Report UK US Large scale grid computing experiment

The TeraGyroid experiment – Supercomputing 2003

R.J. Blake[a,*], P.V. Coveney[b], P. Clarke[c] and S.M. Pickles[d]

[a]*Computational Science and Engineering Department, CCLRC Daresbury Laboratory, Warrington, WA4 4AD, UK*

[b] *Centre for Computational Science, Christopher Ingold Laboratories, University College London, 20 Gordon Street, London WC1H 0AJ, UK*

[c]*Department of Physics and Astronomy, University College London, Gower Street, London, WC1E 6BT, UK*
Present address: National e-Science Centre, e-Science Institute, 15 South College Street, Edinburgh, EH8 9AA, UK

[d]*Manchester Computing, University of Manchester, Oxford Road, Manchester, M13 9PL, UK*

Abstract. Amphiphiles are molecules with hydrophobic tails and hydrophilic heads. When dispersed in solvents, they self assemble into complex mesophases including the beautiful cubic gyroid phase. The goal of the TeraGyroid experiment was to study defect pathways and dynamics in these gyroids. The UK's supercomputing and USA's TeraGrid facilities were coupled together, through a dedicated high-speed network, into a single computational Grid for research work that peaked around the Supercomputing 2003 conference. The gyroids were modeled using lattice Boltzmann methods with parameter spaces explored using many 128^3 and 256^3 grid point simulations, this data being used to inform the world's largest three-dimensional time dependent simulation with 1024^3-grid points. The experiment generated some 2 TBytes of useful data. In terms of Grid technology the project demonstrated the migration of simulations (using Globus middleware) to and fro across the Atlantic exploiting the availability of resources. Integration of the systems accelerated the time to insight. Distributed visualisation of the output datasets enabled the parameter space of the interactions within the complex fluid to be explored from a number of sites, informed by discourse over the Access Grid. The project was sponsored by EPSRC (UK) and NSF (USA) with trans-Atlantic optical bandwidth provided by British Telecommunications.

1. Introduction

The TeraGyroid project was an ambitious experiment to investigate the new opportunities for computational science created by the Grid – for example, demonstrating new scientific capabilities and international collaborative working – at the same time as establishing technical collaborations to support the development of national and international Grids. The federated resources of the UK High-end Computing facilities and the US TeraGrid were harnessed in an accelerated programme of computational condensed matter science that peaked during the Supercomputing 2003 conference (SC2003). The application (LB3D), provided by the RealityGrid project (http://www.realitygrid.org), used lattice-Boltzmann simulations of complex fluids to study the defect dynamics and visco-elastic properties of the gyroid mesophase of amphiphilic liquid crystals [1–4]. The resources available on the combined US-UK grids enabled and accelerated the investigation of phenomena at unprecedented time and length scales that had hitherto been out of reach.

A central theme of RealityGrid is the facilitation of distributed and collaborative exploration of parameter space through computational steering and on-line, high-end visualization [5,6]. The TeraGyroid ex-

*Corresponding author: R.J. Blake, Computational Science and Engineering Department, CCLRC Daresbury Laboratory, Warrington, WA4 4AD, UK. Tel.: +44 (0) 1925 603372; Fax: +44 (0) 1925 603634; E-mail: r.j.blake@dl.ac.uk.

periment realised this theme in dramatic fashion at SC2003. A series of demonstrations under the guise of "Transcontinental RealityGrids for Interactive Collaborative Exploration of Parameter Space" (TRICEPS), won the HPC Challenge competition in the category of "Most Innovative Data Intensive Application" [7]. The same work was also demonstrated live before an Access Grid audience during the SC Global showcase "Application Steering in a Collaborative Environment" [8]. A family of simulations at different resolutions were spawned and steered into different regions of parameter space by geographically distributed scientists on both sides of the Atlantic. The collaborative environment of Access Grid was key, not only for project planning, but also for managing the distributed experiment and discussing the evolution of the simulations. Figure 1 shows a snapshot of the Access Grid screen as seen in Phoenix during Supercomputing Global 2003.

