in this situation. The emergent standard, IMGCIF (Hammersley *et al.*, 2003), could provide some answers to these questions; however, the file size problem would remain. The outcomes of operating in this way would be significant gains in the amount of service time saved, through reduction of repetitions or experiments of little worth.

(*b*) Staging of exported structure files using the *Globus* file handler transfer system. This allows efficient data file management and provides an efficient mechanism for routing data files to pertinent services.

(*c*) Seamless access to computational resources. The melting point calculation performed in the demonstrator provides an example of the efficient staging and routing of data to follow-on grid-based computational services. This has obvious benefits for subject areas that rely on crystal structure determinations as a starting point for study. Moreover, providing this service in a grid-based environment allows access to an enormous amount of computational power and thus facilitates computationally expensive calculations, *e.g.* molecular dynamics simulations.

(*d*) Communication between the user and the laboratory. The multicasting facilities (video and audio), enable vital information that has not been previously provided to be exchanged between the laboratory and the user at precisely the time it is required. Additionally, both parties can be involved in devising a data collection strategy which fits the purpose of the experiment (*e.g.* a simple data collection strategy may be employed if purely the connectivity is required, or conversely a complex strategy would be calculated should the user require a high-resolution study), which would maximize efficient use of the service and its experimental facilities.

(*e*) Transferability between different types of crystallographic experiment. The ECSES demonstrator exemplifies the single-crystal diffraction experiment; however, the infrastructure employed here would be equally applicable to different services, such as those providing protein or powder crystallographic expertise and facilities. The possibility of using such a system in a training or educational role, as well as operating a crystallographic service, exists and will be investigated as part of a following study.

4.2. Issues arising

The intention of the ECSES demonstrator was not only to be a proof of concept, but also to test current technologies for their suitability and highlight aspects that would require attention when constructing a service. As such there are a number of points outlined below that should be considered.

(*a*) User authentication and security. In order that the integrity of a user's data, the system provided by the service and the service providers domain (*e.g.* in this case the university) are not compromised, a security model must be implemented to provide a degree of 'privacy' and guard against malicious attacks. The grid has been designed for the seamless sharing of data between scientists and consequently the issue of security had been somewhat overlooked at the time when the ECSES demonstrator was built. As a result, this model would be contrary to that which would be desirable to industry and commerce. Moreover, research scientists are also operating in increasingly competitive environments and require protection and privacy over their data. Thus there is a need for an infrastructure that provides security yet promotes the exchange of data, and a user of such a system must be able to operate at many different levels. At the time of construction (*Globus*, v1.1.4) the security available was a single sign-on security framework using *Globus Security Infrastructure* (*GSI*), *i.e.* password protection allowing access to files with certain 'permission levels'. It was immediately obvious that this model did not provide a sufficiently flexible fine-grained security structure to take care of the need to differentiate between the external and internal owners of the data. It is clear that a more complex model must be implemented, whilst maintaining ease of use; however, this was too large a task to achieve within the timescale of the project.

(*b*) The use of grid, as opposed to web, services. Grid services (outlined in Appendices *A* and *B*) are essentially an integrated set of components that are configured to run on the same architecture and may be tailored to the requirements of the system being built. Web services, however, are much more bespoke and generally built *ad hoc* for a system and conform to a much broader set of standards, *i.e.* those compatible with the World Wide Web. The ECSES demonstrator set out to evaluate grid services (*Globus*) and thus a direct comparison is not possible. However, some conclusions can be made. It was immediately evident when constructing the demonstrator that the interface between the scientific instrument (and the data that it outputs) and the grid system would provide the most problems. Once data files had entered the grid, their management, ability to perform calculations *etc.*, was a relatively achievable task. Thus the interface between the diffractometer and *Globus* involved construction of bespoke code, which might well have been better performed using web services. However, considerable interaction with diffractometer manufacturers would be required in order to modify their instruments to become grid-compliant and their awareness in this area is only now coming to light.

(*c*) Scheduling the experiment. Commencing the crystallographic experiment during the operation of the demonstrator required precise timing for both parties involved to be available and initiate certain aspects of the system. For the demonstrator this involved a telephone exchange, but it is clear that an operational system would require a more robust and less intrusive method for simultaneous presence and initiation of the remote experiment.

