ELSEVIER

Future Generation Computer Systems 18 (2002) 1017–1031

FGCS
FUTURE GENERATION COMPUTER SYSTEMS

www.elsevier.com/locate/future

The UK e-Science Core Programme and the Grid

Tony Hey*, Anne E. Trefethen

UK e-Science Core Programme EPSRC, Polaris House, North Star Avenue, Swindon SN2 1ET, UK

Abstract

This paper describes the £120M UK 'e-Science' (http://www.research-councils.ac.uk/and http://www.escience-grid.org.uk) initiative and begins by defining what is meant by the term e-Science. The majority of the £120M, some £75M, is funding large-scale e-Science pilot projects in many areas of science and engineering. The infrastructure needed to support such projects must permit routine sharing of distributed and heterogeneous computational and data resources as well as supporting effective collaboration between groups of scientists. Such an infrastructure is commonly referred to as the Grid. Apart from £10M towards a Teraflop computer, the remaining funds, some £35M, constitute the e-Science 'Core Programme'. The goal of this Core Programme is to advance the development of robust and generic Grid middleware in collaboration with industry. The key elements of the Core Programme will be outlined including details of a UK e-Science Grid testbed. The pilot e-Science projects that have so far been announced are then briefly described. These projects span a range of disciplines from particle physics and astronomy to engineering and healthcare, and illustrate the breadth of the UK e-Science Programme. In addition to these major e-Science projects, the Core Programme is funding a series of short-term e-Science demonstrators across a number of disciplines as well as projects in network traffic engineering and some international collaborative activities. We conclude with some remarks about the need to develop a data architecture for the Grid that will allow federated access to relational databases as well as flat files.
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: e-Science; Core Programme; Grid

1. Introduction

The term 'e-Science' was introduced by Dr. John Taylor, Director General of Research Councils in the UK Office of Science and Technology (OST). From his previous experience as Head of Hewlett-Packard's European Research Laboratories, and from his experience as Director General of the OST, Taylor saw that many areas of science are becoming increasingly reliant on new ways of collaborative, multidisciplinary working. The term e-Science is intended to capture these new mode of working [1]:

> '*e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it.*'

The infrastructure to enable this science revolution is generally referred to as the Grid [2]. Two examples of such e-Science projects are from particle physics and astronomy. The world-wide particle physics community is planning an exciting new series of experiments to be carried out on the new 'Large Hadron Collider' (LHC) experimental facility under construction at CERN in Geneva. The goal is to find signs of the Higgs boson, key to the generation of mass for both the vector bosons and the fermions of

* Corresponding author. Tel.: +44-1793-444022; fax: +44-1793-444505.
E-mail addresses: tony.hey@epsrc.ac.uk (T. Hey), anne.trefethen@epsrc.ac.uk (A.E. Trefethen).

0167-739X/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-739X(02)00082-1

the Standard Model of the weak and electromagnetic interactions. The experimental physicists are also hoping for indications of other new types of matter such as supersymmetric particles which may shed light on the 'dark matter' problem of cosmology. These LHC experiments are on a scale never before seen in physics, with each experiment involving a collaboration of over 100 institutions and over 1000 physicists from Europe, USA and Japan. When operational in 2005, the LHC will generate petabytes of experimental data per year, for each experiment. This vast amount of data needs to be pre-processed and distributed for further analysis by all members of the consortia to search for signals betraying the presence of the Higgs boson or other surprises. The physicists need to put in place an LHC Grid infrastructure that will permit the transport and data mining of such distributed data sets. There are a number of funded projects in Europe (EU DataGrid [3], EU DataTag [4], UK GridPP [5] and in the USA (NSF GriPhyN [6], DOE PPDataGrid [7], NSF iVDGL [8]) in which the particle physicists are working to build a Grid that will support these needs. Our second example of e-Science is much more directly data-centric. In the UK, the astronomers are planning to create a 'virtual observatory' in their e-Science AstroGrid project. There are similar initiatives in the USA with the NSF NVO [9] project and the EU AVO [10] project. The goal of these projects is to provide uniform access to a federated, distributed repository of astronomical data spanning all wavelengths from radio waves to X-rays. At present, astronomical data using different wavelengths are taken with different telescopes and stored in a wide variety of formats. Their goal is to create something like a 'data warehouse' for astronomical data that will enable new types of studies to be performed. Again, the astronomers are considering building a Grid infrastructure to support these Virtual Observatories. Later in this article we shall describe other types of e-Science and e-Engineering problems that have more obvious interest to industry.

The vision for a layer of 'Grid' middleware that provides a set of core services to enable such new types of science and engineering is due to Ian Foster, Carl Kesselman and Stephen Tuecke. In their Globus project, they have developed parts of a prototype open source Grid Toolkit [11]. Their choice of the name 'Grid' to describe this middleware infrastructure resonates with the idea of a future in which computing resources, compute cycles and storage, as well as expensive scientific facilities and software, can be accessed on-demand like the electric power utilities of today. These 'e-Utility' ideas are also reminiscent with the recent trend of the Web community towards a model of 'Web services' advertised by brokers and consumed by applications, which have recently been brought together in the Open Grid Services Architecture (OGSA) [12].

In the next section, we outline the general structure of the UK e-Science programme and discuss the 'existence proof' of the NASA Information Power Grid (IPG) [13] as the closest example of a working 'production Grid'. We also set our activity in context by listing some of the other funded Grid projects around the world. In Section 3, we describe the UK e-Science Core Programme in some detail and in Section 4 we describe the presently funded UK e-Science pilot projects in engineering and the physical sciences. We conclude with some remarks about the evolution of Grid middleware architecture and the possible take-up of the Grid by industry.