The remainder of this report is organised as follows. In Section 2, we summarise the participants within the Project and their respective roles. In Section 3 we describe the science behind the experiment, the LB3D simulation software, the encapsulating RealityGrid environment and the porting of this application to, and performance on, the various facilities. In Section 4 we discuss the Grid Software Infrastructure, Globus and the components used within the project and developments that had to be undertaken. Section 5 discusses the visualization of the output from the various facilities, in Section 6 the networking infrastructure implemented for the experiment is described. Section 7 overviews the scientific achievements. Section 8 summarises management information related to the organisation, promotion and resources used within the project and Section 9 summarises conclusions and lessons learned.

2. The TeraGyroid testbed and project partners

The TeraGyroid project brought together three distinct groups:

– **RealityGrid Partners:**
   University College London (UCL) – Application, Visualization and Networking
   University of Manchester (UoM) – Grid Middleware, Visualization and Networking
   Edinburgh Parallel Computing Centre (EPCC) – Application
   Tufts University (TU) – Application

– **TeraGrid sites at:**
   Argonne National Laboratory (ANL) – Visualization and Networking
   National Center for Supercomputing Applications (NCSA) – Compute
   Pittsburgh Supercomputing Center (PSC) – Compute and Visualization
   San Diego Supercomputer Center (SDSC) – Compute
– **UK High-end Computing Services**
   HPCx (run by the University of Edinburgh and CCLRC Daresbury Laboratory) – Compute, Networking and Coordination
   CSAR (run by the University of Manchester and CSC) – Compute and Visualization

The principal computational resources integrated into the testbed are listed in Table 1. We also used a number of visualization systems including SGI Onyx systems at Manchester, UCL, NCSA and Phoenix (on loan from SGI and commissioned on site), and the TeraGrid visualization cluster at ANL. The testbed and networks are depicted schematically (in much simplified form) in Fig. 2.

It was known from previous experience [9] that the end-to-end bandwidth available on production networks, coupled with the need to share this fairly with other competing traffic, would seriously hamper the ability to migrate jobs – which requires transfer of large (1 TByte for the largest 1024^3 simulation) checkpoint files – across the Atlantic, and to generate data on one continent and visualize it on the other. We never sought to migrate such large files. For systems of size 128^3, 256^3 we're in the GByte range and that was doable. Fortunately, BT donated two 1 Gbps links from London to Amsterdam, which, in conjunction with the MB-NG project (http://www.mb-ng.net), the SuperJANET development network, the high bandwidth SurfNet provision and the TeraGrid backbone completed a dedicated experimental network to the SC2003 exhibition floor at Phoenix.

3. The science and the lattice-Boltzmann model

An amphiphile is a chemical species whose molecules consist of a hydrophobic tail attached to a hydrophilic head. This dichotomy causes the molecules to self-assemble into complex morphologies when dispersed in solvents, binary immiscible fluid mixtures or melts. Mesophases have been studied in detail for a

Table 1
Principal computational resources in the testbed.

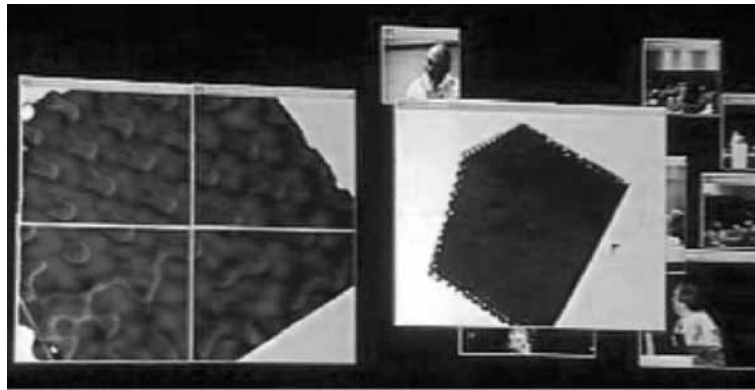| Site	Architecture	Processors	Peak (TF)	Memory (TB)
HPCx (Daresbury)	IBM Power 4 Regatta	1024	6.6	1.024
CSAR (Manchester)	SGI Origin 3800	512	0.8	0.512 (shared)
CSAR (Manchester)	SGI Altix	256	1.3	0.384
PSC	HP-Compaq	3000	6	3.0
NCSA	Itanium 2	256	1.3	0.512
SDSC	Itanium 2	256	1.3	0.512
Total		5304	17.3	5.944