(*d*) Remote instrument control. This aspect of the demonstrator has many facets to be considered and highlighted a number of issues. The first consideration is that of operating a service by remote instrument control and automated routines ('dark laboratory') against that of utilizing multicasting technologies to enable collaboration between the laboratory and its distant user. The ECSES demonstrator opted for the latter, as there are aspects of the dark-laboratory approach that would be difficult to realise. These issues include: health and safety, where the consequences of an experiment going wrong could endanger life (*e.g.* a shutter being open when interlocks are also open); legal implications with working out of hours; potential compromise or damage of the diffraction apparatus, whether accidental or malicious; the inability to communicate with the laboratory should there be logistical or scientific problems, *etc.* Therefore an environment was constructed whereby the remote user could communicate and collaborate with a crystallography service; however, there are also issues with this mode of operation.

(*e*) Video conferencing, audio multicasting and instant messaging. These services require a reasonable amount of management during a session from both ends of the exchange. The service user must install software and hardware to enable a full two-way exchange with the service and then operate it during the session, and whilst the service must do likewise, it must also install video-conferencing facilities that enable the user to visualize the experiment. The multicasting infrastructure was available as part of the Access grid rollout in the UK, with a unicast bridge available where multicast facilities could not be provided. However, the greatest problem with multicasting facilities was that of network bandwidth. The issue of insufficient, or irregular, bandwidth can greatly influence how effective an interaction can be, as transmissions slow down or become intermittent with fluctuations in availability. Using the example of the demonstrator operating between the National e-Science Centre and the NCS (at Southampton University), the ECSES system ran perfectly until an Access grid lecture was broadcast to the National e-Science Centre from Southampton University. In this case, the Southampton University institutional firewall limited the effective bandwidth. It is likely to be the situation that whilst a service can maximize available bandwidth, its users will not necessarily be able to do this. In addition, mismatched bandwidth availability between two sites is likely to cause slowing problems if additional rate-limiting hardware is not incorporated into the system.

(*f*) System security. The *Globus* (v1.1.4) grid operating system operates a single sign-on security model where a user can authenticate with the grid and then seamlessly access any resources. Whilst this approach worked for the ECSES demonstrator, it was clear that the design of an operational service would have to consider a much tighter security model. The aspects of the service that would have to be protected include confidentiality of data and protection of associated proprietary rights, protection of the service hardware and software, and protection of the institutional computing infrastructure from a malicious attack.

(*g*) Software licensing and provision. The seamless provision of data workup and analysis software would be difficult to provide without compromising software licence agreements. For example, the ECSES demonstrator interrogated the CSD to find similar structures and if this service were made available to all NCS users, irrespective of whether they were licensed to use it or not, then the NCS would be in breach of their agreement with CCDC. Moreover, considerable work had to be performed on the internals of the CSD software so that it could interface and network with grid software. Both of these points are serious issues for grid and e-science communities to overcome, but should begin to become less severe as these scientific practices become more common and software providers build and license their code for use on the grid.

5. Conclusions

The demonstrator service presented describes a 'proof of concept' of remote monitoring, interaction and control of an experiment being conducted by an existing crystallographic service. Whilst it is clear that there are a number of issues to be addressed before such an infrastructure may be universally deployed, the demonstrator has highlighted not only the problems to be overcome but also directions in which service crystallography can develop. Issues such as secure and authenticated access to the system, traversing multiple firewalls and full integration into an existing experimental service have arisen and will require considerable attention. However, it is clear that the seamless interaction between a remote user (either a trained crystallographer or the synthetic chemist who provides the sample) and experiment is highly desirable as the client may provide information which is key to the success of the experiment. In addition, such a facility promotes more user engagement with a crystallographic service and generates a more responsive service which can provide instantaneous feedback to its users. The use of such a capability for training purposes is not a further major advantage. Moreover, in an e-science-enabled automated, or semi-automated, environment, a trained remote user would be able to conduct a full crystallographic structure determination and perform further *in silico* investigations on the result at the same or other sites on the grid.