2. The UK e-Science programme

2.1. Funding and structure of the UK e-Science programme

Under the UK Government's Spending Review in 2000, the OST was allocated £98M to establish a 3-year e-Science R&D Programme. This e-Science initiative spans all the Research Councils—the Biotechnology and Biological Sciences Research Council (BBSRC), the Council for the Central Laboratory of the Research Councils (CCLRC), the Engineering and Physical Sciences Research Council (EPSRC), the Economic Social Research Council (ESRC), the Medical Research Council (MRC), the Natural Environment Research Council (NERC) and the Particle Physics and Astronomy Research Council (PPARC). A specific allocation was made to each Research Council (see Fig. 1), with PPARC being allocated the lion's share (£26M) so that they can begin putting in place the infrastructure necessary to support the LHC experiments that are projected to come on stream in 2005. The Central Laboratories at Daresbury and Rutherford (CLRC) have been

allocated £5M specifically to 'Grid-enable' their experimental facilities. The sum of £10M has been specifically allocated towards the procurement of a new national Teraflop computing system. The remainder, some £15M, is designated as the e-Science 'Core Programme'. This sum is augmented by an allocation of £20M from the Department of Trade and Industry making a total of £35M for the Core Programme.

As is common in DTI programmes, the DTI contribution of £20M requires a matching contribution from industry. It is also expected that there will be industrial contributions to the individual Research Council e-Science pilot projects making a total industrial commitment to the e-Science programme of well over £20M. The goal of the Core Programme is to support the e-Science pilot projects of the different Research Councils and work with industry in developing robust, 'industrial strength' generic Grid middleware. Requirements and lessons learnt in the different e-Science applications will inform the development of more stable and function Grid middleware that can assist the e-Science experiments and be of relevance to industry and commerce.

The management structure is also indicated in Fig. 1. The Director of the UK Core Programme is advised by a Grid Technical Advisory Group. Tony Hey has been appointed as Director of the Core Programme. The different Research Council e-Science programmes are represented on an e-Science Steering Committee, chaired by David Wallace, with the Core Programme Director present as a member. The EPSRC are providing programme management for the Core Programme on behalf of all the Research Councils.

2.2. NASA's IPG

Over the last 3 years, NASA has pioneered a new style of computing infrastructure by connecting the computing resources of several of its R&D Laboratories to form the IPG [13] (see Fig. 2).

The vision for the IPG has been enunciated by Bill Johnston, the leader of the activity as being intended to:

> '*promote a revolution in how NASA addresses large-scale science and engineering problems by providing "persistent infrastructure" for "highly capable" computing and data management services that, on-demand, will locate and co-schedule the multi-Centre resources needed to address large-scale and/or widely distributed problems, and provide the ancillary services that are needed to support the workflow management frameworks that coordinate the processes of distributed science and engineering problems.*'

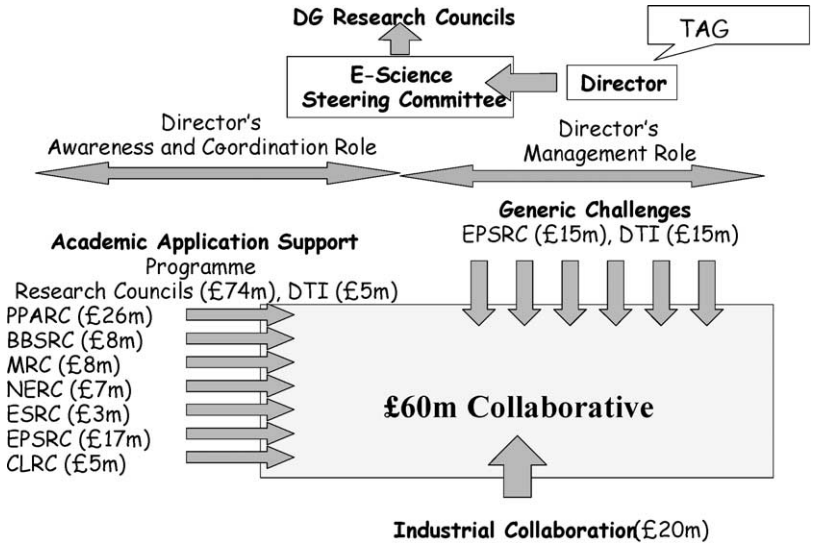

Fig. 1. Structure and funding for UK e-Science Programme.

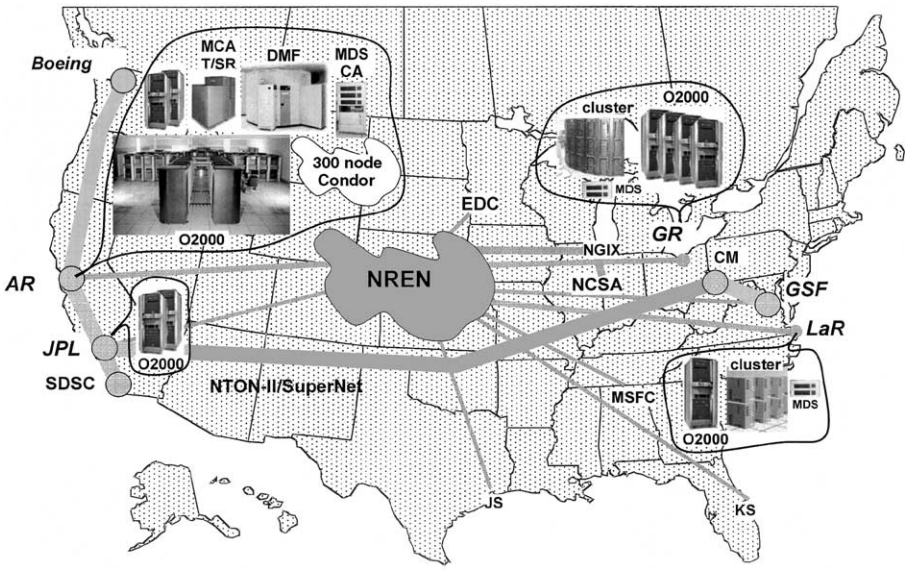

Fig. 2. The NASA IPG.

In NASA's view, such a framework is necessary for their organisation to address the problem of simulating whole systems. It is not only computing resources that are distributed but also expertise and know-how. In order to simulate a whole aircraft—wings, CFD, engines, landing gear and so on—NASA must be able to bring together not only the necessary computing resources but also establish mechanisms by which engineers and scientists at different sites can collaborate. This is the ambitious goal of their 'intra-Grid' connecting different NASA sites with different expertise and hardware resources.