Fig. 1. Access Grid screen as seen in Phoenix during the SC Global session on application steering.
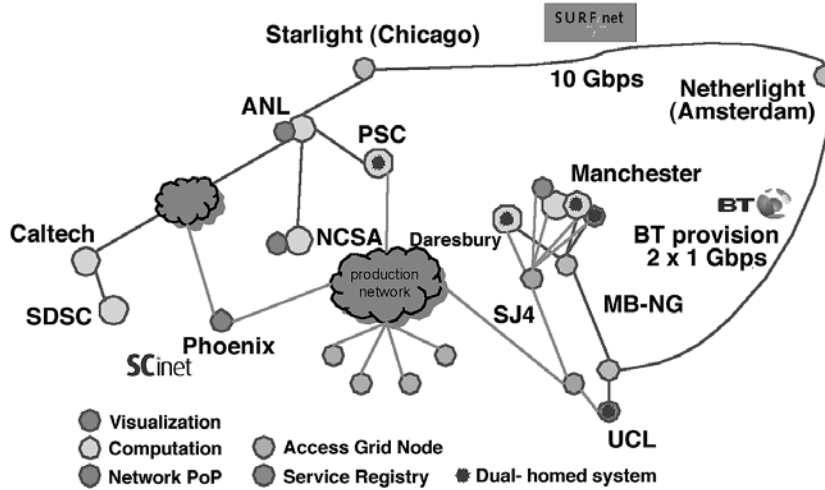

Fig. 2. Schematic of TeraGyroid testbed for SC'03 demonstration.

long time and are a subject of vast technological significance. Some mesophases are liquid crystalline, with features intermediate between a liquid and a solid. The gyroid is one of these: similar to phases ubiquitous in biological systems, it has important applications in membrane protein crystallisation, controlled drug release and biosensors [10]. The gyroid exhibits weak crystallinity and the presence of defects, which play an important role in determining its mechanical properties.

Gyroid mesophases have been observed to form in certain regions of the LB3D model parameter space. Instead of achieving "perfect" crystalline gyroids, with no defects in their structure, several regions containing a gyroid phase are generated, each with a slightly different orientation; there are interface or "defect" regions between the individual gyroid domains, a situa-
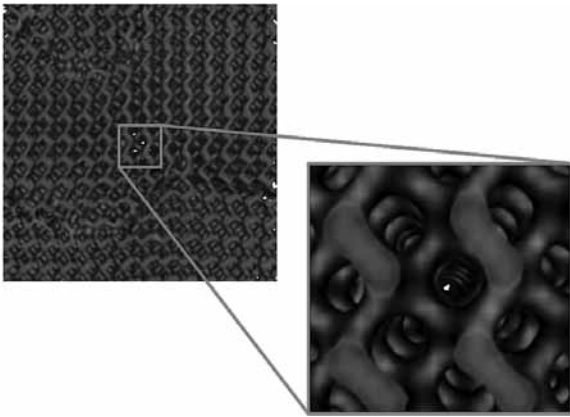
Fig. 3. A volume rendered dataset of a 128^3 system after 100000 simulation timesteps. Multiple gyroid domains have been formed and the close-up shows the extremely regular, crystalline, gyroid structure within a domain.

tion analogous to magnetic domains. Figure 3 shows a volume rendered dataset of a 128^3 system, the minimum size that confers stability on the gyroid phase, avoiding appreciable finite size effects. One can easily see the different domains that stay intact even after very long simulation times. The close-up shows how perfectly the gyroid forms within a given domain. Our previous simulations of liquid crystalline mesophases show that surfactant-surfactant interactions need not be long ranged in order for periodically modulated, long-range ordered structures to self-assemble [2]. The objective of TeraGyroid is to study defect pathways and dynamics in gyroid self-assembly via the largest set of lattice-Boltzmann (LB) simulations ever performed.