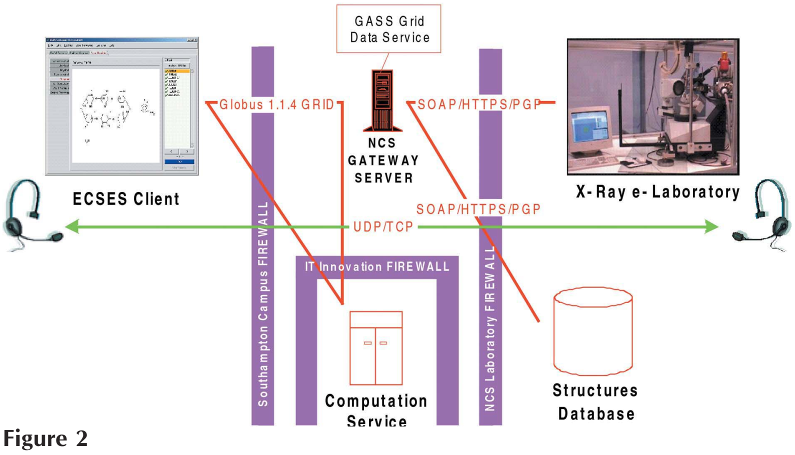

Figure 2
Basic ECSES architecture and network, covering the process for receipt of sample, structure determination and subsequent property calculation.

APPENDIX *A*
Workflow analysis

A detailed analysis of both the human and machine workflows required to obtain and reuse crystallographic data at the NCS was undertaken and abstracted for the ECSES demo. The resulting pathway is shown from several perspectives in figures deposited as supplementary material.[1] These workflows highlight the steps that humans and automated routines need to take. It is absolutely imperative that the workflow is accurately captured and assessed before embarking on the design of such a complex computing infrastructure. Figs. A and B outline the interplay between different components of the laboratory workflow and practice, whilst Fig. C relates each component arising from a step in the workflow to a 'state' that the sample is in at that point in the whole experimental process (figures deposited).

APPENDIX *B*
Architecture, security and grid technologies

B1. Introduction

Fig. 2 gives a high-level overview of the ECSES architecture, showing the main components and their arrangement within the Southampton University network and its respective firewalls. Since the intention was to provide access to NCS facilities from outside the university, it was considered sensible to restrict access further by adding a firewall at NCS; this would also secure the rest of the university from any potential breakout from the NCS systems if they were hacked. Scanning attempts at just such attacks were registered when ECSES was demonstrated at the Supercomputing 2003 conference. An SGI PowerChallenge server was made available for running calculations in a remote location outside the NCS domain (IT-Innovations Centre, Southampton) and provided a demonstration of how one might employ structure data generated by the NCS by running calculations at a remote site *via* the grid infrastructure.

The structural data would normally be stored securely within the NCS domain, so direct access by a remote property (*i.e.* melting point) calculation service would not be possible or advisable. This indicated the need for some form of data-staging post, where data could be extracted from the NCS database and stored temporarily in such a way that it was directly accessible to the computation service. For this reason, a gateway server (or 'stepping stone') was placed physically at the NCS, but logically outside the NCS firewall (though inside the university firewall).

The client end (*i.e.* the remote researcher making use of the NCS facilities) was the main focus of the ECSES demonstrator, and was therefore one of the most highly developed parts of the system, in order to show the potential users of the system what it would be like to interact with a grid-based instrument service. The client software consisted of several components, some of which were proprietary software, the modification of which involved collaboration with the providers in order to ensure compatibility with grid technologies. The principal components were as follows.

(i) The e-science portal provided an interface to the grid (*Globus*) system, together with a graphical user interface (GUI) front end to the NCS service. Various items of experimental data were displayed in different panels (*e.g.* X-ray images, unit-cell data), and most of the user control was provided through this interface. The e-science proxy also drove remote melting-point calculations and displayed the resulting data in a table.

(ii) *ConQuest* (CCDC) is the primary program for searching and retrieving information from the Cambridge Structural Database (CSD). *ConQuest* provided a full range of text/numeric database search options, in addition to more complex search functionality such as chemical substructure searches.

(iii) *Mercury* (CCDC) is a comprehensive range of tools for structure visualization and the exploration of crystal packing, which were used to visualize the CIF structure file(s) returned from NCS.

(iv) *VIC* (*Video Conferencing Tool*; http://www-nrg.ee.lbl.gov/vic/) was developed by the Network Research Group at the Lawrence Berkeley National Laboratory in collaboration with the University of California, Berkeley. It provided video feeds from webcams both at the client end and at the NCS.