In his presentations on the IPG, Bill Johnston makes the point that although technology capable of enabling such distributed computing experiments has been around for many years, such experiments tended to be 'one off' demonstration systems requiring the presence and knowledge of expert distributed system engineers. He argues that what is needed now is middleware that makes the construction of such systems *routine*—with no need for experts to build and maintain the system. The IPG is the first step along the road towards a 'production-quality' Grid. The IPG middleware is based on the Globus Toolkit that offers secure single sign-on through digital certificates (Grid Security Infrastructure, GSI), fast secure file transfer (GridFTP) and batch queueing resource management of computational jobs (Globus Resource Allocation Manager, GRAM). The Globus middleware can co-exist with CORBA middleware applications and is supplemented with Miron Livny's long established Condor software [14] and Reagan Moore's Storage Resource Broker package [15]. The Condor software allows pools of workstations to be designated and used as a source of 'idle cycles' for computational jobs submitted to the pool. The SRB software, on the other hand, goes some way towards federating access to data held in file systems, databases and archival systems. It does this by means of an XML metadata catalogue that allows the logical separation of content from its physical storage.

In the UK context, we regard the IPG as an 'existence proof' that the Grid vision can be realised. Clearly, the present implementation of Grid middleware is still rudimentary and there are many deficiencies. Nevertheless, since we are about to launch a major series of e-Science projects in the UK that will require a similar Grid infrastructure, it seems prudent to build on the experience of NASA and use the Globus, Condor and SRB Grid middleware as a starting point.

Table 1
US funded Grid projects

Project	Funding agency	URL
IPG	NASA	http://www.nas.nasa.gov/About/IPG/ipg.html
Science Grid	DOE	http://www-itg.lbl.gov/Grid/
GriPhyN Project	NSF	http://www.griphyn.org/
PPDataGrid	DOE	http://www.ppdg.net/
NVO	NSF	http://www.srl.caltech.edu/nvo/
NEESGrid	NSF	http://www.neesgrid.org/html/np.html
Distributed Terascale Facility (TeraGrid)	NSF	http://www.teragrid.org/
DISCOM (ASCI)	DOE	http://www.cs.sandia.gov/discom/
Earth Systems Grid	DOE	http://public.lanl.gov/radiant/research/grid.html
FusionGrid	DOE	http://www.fusiongrid.org/
BIRN	NIH	http://birn.ncrr.nih.gov/
iVDGL	NSF	http://www.ivdgl.org/
GridCenter	NSF	http://www.grids-center.org/
GrADS	NSF	http://nhse2.cs.rice.edu/grads/

2.3. National and international Grid projects

There are now significant investments being made in Grid research and development around the world. To set the UK programme in context we present a list of funded Grid projects in the USA (Table 1) and in Europe (Table 2). With the recent announcements of both IBM and Sun's support for the Grid concept, it is clear that the Grid is likely to become the key middleware not only for e-Science and but also for industry and commerce as the middleware matures and gains more functionality.

The majority of these Grid infrastructure and application projects will use the Globus Toolkit as the starting platform on which to provide Grid services.

In addition to these US and European Union projects, there are a number of other European national projects besides the UK initiative. These include projects in the Netherlands (VLAM [16] and DutchGrid [17]), Germany (Unicore [18]), Italy (INFN Grid [19]) and Poland (PIONIER Grid [20]). There are also Grid projects in Eire and Hungary, and funding approved for Grids in France and Switzerland. There is also an aspiration to construct an Asia-Pacific Grid. A word of warning is in order before making comparisons to compare the funding available for Grid middleware development. The announced funding for US projects usually includes a substantial amount for hardware purchases. In the UK by contrast, almost all of the £120M is for middleware development and e-Science software tools.

Table 2
European Union funded projects

Project	Funding agency	URL
DataGrid (CERN)	European Commission	http://www.datagrid.cnr.it/, http://www.cern.ch/grid/
EuroGrid (Unicore)	European Commission	http://www.eurogrid.org/
Damien (Metacomputing)	European Commission	http://www.hlrs.de/organization/pds/projects/damien/
AVO (Virtual Observatory)	European Commission	http://www.astro-opticon.org/archives.html
GRIP (Unicore/Globus)	National Center for Research Resources	http://www.unicore.org/links.htm
GridLab (Cactus Framework)	European Commission	http://www.gridlab.org/
CrossGrid (infrastructure components)	European Commission	http://www.crossgrid.org/crossgrid/crossgrid.html
Grid-Ireland		http://www.cs.tcd.ie/coghlan/, http://www.cuc.ucc.ie/
Grid for remote computing		http://sara.unile.it/grb/grb.html

3. The UK e-Science pilots

3.1. Introduction

As noted above, within the UK e-Science programme each research council will fund a number of pilot projects in their own application areas. The research councils are each taking a slightly different approach in the selection of pilot projects, which has the consequence that of the more than 20 anticipated projects, eight are already underway and the rest will follow over the next 6 months. PPARC and EPSRC have selected projects at this time. PPARC are funding two e-Science pilots and EPSRC six projects. These two research councils support very different communities. PPARC is focussed on particle physics and astronomy: both have very tightly knit communities but both also have an increasingly global scope. By contrast, the EPSRC community is much broader and including disciplines ranging from basic science to engineering.

In this section, we give an overview of the pilot projects that have been funded to set the context for the e-Science Core Programme. The two PPARC projects have been discussed in the introduction. GridPP is a collaboration of UK particle physicists to build the UK component of a world-wide Grid to support the LHC experiments. The UK project is working very closely with the EU DataGrid project and with CERN. The UK astronomers are coming together in the AstroGrid 'Virtual Observatory' project. Clearly, there will be close collaboration with the US NVO project and the European AVO project. PPARC are also working closely with the e-Science Core Programme to develop a common architectural vision and a common base of core Grid middleware.

3.2. Engineering and physical science pilot projects

The six pilots funded by EPSRC are as follows.

3.2.1. The RealityGrid: a tool for investigating condensed matter and materials

This project is led by Professor Peter Coveney and involves a University consortium comprising QMW, the Universities of Edinburgh, Loughborough, Manchester, and Oxford. The goal of this pilot project is to enable the realistic modelling of complex condensed matter systems at the molecular and mesocale levels, and to provide the setting for the discovery of new materials. Integration of high performance computing and visualisation facilities are critical to this pilot, providing a synthetic environment for modelling the problem that will be compared and integrated with experimental data. The RealityGrid involves the active collaboration of industry: AVS, SGI and Fujitsu are collaborating on the underlying computational and visualisation issues, Schlumberger and the Edward Jenner Institute for Vaccine Research will provide end-user scientific applications to evaluate and test the environment and tools produced by the project.