The LB3D lattice-Boltzmann model provides a hydrodynamically correct mesoscale fluid simulation method which describes the equilibrium, kinetic, and flow properties of complex surfactant-containing fluids with applications to bulk and confined geometries [3]. The mesoscopic, particulate nature of the lattice-Boltzmann model means that it can be applied to the study of condensed matter on length and timescales intermediate between the microscopic and the macrosopic, while retaining the hydrodynamic interactions that are critical to the evolution of structure. A distinguishing feature of the model used in this work is that it takes a "bottom-up" approach, simulating the behaviour of fluids by specifying mesoscale interactions between fluids, rather than imposing macroscopic behaviour.

The lattice-Boltzmann model captured within the LB3D code, which is fully computationally steerable using a library developed within RealityGrid, allows

researchers to interact with the simulations from local workstations while the compute jobs are running on a massively parallel platform. LB3D was recently awarded a gold star rating for its supreme scaling characteristics on 1024 processors of HPCx. The performance of the code has been optimised by EPCC, who also added parallel HDF5 support. Parallel HDF5 is a data format that allows every processor to write to the same file at the same time. It helps us to reduce memory requirements of the code and improves its scalability substantially. Such a scalable code, combined with access to massive computing power, allows much larger amounts of condensed matter to be probed, effectively bridging the gap in length and time scales between microscopic and macroscopic descriptions that has afflicted simulations for decades.

The TeraGyroid project undertook simulations involving lattices of over one billion sites and for extended simulation times to follow the slow dynamics. It is not only the size of the simulations that is unprecedented: we are also not aware of any other "bottom-up" approaches to study defect dynamics in self-assembled mesophases. This is not surprising since the computational effort that would have to be invested is surely at least on the scale of what we have available within the TeraGyroid project.

The TeraGyroid experiment represents the first use of collaborative, steerable, spawned and migrated processes based on capability computing. Scientific scenarios within the experiment included:

 – exploration of the multi-dimensional fluid coupling parameter space with 64^3 simulations accelerated through steering;
 – study of finite size periodic boundary condition effects, exploring the stability of the density of defects in the 64^3 simulations as they are scaled up to 128^3 and 256^3;
 – exploring the stability of crystalline phases to perturbations as the lattice size is increased over long time scales; and,
 – exploring the stability the crystalline phases to variations in effective temperature through simulated annealing experiments.

The LB3D code is a state-of-the-art application written in Fortran90 with dynamic memory allocation. The model is characterised by the number of grid points in three dimensions multiplied by the number of variables associated with each grid point. For a three dimensional grid with 1024 grid points in each dimension there are approximately 109 points (1 Gpoint) and the

code uses up to 128 variables at each point resulting in double precision (8 bytes per variable) in 1 TByte of main memory. This ignores the requirement for temporary arrays and the vagaries of how memory is allocated and de-allocated dynamically. Problems were experienced in porting the application to TeraGrid sites but these were cleared up once a version of the Intel compiler was found that compiled the code correctly. Site configuration issues were found to be important as most applications assume a system performance model, which is usually tailored (for obvious reasons) to the centre where most of the work is undertaken. It became clear during the project that high performance applications need to characterise performance on a range of target systems in terms of CPU, memory bandwidth, local I/O bandwidth, parallel file system bandwidth and network interconnect speeds between the file system and the outside world.

Inevitably, the application and visualization must be ported to, and deployed on, each resource on the grid where they are to be run. We do not find that Globus-based grids today do much to facilitate this. Instead, the increasing variety of resources that the grid makes accessible means that users and developers are spending an ever-increasing fraction of their time porting applications to new systems and configuring them for the vagaries of each.

One of the major achievements of the project was to undertake (possibly) the world's largest ever Lattice-Boltzmann complex of fluid simulation with:

1024^3 lattice sites
finite-size effect free dynamics
2048 processors
1.5 TB of memory
1 minute per time step on 2048 processors
3000 time steps
1.2TB of visualization data

running on LeMieux at Pittsburgh Supercomputing Centre. This is the class of simulation that should be running routinely on the next generation of UK high-end computing facility.