(v) *RAT* (*Robust Audio Tool*) is an open-source audio conferencing tool (http://www-mice.cs.ucl.ac.uk/multimedia/software/), which provided real-time verbal communication.

B2. Grid technologies

At the beginning of the ECSES project, *Globus* (http://www.globus.org) (Foster & Kesselman, 1997) was emerging as one of the leading grid technologies, and provided probably

[1] Supplementary data are available from the IUCr electronic archives (Reference: HE5333). Services for accessing these data are described at the back of the journal.

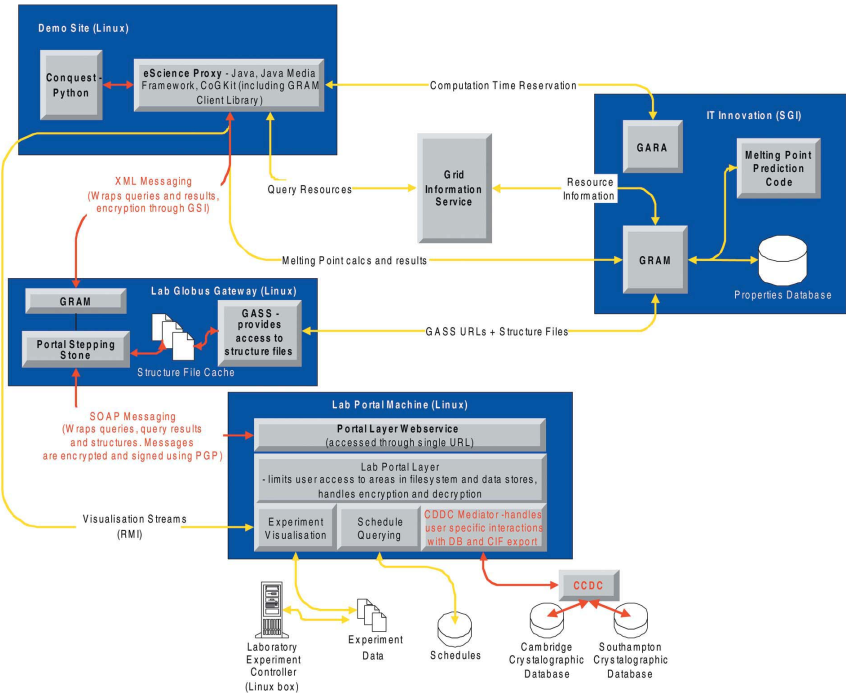
Demo Site (Linux)
Conquest-Python
eScience Proxy - Java, Java Media Framework, CoG Kit (including GRAM Client Library)
Computation Time Reservation
IT Innovation (SGI)
GARA
Melting Point Prediction Code
XML Messaging (Wraps queries and results, encryption through GSI)
Query Resources
Grid Information Service
Resource Information
Melting Point calcs and results
GRAM
Properties Database
Lab Globus Gateway (Linux)
GRAM
Portal Stepping Stone
GASS - provides access to structure files
Structure File Cache
GASS URLs + Structure Files
Lab Portal Machine (Linux)
SOAP Messaging (Wraps queries, query results and structures. Messages are encrypted and signed using PGP)
Portal Layer Webservice (accessed through single URL)
Lab Portal Layer - limits user access to areas in filesystem and data stores, handles encryption and decryption
Visualisation Streams (RMI)
Experiment Visualisation
Schedule Querying
CDDC Mediator -handles user specific interactions with DB and CIF export
CCDC
Laboratory Experiment Controller (Linux box)
Experiment Data
Schedules
Cambridge Crystalographic Database
Southampton Crystalographic Database

Figure 3
Detailed ECSES architecture, showing in particular the different administrative domains and the firewalls that separate them.

the most complete grid environment available at that time (Takemiya *et al.*, 2003). For this reason, it was adopted as the preferred grid technology for early UK pilot grid projects (including ECSES) with the hope that, by standardizing on technology, the various grid services being developed would eventually become interoperable.

Globus (v1.1.4) consists of the following core components.

(i) *GRAM* (*Globus Resource Allocation Manager*) (Nabrzyski *et al.*, 2003) provides a common interface for remote submission, monitoring and control of computation jobs, within a *Globus* environment.