3.2.2. Comb-e-Chem—structure-property mapping: combinatorial chemistry and the Grid

The Comb-*e*-Chem project is concerned with the synthesis of new compounds by combinatorial methods. It is a collaboration between the Universities of Southampton and Bristol, led by Dr. Jeremy Frey. The university consortium is working together with Roche Discovery, Welwyn, Pfizer, and IBM. Combinatorial methods provide new opportunities for the generation of large amounts of original chemical knowledge. To this end, an extensive range of primary data needs to be accumulated, integrated and relationships modelled for maximum effectiveness. The project intends to develop an integrated platform that combines existing structure and property data sources within a Grid-based information and knowledge-sharing environment. The first requirement for this platform is to support new data collection, including process as well as product data, based on integration with electronic lab and e-logbook facilities. The next step is to integrate data generation on-demand via Grid-based quantum and simulation modelling to augment the experimental data. For the environment to be usable by the community at large, it will be necessary to develop interfaces that provide a unified view of resources, with transparent access to data retrieval, online modelling, and design of experiments to populate new regions of scientific interest. The service-based Grid computing infrastructure required will extend to devices in the laboratory as well as data bases and computational resources.

3.2.3. Distributed aircraft maintenance environment (DAME)

The collaborating universities on this pilot are York, Oxford, Sheffield, and Leeds, and the project is led by Professor Jim Austin. The project aims to build a Grid-based distributed diagnostics system for aircraft engines. The pilot is in collaboration with Rolls Royce and is motivated by the needs of Rolls Royce and its information system partner Data Systems and Solutions. The project will address performance issues such as large-scale data management with real-time demands. The main deliverables from the project will be a generic Distributed Diagnostics Grid application; an Aero-gas turbine Application Demonstrator for the maintenance for aircraft engines; and techniques for distributed data mining and diagnostics. Distributed diagnostics is a generic problem that is fundamental in many fields such as medical, transport and manufacturing and it is hoped that the lessons learned and tools created in this project will be suitable for application in those areas.

3.2.4. Grid enabled optimisation and design search for engineering (GEODISE)

GEODISE is a collaboration between the Universities of Southampton, Oxford and Manchester, together with BAE Systems, Rolls Royce, Fluent. The goal of this pilot is to provide Grid-based seamless access to an intelligent knowledge repository, a state-of-the-art collection of optimisation and search tools, industrial strength analysis codes, and distributed computing and data resources.

Engineering design search and optimisation is the process whereby engineering modelling and analysis are exploited to yield improved designs. In the next 2–5 years, intelligent search tools will become a vital component of all engineering design systems and will steer the user through the process of setting up, executing and post-processing design, search and optimisation activities. Such systems typically require large-scale distributed simulations to be coupled with tools to describe and modify designs using information from a knowledge base. These tools are usually physically distributed and under the control of multiple elements in the supply chain. Whilst evaluation of a single design may require the analysis of gigabytes of data, to improve the process of design can require assimilation of terabytes of distributed data. Achieving the latter goal will lead to the development of intelligent search tools. The application area of focus is that of computational fluid dynamics (CFD) which has clear relevance to the industrial partners.

3.2.5. Discovery net: an e-Science testbed for high throughput informatics

The Discovery Net pilot has a completely different aim than those of the other projects, in that it is focussed on high throughput. It aims to design, develop and implement an advanced infrastructure to support real-time processing, interpretation, integration, visualisation and mining of massive amounts of time critical data generated by high throughput devices. The project will cover new technology devices and technology including biochips in biology, high throughput screening technology in biochemistry and combinatorial chemistry, high throughput sensors in energy and environmental science, remote sensing and geology. A number of application studies are included in the pilot—analysis of Protein Folding Chips and SNP Chips using LFII technology, protein-based fluorescent micro array data, air sensing data, renewable energy data, and geohazard prediction data. The collaboration on this pilot is between groups at Imperial College London, by Dr. Yike Guo, and industrial partners Infosense Ltd., Deltadot Ltd., and Rvco Inc.

3.2.6. myGrid: directly supporting the e-Scientist

This pilot has one of the larger consortiums comprising the Universities of Manchester, Southampton, Nottingham, Newcastle, and Sheffield together with the European Bioinformatics Insitute. The goal of myGrid is to design, develop and demonstrate higher level functionalities over an existing Grid infrastructure that support scientists in making use of complex distributed resources. An e-Scientist's workbench will be developed, which, not unlike that of Comb-*e*-Chem, aims to support the scientific process of experimental investigation, evidence accumulation and result assimilation, the scientist's use of the community's information, and scientific collaboration, allowing dynamic groupings to tackle emergent research problems.

A novel feature of the proposed workbench is provision for personalisation facilities relating to resource selection, data management and process enactment.

The myGrid design and development activity will be driven by applications in bioinformatics. myGrid will develop two application environments, one that supports the analysis of functional genomic data, and a second that supports the annotation of a pattern database. Both of these tasks require explicit representation and enactment of scientific processes, and have challenging performance requirements. The industrial collaborators on this project are GSK, AstraZeneca, IBM and SUN.

3.3. Particle physics and astronomy pilot projects

PPARC does not support so broad community as ESPRC (http://www.epsrc.ac.uk) and hence there are only two pilot projects—one for each of their communities. The two pilots funded are as follows.

3.3.1. GridPP

GridPP is a collaboration of Particle Physicists and Computer Scientists from the UK and CERN. The goal of GridPP is to build a UK Grid and deliver Grid middleware and hardware infrastructure to enable testing of a prototype of the Grid for the LHC project at CERN, as described in the introduction to this paper. The GridPP project is designed to integrate with the existing Particle Physics programme within the UK, thus enabling early deployment and full testing of Grid technology. The project is using the Globus toolkit and many of the GridPP centres in the UK are also UK e-Science Grid centres hence allowing many ways of leveraging effort and technology.

