The UK e-Science Core Programme and the Grid

Tony Hey¹ and Anne E. Trefethen²

¹Director UK e-Science Core Programme EPSRC, Polaris House, North Star Avenue
Swindon SN2 1ET, UK,
Tony.Hey@epsrc.ac.uk

² Deputy Director UK e-Science Core Programme EPSRC, Polaris House, North Star
Avenue Swindon SN2 1ET, UK,
Anne.Trefethen@epsrc.ac.uk

Abstract. e-Science encompasses computational science and adds the new dimension of the routine sharing of distributed computational, data, instruments and specialist resources. This paper describes the £120M UK ‘e-Science’ initiative, focusing on the ‘Core Programme’. The majority of the funding is for large-scale e-Science pilot projects in many areas of science and engineering. The pilot e-Science projects that have been announced to date are 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. The requirements of these projects and the recent directions in developments of open source middleware have provided direction for the roadmap of development within the Core Programme. The middleware infrastructure needed to support e-Science projects must permit routine the sharing of any remote resources as well as supporting effective collaboration between groups of scientists. Such an infrastructure is commonly referred to as the Grid. The focus of the paper is on the Core Programme that has been created in order 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.

1. Introduction

e-Science in many ways seems a natural extension to computational science. It is of course much more than this. Dr John Taylor, Director General of Research Councils in the UK Office of Science and Technology (OST), introduced the term e-Science. 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 this 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]. The use of the term *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. We will illustrate toward the end of this paper how these two models are likely to intersect.

In the next section we outline the general structure of the UK e-Science programme. Following on from that we describe the UK funded e-Science Pilot projects and discuss the likely requirements from these projects. In Section 3 we describe the UK e-Science Core Programme in some detail.. We conclude with some remarks about the evolution of grid middleware architecture and the roadmap for contributions in the UK.

2. The UK e-Science Programme

Under the UK Government’s Spending Review in 2000, the Office of Science and Technology (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 Figure 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.

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

3. The UK e-Science Pilot Projects

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 whilst some research councils have committed all their e-Science funds, others are only partially through that process. We describe below those projects that have been funded so far.

3.1 Engineering and Physical Sciences Pilot Projects

The EPSRC did not mandate the areas that should be covered in the pilot projects but required that the proposal should be consortia based proposals including consortia of researchers spanning application scientists and engineers, computational scientists and computer scientists and those from industry and commerce. They also required the proposals to be focused on real world challenges and be able to demonstrate, within three years, working systems of Grid technologies. The six pilots funded by EPSRC are:

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 mesoscale levels, and to provide the setting for the discovery of new materials. Integration of high performance computing and visualisation facilities is 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.

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 databases and computational resources.

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.

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.

DiscoveryNet: An e-Science Testbed for High Throughput Informatics

The DiscoveryNet pilot has a slightly different aim than those of the other projects, in that it is focused on high throughput. It aims to design, develop and implement an advanced infrastructure to support real-time processing, interpretation, integration, visualization 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, lead by Dr Yike Guo, and industrial partners Infosense Ltd, Deltadot Ltd, and Rvco Inc.

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 Institute. 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 in this pilot project. The workbench, 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 another 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.

3.2 Particle-Physics and Astronomy Pilot Projects

The Particle Physicists and Astronomers are communities that perhaps have the most easily recognizable requirement for e-Science. 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 the Standard Model of the weak and electromagnetic interactions. 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. The physicists need to put in place an LHC Grid infrastructure that will

permit the transport and data mining of such distributed data sets, in which the particle physicists are working to build a Grid that will support these needs. Similarly the astronomers have the goal of providing 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. The two projects being funded by PPARC are:

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 Large Hadron Collider (LHC) project at CERN, as described above. 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. GridPP is very closely integrated with the EU DataGrid project [3] and, EU DataTag [4], and related to the USA projects NSF GriPhyN [5], DOE PPDataGrid [6], and NSF iVDGL [7]

AstroGrid

The AstroGrid project has been described early in terms of the requirements within this community. The project is 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 collaborating closely with the US NVO [8] project and the European AVO [9] project.

3.3 Biotechnology and Biological Sciences Research Council

Biological science is now one of the major areas of focus of scientific research worldwide. This effort is largely interdisciplinary teams working on specific problems of significant biological interest that underpin UK industry: pharmaceuticals, healthcare, biotechnology and agriculture. BBSRC has made a significant investment in bioinformatics research in recent years to establish a strong bioinformatics research community and have built on this investment with the e-Science funding focusing on post genomics, structural biology/structural genomics and generic technology development. The two pilots funded are described below:

A GRID Database for Biomolecular Simulations

This project is led by Prof Mark Sansom of Oxford University with collaborators at Southampton University, Birkbeck College, York University, Nottingham University and the University of Birmingham. The aims of the project include establishing a consortium for biomolecular simulations based on distributed, shared data structures; to establish a biomolecular simulation database; to establish metadata and tools to provide interrogation and datamining of federated datasets. This GRID approach with associated metadata facilities will facilitate access to simulation results by non-experts eg in the structural biology and genomics communities.