In RealityGrid, an application is instrumented for computational steering through the RealityGrid steering library. A fully instrumented application, such as LB3D, supports the following operations:

– Pause, resume, detach and stop
– Set values of steerable parameters
– Report values of monitored (read-only) parameters
– *Emit* "samples" to remote systems for e.g. on-line visualization
– *Consume* "samples" from remote systems

– Checkpoint and windback.

Emit and *consume* semantics are used because the application should not be aware of the destination or source of the data. *Windback* here means revert to the state captured in a previous checkpoint without stopping the application. In RealityGrid, the act of taking a checkpoint is the responsibility of the application, which is required to register each checkpoint with the library. LB3D supports *malleable checkpoints*, by which we mean that the application can be restarted on a different number of processors on a system of different architecture.

Checkpoint/recovery is a key piece of functionality for computational steering. Sometimes the scientist realises that an interesting transition has occurred, and wants to study the transition in more detail; this can be accomplished by winding back the simulation to an earlier checkpoint, and increasing the frequency of sample emissions for on-line visualization. An even more compelling scenario arises when computational steering is used for parameter space exploration, as in TRICEPS.

The simulation evolves under an initial choice of parameters until the first signs of emergent structure are seen, and a checkpoint is taken. The simulation evolves further, until the scientist recognises that the system is beginning to equilibrate, and takes another checkpoint. Suspecting that allowing the simulation to equilibrate further will not yield any new insight, the scientist now rewinds to an earlier checkpoint, chooses a different set of parameters, and observes the system's evolution in a new direction. In this way, the scientist assembles a tree of checkpoints – our use of checkpoint trees was inspired by GRASPARC [11] – that sample different regions of the parameter space under study, while carefully husbanding his or her allocation of computer time. The scientist can always revisit a particular branch of the tree at a later time should this prove necessary. This process is illustrated in Fig. 4, in which a Lattice-Boltzmann simulation is used to study the phase behavior of a mixture of fluids. Here one dimension of the parameter space is explored by varying the surfactant-surfactant coupling constant.

Figure 5 shows a schematic representation of the LB3D three-tier architecture for computational steering for the case in which a visualization component is connected to the simulation component. One or both of these components may be steered using the steering client. Both components are started independently and can be attached and detached dynamically. The service-oriented middle tier solves numerous issues that are
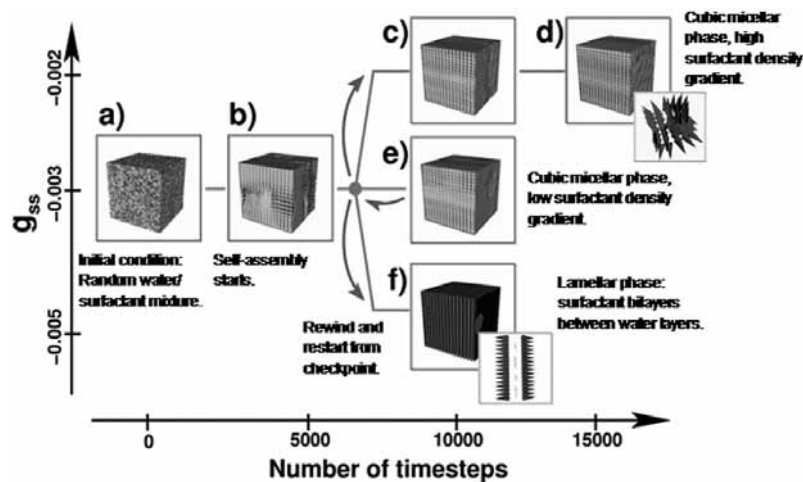
Fig. 4. Parameter space exploration gives rise to a tree of checkpoints. (From Chin et al. [6]).