3.3.2. AstroGrid

The AstroGrid project has been described early in terms of the requirements within this community. The project is a collaboration between astronomers and computer scientists at the universities of Edinburgh, Leicester, Cambridge, Queens Belfast, UCL and Manchester, together with RAL. The goal of AstroGrid is to build a Grid infrastructure that will allow a ‘Virtual Observatory’, unifying the interfaces to astronomy databases and providing remote access as well as assimilation of data. The Virtual Observatory is a truly global problem and AstroGrid will be the UK contribution to the global Virtual Observatory.

3.4. Concluding remark

These PPARC and EPSRC pilots cover a range of scientific disciplines and Grid technology areas. One conspicuous requirement throughout the applications is the need for support for data management and federated database access. The present Grid middleware offers only very limited support in this area.

4. The UK e-Science Core Programme

4.1. Structure of the Core Programme

As we have explained, the goal of the e-science Core Programme is to identify the generic middleware requirements arising from the e-Science pilot projects. In collaboration with scientists, computer scientists and industry, the Director has a mandate to develop a framework that will promote the emergence of robust, industrial strength Grid middleware that will not only underpin individual application areas but also be of relevance to industry and commerce.

The Core Programme has been structured around six key elements:

1. Implementation of a National e-Science Grid testbed based on a network of e-Science Centres.
2. Promotion of generic Grid middleware development.
3. Interdisciplinary Research Collaboration (IRC) Grid projects.
4. Establishment of a support structure for e-Science pilot projects.
5. Support for involvement in international activities.
6. Support for e-Science networking requirements.

We briefly discuss each of these activities below.

4.2. The UK e-Science Grid and the e-Science centres

Nine e-Science Centres have been established at the locations shown on the map of the UK in Fig. 3.

A National e-Science Centre has been established in Edinburgh, managed jointly by Glasgow and Edinburgh Universities. Eight other regional centres have been established—in Belfast, Cardiff, Manchester, Newcastle, Oxford, Cambridge, London (Imperial

College) and Southampton—giving coverage of most of the UK. Manchester also currently operates the UK's national Supercomputer service. The Centres have three key roles:

(i) to allocate substantial computing and data resources and run a standard set of Grid middleware to form the basis for the construction of the UK e-Science Grid;
(ii) to generate a portfolio of collaborative industrial Grid middleware and tools projects;
(iii) to disseminate information and experience of the Grid within their local region.

The centres have a pre-allocated budget for industrial collaborative Grid projects of £1M (£3M for the National Centre) requiring matching funds in cash or kind from industry.

Fig. 3 also shows the Rutherford and Daresbury Laboratory sites of CLRC. These national laboratories are key sites of the UK e-Science Grid. The Hinxton site near Cambridge is also shown in Fig. 3: Hinxton hosts the European Bioinformatics Institute (EBI), the Sanger Centre and an MRC Institute. This constitutes one of the major centres of genomic data in the world. It is therefore important that this site is linked to the e-Science Grid with sufficient bandwidth to support a number of e-Science bioinformatics projects.

The National Centre in Edinburgh has also been funded to establish an 'e-Science Institute'. This institute will organise a series of multidisciplinary research seminars covering a wide range of topics, with scientists and experts from all over the world. The seminars can have many different formats, from a one-day workshop to a month-long 'e-Science Festival', planned for summer 2002. Their brief is to make the Institute an internationally known centre for stimulating intellectual debate on all aspects of e-Science.

In addition, AccessGrid nodes have been established in each Centre to aid collaboration both within and outside the UK (Fig. 4). The AccessGrid system was developed at Argonne National Laboratory in the USA and makes use of MBONE and Multicast technologies to provide a more natural video-conferencing experience between multiple sites that allows direct integration of Grid simulations and visualisation [21]. This system allows easy interaction between the Centres and will be used to experiment with innovative ways of working and teaching.

4.3. Promotion of Grid middleware development

In order for the UK programme to be successful, we must provide projects with a common starting point and educate a core team of people with the necessary

Fig. 3. The UK e-Science Grid.

Fig. 4. AccessGrid Session at Manchester e-Science Centre.

experience in building and running a Grid. The UK e-Science Grid linking the nine centres and the two laboratories of the CLRC will be used as a testbed in two ways. As discussed earlier, the initial Grid middleware selected is the same as that used by NASA in their IPG. The IPG may be termed an 'intra-Grid'—since the NASA laboratories connected are all part of one organisation. In the case of the UK e-Science Grid, the Grid connects different universities with different IT policies, firewalls and so on. This is a good test of the basic Globus infrastructure and the digital certificate based security system. The centres are contributing a heterogeneous collection of resources to this Grid, including supercomputer and cluster computing systems as well as diverse data storage systems. This Grid can be used both as a test platform for new Grid middleware and as a resource available for centre industrial projects.

The Core Programme is in discussions with major IT companies such as IBM, Sun, Oracle and Microsoft, as well as with the Globus, Condor and SRB teams and others, concerning the future development of Grid middleware. In this respect, it is encouraging that both IBM and Sun have given strong endorsements to working with the Globus team to take forward the production of improved and robust Grid middleware. The software that will emerge will offer considerably more functionality than the present Grid middleware and will also be produced to industrial quality. In a recent interview, Irving Wladawsky-Berger made the commitment that 'all of our systems will be enabled to work with the Grid, and all of our middleware will integrate with the software' [22]. He also saw that there would be 'early adopter' industries such as petro-chemical, pharmaceutical and engineering design with the Grid making significant impact on more general commerce and industry by 2003 or 2004. This is certainly what we are seeing in the UK e-Science Grid projects described above.

In the Core Programme, a total of £11M has been allocated for collaborative industrial Grid middleware projects via the individual e-Science Centres. An additional £5M (plus a matching industrial contribution) is available through an 'Open Call' with no deadlines. A major task for the Core Programme is the capture of requirements for the Grid infrastructure from each of the e-Science pilot projects. These include computational, data storage and networking requirements as well as the desired Grid middleware functionality. In order that the projects do not dissipate their energies by fruitless re-explorations of common ground, the Core Programme has commissioned a number of reports on the present state of Grid middleware. Reports on Globus, SRB/Databases [23], and .NET are currently available (www.nesc.ac.uk), and evaluations of Sun GridEngine [24] and Avaki's product offerings [25] are in progress. In addition, a Grid DataBase Task Force (DBTF), led by Norman Paton from Manchester, has been set up to

examine the question of Grid middleware interfaces to Relational DataBase Management Systems and the federation of different data sources [26]. Their preliminary ideas have been discussed with the Globus team and with IBM, Oracle and Microsoft. The DBTF has both a short-term remit—to look at developing an interface with some minimal useful functionality as soon as possible—and a longer-term remit—to look at research issues beyond flat files and relational data.