(ii) *GARA* provides programmers with convenient access to end-to-end quality of service (QoS) for programs (Foster *et al.*, 2004); it provides mechanisms for making QoS reservations for computing resources.

(iii) *GIS* (*Grid Information Service*) provides details about the properties of computers and networks within a *Globus* environment (Czajkowski *et al.*, 2001).

(iv) *GSI* (*Grid Security Infrastructure*) provides generic security services for *Globus* applications, *e.g.* secure login (authentication) to a site (Welch *et al.*, 2003). Authentication is done using X.509 certificates, assigned by the *Globus Certification Authority* (*Globus CA*). Sites also maintain a 'map file', which contains a list of *Globus* users (as distinguished names, DNs) who are allowed access.

(v) *GASS* (*Globus Access to Secondary Storage*) provides programs and advanced programming interfaces (APIs) for remotely accessing data. Data on a particular machine are managed *via* a *GASS* server, which may be contacted from a *GASS*-enabled program to read or write data on the remote machine.

The *Globus Toolkit* provides programmers with an implementation of *Globus*. Various APIs are available, including C and Java interfaces to *Globus* functionality. In particular, the *Globus CoG Kit* provides Java libraries which may be used to enable a Java program to work with *Globus*.

B3. Detailed architecture and implementation

A considerable understanding was required for the design, analysis, implementation and operation of such a service. Fig. 3 shows a detailed picture of the overall architecture of the ECSES system. Interaction between the client and the NCS laboratory was achieved using the following steps.

(i) e-Science proxy generates an XML (extensible markup language) file containing a query (or message) for the laboratory.

(ii) The XML query is staged to the *Globus Gateway*.

(iii) 'Stepping-stone' software is launched on the gateway, using the staged XML query as input.

(iv) Stepping-stone software converts XML into a SOAP (Secure Object Access Protocol) message and sends this to the NCS web service (encrypted and signed using PGP, Pretty Good Privacy).

(v) The NCS web service translates the SOAP message into a call to a local operation.

(vi) Query results are returned using the reverse procedure to the above.

B3.1. Melting-point calculation service. The melting-point prediction is a simulated calculation (as no such code currently exists) that was designed solely for demonstration purposes to show what could be achieved in a grid environment. The code only had to provide relatively meaningful melting points, in a reasonable range, and be related to the structure of the compound. To enable this program to interact with a *Globus* environment, the standard C I/O functions (*i.e.* fopen and fclose) were replaced with equivalents from the *Globus* toolkit (*e.g.* globus_gass_fopen and globus_gass_fclose). Whilst the standard fopen command opens a file, given its local file name (*e.g.* 'aspirin.cif'), globus_gass_fopen command takes a *Globus* URL (*e.g.* https://nevis.chem.soton.ac.uk:1505/home/kem/steppingStone/aspirin.cif), which identifies a remotely staged file. In this example a *Globus GASS* server was running on the gateway server, which accepts requests on port 1505. In a similar way, a melting-point result could be output to a remote URL on the client machine, *e.g.* https://sauvignon.it-innovation.soton.ac.uk:1502/usr/local/ecses/escienceproxy/rundir/aspirin.mp.

B4. Implementation problems

A number of fundamental technical issues, discussed below, needed to be resolved during the designing and setup (or relocation) of the ECSES demo. These points were often difficult to overcome and contributed heavily to design strategies for systems built after the ECSES demonstrator.

B4.1. Audio/video conferencing. Although the *VIC* and *RAT* video and audio conferencing tools were well suited for providing a communications link to the NCS laboratory, they were sometimes unreliable due to incompatible sound-card drivers for Linux. A solution was found in the *Open Sound System* (*OSS*) from 4Front Technologies, which is a commercial (supported) version of the Linux sound drivers.

For video conferencing, webcams were located at the NCS client (remote user) and in the NCS laboratory (for both the NCS operator and the diffractometer). The laboratory technician was therefore able to change the video feeds that were transmitted to the remote user to enable different webcam perspectives of the experiment and technician.