problematic in a two-tier architecture. The services that support the checkpoint tree are not shown in the figure. Each node in the tree is implemented as a persistent, long-lived Grid service containing metadata about the simulation, such as input decks, location of checkpoint files and so on. The persistence of the tree is achieved by exploiting the ability of OGSI::Lite to mirror, transparently, the service data of a Grid service in a database. On-line, real-time visualization is an important adjunct for many steering scenarios, but we avoid confusing the two, deliberately separating visualization from steering. Control and status information are the province of steering. Visualization is concerned with the consumption of samples emitted by the simulation. The reverse route can be used to emit samples from the visualization system to the simulation, where it might be interpreted for example as a new set of boundary conditions. The volume of data exchanged between simulation and visualization is large and high-performance transport mechanisms are needed. We provide several mechanisms, such as: (a) writing to and reading from disk, relying on a shared file system or a daemon responsible for transferring files from source to destination; or (b) direct process-to-process communication using sockets. At the time, implementation of (b) was based on Globus-IO, but this introduces a dependency that greatly complicates the process of building the steering library and any applications that use it.

4. Grid software infrastructure

Simulations and visualizations must be launched, and services must be deployed. We do this using a suite of scripts that use the command-line interfaces to the GRAM and GridFTP components of the Globus Toolkit. We also provide a graphical tool, or "wizard", that allows the user to choose an initial condition (which can be done by browsing the checkpoint tree and selecting a pre-existing checkpoint), editing the input deck, choosing a computational resource, launching the job (automatically starting the Steering Grid Service), and starting the steering client. The wizard also provides capabilities to select a visualization technique and visualization resource, start the visualization component and connect it to the simulation. The wizard handles the task of migrating a running job to a different resource on the grid, which involves taking a checkpoint, transferring the checkpoint files, and re-starting the application on the new host. The wizard shells out to command-line scripts, which in general require some customisation for each application, to accomplish these tasks. With the exception of Globus-IO, we do not program to the Globus APIs directly.

Note that we did not use the MDS components of Globus (GIIS and GRIS). This is not because of any doubt about the importance of such services – we view them as essential for the long-term future of the grid – but because our experiences with MDS as deployed on the grids available to us give us reason to doubt the robustness of the services and the utility of the information published through them. Instead, we maintain lists of computational and visualization resources in client-side configuration files read by the wizard.

The RealityGrid registry ran on a Sony PlayStation 2 in Manchester. The Globus Toolkit was installed on all systems, with various versions (2.2.3, 2.2.4, 2.4.3
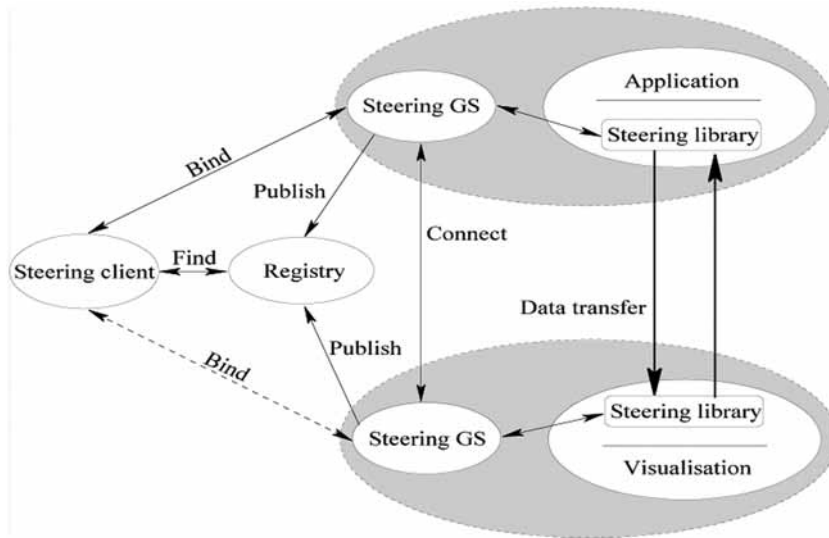
Fig. 5. Schematic architecture of an archetypal RealityGrid configuration.

and 3.1) in use. Our use of Globus (GRAM, GridFTP and Globus-IO) exposed no incompatibilities between these versions, but we note that the GT 3.1 installation included the GT 2 compatibility bundles.