A distributed pipeline for structure-based proteome annotation using GRID technology

This project aims to use GRID technology to provide a structure-based annotation of the proteins in the major genomes (proteomes) and will develop the prototype for a national/distributed proteome annotation GRID. The annotation will primarily employ sophisticated homology and fold recognition methods to assign protein structures to the proteomes and generate 3D models. The consortium plan to establish local databases with structural and function annotations; to disseminate to the biological community the proteome annotation via a single web-based distributed annotation system (DAS); to share computing power transparently between sites using GLOBUS; and to use the developed system for comparison of alternative approaches for annotation and thereby identify methodological improvements. The consortium comprises University College London, the European Bioinformatics Institute, and is led by Prof Mike Sternberg of Imperial College.

3.4 Medical Research Council Pilot Projects

The Medical Research Council has recently funded the projects described below and anticipates funding further pilots in the near future. MRC like other councils required multidisciplinary consortia but restricted the problem areas to cancer or brain sciences for this call. The four proposals funded to date are described below.

Co-ordination, integration and distribution of sequence and structural family data

This project is led by Dr A Bateman of the European Bioinformatics Institute, together with the Sanger Centre, the MRC Centre for Protein Engineering and University College London. The goal of the project is to bridge the gap between sequence and structural information to improve access and utility of this data for biologists. Using Grid technologies the consortium will co-ordinate and integrate information from five existing databases (SCOP, CATH, EMSD, PFAM, and Interpro) allowing researchers from the UK and around the world to increase the pace of their basic research. They plan to facilitate genome annotation by extending the functional annotations associated with these integrated resources.

Bioinformatics, Molecular Evolution and Molecular Population Genetics.

Professor J Hein at Oxford University plans to employ Statistical Alignment to improve comparative versions of problems such as gene finding, protein and RNA secondary prediction. His project also includes the application of coalescent theory to variation in the human genome. This would focus on finding computational accelerations to present methods and would be applied to data generated by the Sanger Centre. He will also study models of protein structure evolution and pathogen sequence evolution. Sequences of pathogens are being determined on a huge scale world wide and issues such as intrapatient evolution, origin of pathogens, clonality of pathogens and much more can be addressed by such data using the proper methods.

Artificial Neural Networks in Individual Cancer Management

The objective of this project is to develop and train a system of artificial neuronal networks (ANNs) providing information on prognosis, staging and optimal (multidisciplinary) management in patients with (a) breast, (b) upper GI and (c) colorectal cancers. In the first part of the study, the neuronal net system will be developed based on an existing system used by the particle physicists in Manchester by working in cooperation with academic clinical oncologists and a research nurse in Dundee using three national patient databases that are available for the project. The patient data used for development and training of the ANNs will be obtained by random retrieval from the three databases. In the second part of the project, validation of the ANNs for Cancer Management will be evaluated (in terms of staging, predicted optimal treatment and prognosis) with the observed management and outcome of the patients. The project is a collaboration between the universities of Dundee and Manchester and is led by A Cuschieri of the former.

Biology of Ageing E-Science Integration and Simulation System (BASIS)

Prof T Kirkwood of the University of Newcastle is employing e-Science techniques to the issue of ageing, which is a complex biological process involving interactions between a number of biochemical mechanisms whose effects occur at the molecular, cell, tissue and whole organism levels. The project will look to providing integration of data and hypotheses from diverse sources and to develop an innovative biomathematical and bioinformatics system.

3.5 Natural Environment Research Council

NERC has a ten year vision for the development of coupled Earth System models that improve both the understanding of the complex interactions within and between the biosphere, the geosphere, the atmosphere and the hydrosphere and the ability to confidently predict and resolve environmental problems. e-Science and Grid technologies clearly play an important role in advancements in this area. NERC has funded 3 pilots projects to date and plan to fund a similar number in a few months time.