A Grid Architecture Task Force has also been created, led by Malcolm Atkinson, Director of the National e-Science Centre in Edinburgh, to look at overall architectural directions for the Grid. They are tasked with producing a 'UK Grid Road Map' for Grid middleware development. Again, the ATF is tasked with identifying some specific short-term goals as well as identifying longer-term research issues. Initial ideas from the DBTF point towards the implementation of a database interface in terms of a 'Grid Services' model along the lines of Web Services. This leads to the idea of the Grid middleware designed as a 'Service Oriented Architecture' with Grid services consumed by higher level applications. Reports from both task forces will be submitted to the Global Grid Forum as white papers for discussion in the near future. We expect to see the eventual emergence of agreed 'Open Grid Services' protocols [12], with specified interfaces to operating systems and RDBMS below this layer and offering a set of Grid services to applications above. It is important that at least a subset of standard protocols that go beyond the present Globus model are agreed as quickly as possible in order to ensure that the Grid middleware development in the application projects can proceed effectively. We intend to collaborate with the Globus team and assist in taking forward the open source implementation of these standards.

4.4. IRC Grid projects

The EPSRC in the UK has funded 3, 6-year, computer science (CS) oriented, IRCs. These are major projects that fund key CS research groups from a number of universities to undertake long-term research in three important areas. The Equator project, led by Tom Rodden from Nottingham, is concerned with technological innovation in physical and digital life [27]. The Advanced Knowledge Technologies (AKT) project led by Nigel Shadbolt from Southampton is concerned with the management of the knowledge life cycle [28]. Lastly, the DIRC project, led by Cliff Jones from Newcastle and Ian Sommerville from Lancaster, is concerned with the dependability of computer-based systems. A fourth IRC, jointly funded by EPSRC and the MRC, is the MIAS project led by Mike Brady from Oxford. This application focussed IRC is concerned with translating data from medical images and signals into clinical information of use to the medical profession.

These IRCs were selected after an open competitive bidding process and represent a unique pool of expertise in these three key software technologies and in the important multidisciplinary application area of medical informatics. For this reason, the Core Programme is funding projects with each of these IRCs to enable them to consider the implications of Grid technology on their research directions. In addition, we are funding two collaborative projects combining the software technologies of the Equator and AKT IRCs with the MIAS application project. In effect these projects constitute a sort of 'Grand Challenge' pilot project in 'e-Healthcare'.

As a concrete example of the potential of Grid technologies for e-Healthcare, consider the problem that Mike Brady and his group at Oxford are investigating in the MIAS IRC. They are performing sophisticated image processing techniques on mammogram and ultrasound scans for breast cancer tumours. In order to locate the position of possible tumours with maximum accuracy, it is necessary for them to construct a Finite Element Model of the breast using an accurate representation of the properties of breast tissue. The informatics challenge for the Grid middleware is to deliver accurate information on the position of any tumour within the breast to the surgeon in or near the operating theatre. Clearly there are many issues concerning security and privacy of data to be resolved, but the successful resolution of such issues must be a high priority not only for healthcare in the UK but also world-wide. This also illustrates another aspect of the Grid. Hospitals do not particularly wish to be in the business of running major computing centres to do the required modelling and analysis required by modern medical technologies. They would much rather be in the position of being able

to purchase the required resource as a 'Grid service' on a pay-as-you-go basis. This illustrates the way in which Grid Services middleware can encourage the emergence of new 'Application Service Provider' businesses.

4.5. Support for e-Science projects

In order to provide support for e-Science application projects, a Grid Support Centre has been established, jointly funded by the Core Programme