B4.2. *Globus*. Installation and configuration of *Globus* (v1.1.4) was far from simple, and took considerable time to compile and deploy. This had to be done on each *Globus* node in the system. Each machine required its own *Globus* certificate, as did each *Globus* user in the system. Once *Globus* was operating between the various machines, it then proved difficult to relocate the client machine without breaking the system (the client had to be moved to different locations for testing, and then eventually to demonstration sites). The *Globus* certificate was tied to the full host name of the machine, but *Globus* authentication is not dependent on IP address so the problem was solved by keeping the same host name for the client, but changing its IP address for its new location. The disadvantage of this approach is the necessary reconfiguration in the *Globus* network of the client IP address, which is an issue for IP-dependent software.

B4.3. Firewall configuration. It was necessary to configure various firewalls to allow access to various *Globus* services (see Fig. A of the supplementary material) and it was also necessary to allow the audio and video streams through using UDP on ports 10000–10003. *Globus* required a TCP tunnel to port 2119 (*Globus Gatekeeper*) and return connections to a range of ports (*e.g.* 15000–15009). *Globus* was to be configured to restrict its choice of return ports to this range (*via* the GLOBUS_TCP_PORT_RANGE parameter) to ensure that it stayed within the range that had been opened.

The authors are indebted to the considerable assistance of system administration personnel both within the Southampton domain (at Information Systems Services and NCS) and at the various demonstration sites, notably the UK National e-Science Centre, Supercomputing 2003 Baltimore and IBM Hursley. Valuable assistance with the staging of *ConQuest* search results to the e-science proxy was provided by the Cambridge Crystallographic Data Centre.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Berman, F., Hey, A. J. G. & Fox, G. C. (2003). Editors. *Grid Computing: Making the Global Infrastructure a Reality*, pp. 9–50. Chichester: John Wiley. (Wiley Series in Communications Networking and Distributed Systems.)

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.

Czajkowski, K., Fitzgerald, S., Foster, I. & Kesselman, C. (2001). *Proceedings of the Tenth IEEE International Symposium on High-Performance Distributed Computing (HPDC-10)*. Piscataway, NJ: IEEE Press.

Foster, I. (2002). *Phys. Today*, **55**(2), 42–47.

Foster, I. (2003). *Grid Computing: Making the Global Infrastructure a Reality*, edited by F. Berman, A. J. G. Hey & G. C. Fox, pp. 51–64. Chichester: John Wiley. (Wiley Series in Communications Networking and Distributed Systems.)

Foster, I., Fidler, M., Roy, A., Sander, V. & Winkler, L. (2004). *Comput. Commun.* **27**, 1375–1388.

Foster, I. & Kesselman, C. (1997). *Int. J. Supercomput. Appl.* **11**, 115–128.

Frey, J. G., Bradley, M., Essex, J. W., Hursthouse, M. B., Lewis, S. M., Luck, M. M., Moreau, L, De Roure, D. C., Surridge, M. & Welsh, A. H. (2003). *Grid Computing: Making the Global Infrastructure a Reality*, edited by F. Berman, A. J. G. Hey & G. C. Fox, pp. 945–962. Chichester: John Wiley. (Wiley Series in Communications Networking and Distributed Systems.)

Hammersley, A. P., Bernstein, H. J. & Westbrook, J. D. (2003). *Image CIF Dictionary (imgCIF) and Crystallographic Binary File Dictionary (CBF)*, Version 1.3.1, ftp//ftp.iucr.org/pub/cif_img_1.3.1.dic.

Laszewski, G. von, Westbrook, M., Foster, I., Westbrook, E. & Barnes, C. (2000). *Cluster Comput.* **3**, 187–199.

Nabrzyski, J., Schopf, J. M. & Weglarz, J. (2003). Editors. *Grid Resource Management*. Dordrecht: Kluwer.

Surridge, M. (2002). *A Rough Guide to Grid Security*, http://eprints.ecs.soton.ac.uk/7286/.

Takemiya, H., Shudo, K., Tanaka, Y. & Sekiguchi, S. (2003). *J. Grid Comput.* **1**, 117–131.

Welch, V., Siebenlist, F., Foster, I., Bresnahan, J., Czajkowski, K., Gawor, J., Kesselman, C., Meder, S., Pearlman, L. & Tuecke, S. (2003). *Twelfth International Symposium on High Performance Distributed Computing (HPDC-12)*. Piscataway, NJ: IEEE Press.