Climateprediction.com: Distributed computing for global climate research

The *climateprediction.com* project aims to perform a probabilistic forecast of climate change, by exploring parameter-space in a coupled AOGCM (the UM) as a Monte Carlo simulation. The model will be distributed to businesses and individuals to run on their PCs, using up idle cycles. The runs will perform a pre-packaged 100-year simulation 1950-2050. Individual forecasts will be weighted by the quality of fit to observations (1950-2000), yielding a probabilistic forecast independent of expert judgment. Initial runs with a "slab" (thermodynamic ocean) version will help prioritise parameter perturbations. The project will have the challenging problem of tera-scale, distributed data archival and analysis. The project lead is Myles Allen at Oxford University who has collaborators at RAL and Open University.

GRID for Environmental Systems Diagnostics and Visualisation

The aim of this project is to develop a grid for ocean diagnostics with high-resolution marine data from the NERC OCCAM model and data from the Met Office Forecasting Ocean-Atmosphere Model. A prototype of the system can be seen at www.nerc-essc.ac.uk/las. Integrated in this environment will be sophisticated remote visualisation tools to allow the visualisation of large amounts of environmental data created in oceanic and atmospheric science. The project includes environmental research through the exploration of the thermohaline circulation and the calculation of water transformation budgets from the models and data assimilation techniques; and new visualisation techniques for unstructured grid fluid models.

The project is being led by Keith Haines of the University of Reading and includes collaborations with the Oceanographic centre at Southampton University, Imperial College, Manchester University, CLRC, the Meteorological Office, SGI, AVS and BMT Ltd.

Environment from the molecular level: an e-Science project for modelling the atomistic processes involved in environmental issues

The consortium in this project will develop are focusing on the processes at an atomistic level for modeling environmental problems. The e-Science challenge is to scale up the length and times scales of the simulation techniques from the molecular level, through the mesoscopic scale and towards the length and times scales of human experiences. The project will develop a collaboratory toolkit for archiving and mining of simulation output and visualisation of critical events together with metadata creation for the simulation of processes at the molecular level. The driving environmental applications are transport of pollutants and containment of radioactivity. This is a large consortium involving individuals from University College London, Bath University, Reading University, the Royal Institution and CLRC Daresbury.

3.6 Concluding Remarks

EPSRC and PPARC were the first to fund pilot projects and have almost completed the process of collecting requirements and beginning to provide direction for development of middleware. The other projects have not yet had time to complete this process. It is clear, however, that an overwhelming theme throughout the pilots is the need to access, manage and mine large amounts of disparate datasets. The datasets are, for some applications such as GridPP, largely in flat files, whilst for others, such as the bioinformatics, are in a variety of databases. The present generation of Grid middleware does not provide the tools required for these data-centric applications. This has been acknowledged and in the later section on directions for the Grid we will provide some suggestions of the UK can play a leading role in addressing this issue.

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, we wish 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 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 Figure 2.

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 currently operates the UK's national Supercomputer service. The Centres have three key roles:

1. 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 testbed;
2. to generate a portfolio of collaborative industrial Grid middleware and tools projects;
3. 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. 2. The UK e-Science Grid

Figure 2 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 Figure 2: 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 the e-Science bioinformatics projects.

The National Centre in Edinburgh has also been funded to establish an 'e-Science Institute'. This Institute is organizing a series of multidisciplinary research seminars covering a wide range of topics, with scientists and experts from all over the world. The seminars 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 (Figure 3). 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 [10]. This system allows easy interaction between the Centres and will be used to experiment with innovative ways of working and teaching.

Fig. 3. An AccessGrid Session at the NorthWest e-Science Centre

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 experience in building and running a Grid. The starting point of middleware has been chosen as Globus [11], Condor [12] and SRB [13]. This choice has been based upon the NASA Information Power Grid (IPG) [14] which is based on this set of

middleware and has therefore proven them to some extent. A difference between the IPG and the UK e-Science Grid is that the former consists of a homogeneous set of compute resources whereas 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. Hence we will be stretching the capabilities of the starting middleware. An Engineering Task Force (ETF) has been created in order to ensure that the centres are able to overcome any obstructions this might create. The UK e-Science Grid will certainly be a good test of the basic Globus infrastructure and the digital certificate based security system.

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 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' [15]. 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. A recent paper [16] co-authored by the Globus team and individuals from IBM sets the direction for this new evolution of grid middleware. The recent paradigm is referred to as the 'Open Grid Software Architecture' (OGSA) and describes the drawing together of two threads of recent development, namely webservices and grid technologies.

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, which are currently available at the National Centre website (www.nesc.ac.uk). 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. 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. The task

force has produced a paper on the initial interface definitions for discussion at the next Global Grid Forum (GGF4) [17].