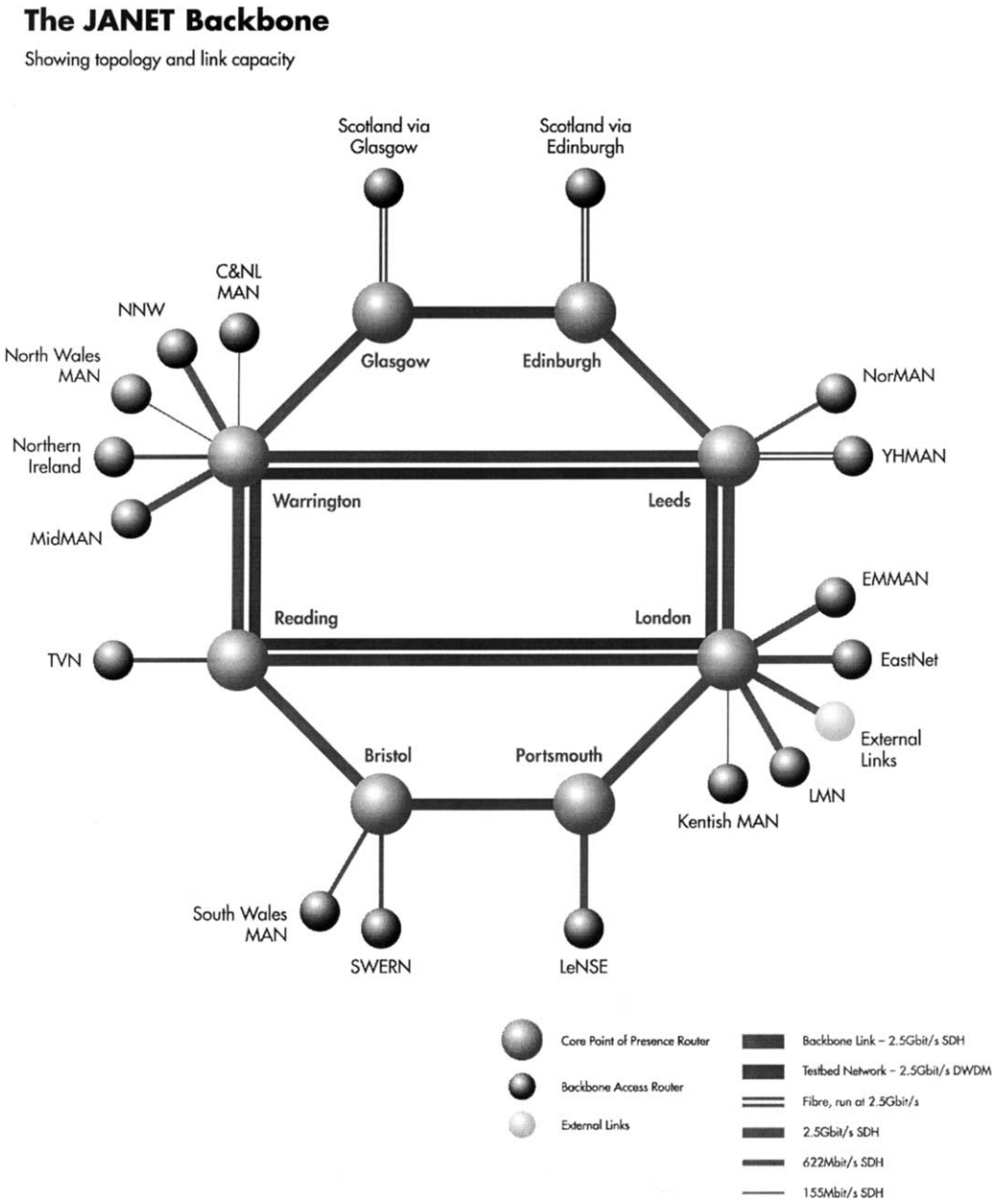

Fig. 5. SuperJANET4 projected network capacity.

and PPARC (www.grid-support.ac.uk). The Centre will provide support for the UK 'Grid Starter Kit' which initially consists of the Globus Toolkit, Condor and the Storage Resource Broker middleware. The support team will be available to answer questions and resolve problems for Grid application developers across the UK on a 9 to 5 basis. They will also test and evaluate new middleware software and control future upgrades to the Grid Starter Kit. A further role of the team is to help educate the systems support staff at the UK e-Science Centres.

4.6. International collaboration

It is important to ensure that the UK e-Science community are actively communicating and collaborating with the international community. It is therefore desirable to encourage the development of an informed UK community on Grid technologies and provide funding for them to play an active role in the development of internationally agreed Grid protocols at the Global Grid Forum. We have therefore funded a 'GridNet' network project which has a substantial travel budget for attendance of UK experts at relevant standards bodies—the Global Grid Forum, the IETF and W3C, for example.

The UK programme is also concerned to create meaningful links to international efforts represented by projects such as the EU DataGrid and the US iVDGL projects. We are therefore funding Grid fellowships for young computer scientists from the UK to participate in these projects. The National e-Science Centre is also tasked with establishing working agreements with major international centres such as Argonne National Laboratory, San Diego Supercomputing Center and NCSA in the USA. We are also looking to establish other international links and joint programmes.

4.7. Networking

The UK e-Science application projects will rely on the UK universities network SuperJANET4 for delivering the necessary bandwidth. The backbone bandwidth of SuperJANET4 is now 2 Gbps and there is funding in place to upgrade this to 10 Gbps by mid 2002 (Fig. 5).

In order to gather network requirements from the UK e-Science project—in terms of acceptable latencies and necessary end-to-end bandwidth—the Grid Network Team (GNT) has been established. Their short-term remit is to identify bottlenecks and potential bottlenecks as well as look to longer term Quality of Service issues. A £500K project with UKERNA and CISCO will be looking at such traffic engineering issues and another project will be looking at the question of bandwidth scheduling with the EU DataGrid project. Both of these network R&D projects are jointly funded by the Core Programme and PPARC. A potential problem for the UK programme is the 'balkanisation' of the UK network into a Super JANET4 backbone run by UKERNA and Metropolitan Area Networks run by university consortia. It may be difficult to provide end-to-end bandwidth QOS guarantees with such a structure.

4.8. Demonstrator projects

The Core Programme has also funded a number of 'Grid Demonstrator' projects. The idea of the demonstrators is to have short-term projects that can use the present technology to illustrate the potential of the Grid in different areas. We have tried to select demonstrators across a range of applications. Examples include a dynamic brain atlas, a medical imaging project using VR, a robotic telescope project, automated chemical data capture and climate prediction (www.research-councils.ac.uk/escience/).

5. Conclusions

There are many challenges to be overcome before we can realise the vision of e-Science and the Grid described above. These are not only technical issues such as scalability, dependability, interoperability, fault tolerance, resource management, performance and security but also more people-centric relating to collaboration and the sharing of resources and data.

As an example of a technical issue, we believe that realistic performance estimation will be key to the realisation of the vision of the Grid as a global marketplace for resources. The NSF GrADS [29] project

envisages a "performance contract" framework as the basis for the dynamic negotiation process between resource providers and consumers. For realistic performance prediction, we need reliable data collection facilities that are able to monitor the available capacity of distributed resources and forecast future performance levels using statistical models. Performance forecasting and resource management are complex tasks within the Grid environment. As an example of a people-centric issue, consider the motivation for scientists to share their annotated scientific data—simulation or experimental—in federated data repositories. For science to make best use of its limited funds, sharing of such expensively gathered scientific data is clearly of paramount importance. Yet the motivation for any individual scientist is not so clear. Perhaps funding agencies need to add some incentive to encourage such a community-minded approach to sharing scientific data.

There are many other challenges. Two important areas not yet addressed by the UK programme are security and scientific data curation. For the Grid vision to succeed in the industrial and commercial sector, the middleware must be secure against attack. In this respect, the present interaction of Globus with firewalls leaves much to be desired. Another issue concerns the long-term curation and preservation of the scientific data along with its associated metadata annotations. A 3-year programme is clearly unable to provide a solution to such a long-term problem, but it is one that must be addressed in the near future. Our present e-Science funding has enabled the UK to make a start on these problems and for the UK to play a full part in the development of the international Grid community. If the Grid middleware succeeds in making the transition from e-Science to commerce and industry in a few years, there will clearly be many new application areas to explore.