In addition, a Grid Architecture Task Force has been set up, 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 and this fits very well with OGSA, leading to a Grid middleware designed as a 'Service Oriented Architecture' with Grid services consumed by higher level applications. We expect to see the eventual emergence of agreed 'Open Grid Services' protocols, 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 Interdisciplinary Research Collaboration Grid Projects

The EPSRC in the UK has funded three, six-year, computer science (CS) oriented, Interdisciplinary Research Collaborations (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 [19]. The Advanced Knowledge Technologies project (AKT) led by Nigel Shadbolt from Southampton is concerned with the management of the knowledge life cycle [20]. 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 and PPARC (www.grid-support.ac.uk). The Centre provides support for the UK 'Grid Starter Kit' which initially consists of the Globus Toolkit, Condor and the Storage Resource Broker middleware. The support team is available to answer questions and resolve problems for Grid application developers across the UK on a 9 to 5 basis. They also test and evaluate new middleware software and control future upgrades to the Grid Starter Kit. Further roles of the center include to help educate the systems support staff at the UK e-Science Centres and this is being executed through the ETF, and to provide a certificate authority for the UK.

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 2.5 Gbps and there is funding in place to upgrade this to 10 Gbps by mid 2002. 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 joint 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.

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. The demonstrators are being used to illustrate what e-Science and the Grid has to offer, now and in the future. They are suitable for a variety of audience, scientific, non-scientific, academic and industrial. They have also provided a 'bleeding' edge review of the near term requirements for middleware development.

6. Conclusions

There are many challenges to be overcome before we can realize 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 we have noted the UK pilot projects all demand an infrastructure and tools for data management and mining. This implies interfaces for federating database and

techniques for metadata generation along side other data issues. Also as noted, OGSA takes the ideas of web services and extends them to provide support for grid services. The data services within OGSA are yet to be tackled. We have groups in the UK who are already taking a lead on these issues and by working with industry; applications and the Globus team are setting directions for discussion. This is clearly an area where the UK can and will make valuable contributions to the evolving Grid middleware.

Still there are many other challenges yet to be addressed. Two important areas not yet mentioned 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 three-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 and to our colleagues in other research councils who have provided the information regarding research council pilot projects. 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] e-Science definition: <http://www.clrc.ac.uk>
- [2] Foster, I. and Kesselman, C. (eds.). *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, 1999.
- [3] EU DataGrid: <http://www.datagrid.cnr.it/>, <http://www.cern.ch/grid/>
- [4] EU DataTag: <http://www.datatag.org>
- [5] GriPhyn: <http://www.griphyn.org/>
- [6] PPDataGrid: <http://www.ppdg.net/>
- [7] iVDGL: <http://www.ivdgl.org/>
- [8] US NVO: <http://www.srl.caltech.edu/nvo/>
- [9] EU AVO: <http://www.astro-opticon.org/archives.html>

- [10] AccessGrid: <http://www-fp.mcs.anl.gov/fl/accessgrid/>
- [11] Foster I. and Kesselman C., *Globus: A Metacomputing Infrastructure Toolkit*, International Journal of Supercomputer Applications, 11(2): 115-128, 1997.
- [12] M. Litzkow, M. Livny and M. Mutka, 'Condor- A Hunter of Idle Workstations', Proceedings of the 8th International Conference of Distributed Computing Systems, pages 104-111, June 1988.
- [13] Storage Resource Broker: <http://www.npaci.edu/DICE/SRB>
- [14] NASA Information Power Grid: <http://www.nas.nasa.gov/About/IPG/ipg.html>
- [15] IBM position: <http://www.ibm.com/news/us/2001/08/15.html>
- [16] Foster, I., Kesselman, C., Nick, J. and Tuecke, S., *The Physiology of the Grid: Open Grid Services Architecture for Distributed Systems Integration*, to be presented at GGF4, Feb. 2002.
- [17] Paton, N.W., Atkinson, M.P., Dialani, V., Pearson, D., Storey, T. and Watson, P., *Database Access and Integration Services on the Grid*, UK DBTF working paper, Jan. 2002
- [18] Watson, P., Databases and the Grid. Technical Report CS-TR-755, University of Newcastle, 2001.
- [19] Equator Project: <http://www.equator.ac.uk/projects/DigitalCare.html>
- [20] Advanced Knowledge Technologies: <http://www.ecs.soton.ac.uk/~nrs/projects.html>
- [21] De Roure, D., Jennings, N. and Shadbolt, N., *Research Agenda for the Semantic Grid: a Future e-Science Infrastructure*, Report for the UK e-Science Core Programme Directorate and NeSC TR: 2002-1, December 2001, www.semanticgrid.org. [1] <http://www.e-science.clrc.ac.uk/>