Acknowledgements

The authors thank Juri Papay for valuable assistance in preparing this paper and Jim Fleming and Ray Browne for their help in constructing the Core Programme. We also thank Paul Messina for encouragement and insight and Ian Foster and Carl Kesselman for their constructive engagement with the UK programme.

References

[1] John Taylor e-Science definition: http://www.e-science.clrc.ac.uk/.
[2] I. Foster, C. Kesselman (Eds.), The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, Los Altos, CA, 1999.
[3] B. Segal, Grid Computing: The European Data Project, IEEE Nuclear Science Symposium and Medical Imaging Conference, Lyon, October 2000, pp. 15–20.
[4] The DataTag Project: http://www.datatag.org.
[5] GridPP: http://www.gridpp.ac.uk/.
[6] Griphyn: http://www.griphyn.org/.
[7] The Particle Physics DataGrid: http://www.ppdg.net/.
[8] International Virtual Data Grid Laboratory: http://www.ivdgl.org/.
[9] National Virtual Observatory: http://www.srl.caltech.edu/nvo/.
[10] Opticon Astrophysical Virtual Observatory: http://www.astro-opticon.org/archives.html.
[11] I. Foster, C. Kesselman, Globus: a metacomputing infrastructure toolkit, Int. J. Supercomput. Appl. 11 (2) (1997) 115–128. http://www.globus.org.
[12] I. Foster, C. Kesselman, J. Nick, S. Tuecke, The Physiology of the Grid: Open Grid Services Architecture for Distributed Systems Integration, to be presented at GGF4, February 2002.
[13] NASA Information Power Grid: http://www.nas.nasa.gov/About/IPG/ipg.html.
[14] M. Litzkow, M. Livny, M. Mutka, 'Condor—a hunter of idle workstations', Proceedings of the Eighth International Conference of Distributed Computing Systems, June 1988, pp. 104–111. http://www.cs.wisc.edu/condor.
[15] Storage Resource Broker: http://www.npaci.edu/DICE/SRB.
[16] Virtual Laboratory Abstract-Machine: http://www.dutchgrid.nl/VLAM-G/AM/doclive-vlam.html.
[17] DutchGrid: http://www.dutchgrid.nl.
[18] The Unicore Project: http://www.unicore.de/.
[19] Instituto Nationale di Fisica Nucleare: http://www.infn.it/.
[20] PIONIER: Polish Optical Internet Advanced Applications, Services and Technologies for Information Society: http://www.man.poznan.pl/pol34/pionier/english/.
[21] AccessGrid: http://www-fp.mcs.anl.gov/fl/accessgrid/.
[22] IBM Grid Position News: http://www.ibm.com/news/us/2001/08/15.html.
[23] P. Watson, Databases and the Grid, Technical Report CS-TR-755, University of Newcastle, 2001.
[24] Sun Gridware: http://www.sun.com/gridware/.
[25] Avaki: http://www.avaki.com/products/.
[26] N.W. Paton, M.P. Atkinson, V. Dialani, D. Pearson, T. Storey, P. Watson, Database Acccess and Integration Services on the Grid, UK DBTF Working Paper, January 2002.
[27] Digital Care: http://www.equator.ac.uk/projects/DigitalCare/.
[28] AKT IRC: http://www.ecs.soton.ac.uk/~nrs/projects.html.
[29] The GrADS Project: http://nhse2.cs.rice.edu/grads/.

Tony Hey is a Professor of Computation at the University of Southampton and has been Head of the Department of Electronics and Computer Science and Dean of Engineering and Applied Science at Southampton. From 31 March 2001, he has been seconded to the EPSRC and DTI as Director of the UK's Core e-Science Programme. He is a Fellow of the Royal Academy of Engineering, the British Computer Society, the Institution of Electrical Engineers and the Institution of Electrical and Electronic Engineers. Professor Hey is European editor of the journal 'Concurrency and Computation: Practice and Experience' and is on the organising committee of many international conferences.

Professor Hey has worked in the field of parallel and distributed computing since the early 1980s. He was instrumental in the development of the MPI message-passing standard and in the Genesis Distributed Memory Parallel Benchmark suite. In 1991, he founded the Southampton Parallel Applications Centre in 1991 that has played a leading technology transfer role in Europe and the UK in collaborative industrial projects. His personal research interests are concerned with performance engineering for Grid applications, but he also retains an interest in experimental explorations of quantum computing and quantum information theory. As the Director of the UK e-Science Programme, Tony Hey is currently excited by the vision of the increasingly global scientific collaborations being enabled by the development of the next generation 'Grid' middleware. The successful development of the Grid will have profound implications for industry and he is much involved with industry in the move towards OpenSource/OpenStandard Grid software.

Tony Hey is also the author of two popular science books: 'The Quantum Universe' and 'Einstein's Mirror'. Most recently he edited the 'Feynman Lectures on Computation' for publication, and a companion volume entitled 'Feynman and Computation'.

Anne E. Trefethen acts as the Deputy Director of the Core e-Science Programme. She is on secondment from NAG Ltd., where she is the Vice President for Research and Development. Anne joined NAG in 1997 and in her role there she leads technical development in NAG's numerical and statistical products.

Before joining NAG, Anne was the Associate Director for Scientific Computational Support at the Cornell Theory Center. For most of the 1990s, the Cornell Theory Center was one of four NSF funded national supercomputer centres in the USA. During her first 3 years at the centre, she was a research scientist in the performance and algorithms group, working largely on parallel linear algebra and parallel scientific applications, also teaching on the centre's courses on parallel techniques and numerical algorithms. As an Associate Director she was responsible for the division who supported computational scientists across the country in terms of software, application and algorithmic consultancy, education and training, strategic applications support, and outreach programs. At that time, as well as many technical activities, she was very much involved in the design and development of Web-based educational courses and led the research activity, MultiMatlab, to create a distributed Matlab environment for large-scale problem solving.

From 1988 to 1992, Anne was a Research Scientist in the Mathematical and Computational Sciences Group at Thinking Machines Corporation. She was the project leader for the first independent release of the Connection Machine Scientific Software Library (CMSSL).