For the TeraGyroid simulation to work it was necessary to implement a 64-bit version of Globus GT2, on the HPCx IBM system, in particular the globus io libraries which are used to communicate data between the steering code and the computational code running on the HPCx nodes. Whilst we had previous experience of Globus GT2 on AIX systems and had already installed the GT2.2 version released as an installp binary by AIX, this did not contain any 64-bit code and also did not have the required libraries. Progress in establishing the grid software infrastructure took a lot longer than anticipated, nearly a month in total. The following is an example of the steps that had to be taken to enable HPCx for the experiment to work:

i) It was decided, following compatibility tests, to use the latest GT2.4.3 source.

ii) Trial builds were done in 32-bit more on a Power 3 system and the HPCx test and development system. This was successful.

iii) Work then focussed on building a 64-bit code from source. This was unsuccessful as some code would not build and users' certificates could not be read.

iv) It was discovered that the version of OpenSSL (0.9.6i) shipped with GT2.4.3 was obsolete and does not work on AIX 5 in 64-bit mode. The new v0.9.7c did. However it was not straight-

forward to retro-fit the open source distribution into the Globus package framework so that the two would build together.

v) We finally managed to compile OpenSSL independently and change the library names to match the required names for the Globus distribution then build the rest. The build still would not accept UK e-Science certificates.

vi) There was discovered (by Mike Jones, Manchester) to be a problem with the way the old certificates had expected an "eMail=" field. It was found that actually an "=eMailAddress" field was produced which is now the accepted standard. Fortunately it was possible to change a simple character string to make this work.

vii) This resulted in a working version and system changes could be made to enable the Globus gatekeeper to be started from inetd and map user's certificate names onto local user names.

viii) The same software was ported onto login.hpcx. ac.uk.

ix) changes were made to port the software onto a complementary login node using the dns name for the trans-atlantic MB-NG network (changes required in local dns tables to get correct reverse lookup).

x) HPCx has head nodes (login) and worker (development/ interactive/ production) nodes, the latter are on a private network. This is common practice on large systems, especially clusters. Stephen Booth (EPCC) wrote a generic

port forwarder which started using inetd when a node process attempted to contact an external ip address. This was used to transfer the visualization data off the nodes and to send back steering parameter changes.

Unfortunately, a large amount of effort had to be invested in trying to work with or around the Globus Toolkit, supposedly the standard middleware for Grid applications and that underpinning both the UK e-Science Grid and the US TeraGrid. The RealityGrid experience over two years has been that construction of applications or Grid services using Globus is not straightforward. In addition, a vast amount of systems administrator time was required to compile and install it on the many machines involved in the project, often due to its tendency to require custom-patched versions of third-party libraries, rather than building on top of existing infrastructure. We have found that it is much easier to construct Grid services using OGSI::Lite, a lightweight implementation of the emerging Open Grid Service Infrastructure standard [12].

Considerable negotiation was necessary to persuade TeraGrid sites to recognise user and host certificates issued by the UK e-Science Certificate Authority (CA), and to persuade UK sites to recognise the five different CAs used within TeraGrid. In our case, the agreement was facilitated by a sense of common purpose. In general, however, establishing the necessary trust relationships to federate two grids is a process that will scale poorly if every systems administrator has to study the policies of every CA; there is a clear role here for Policy Management Authorities (see http://www.gridpma.org).

5. Visualization

Amphiphilic fluids produce exotic mesophases with a range of complex morphologies. These can only be fully understood using sophisticated visualization software, semi-immersive or fully immersive virtual reality techniques being essential in some cases. The complexity of these data sets makes visualization of large images itself a major graphics challenge. Using the VTK library, we were able to view up to 512^3 data sets of the gyroid morphology using a PC (2 GB memory and hardware volume rendering are minimum requirements). Visualization of billion node models requires 64-bit hardware and multiple rendering units (such as multipipe OpenGL or Chromium, with which SGI IR4

pipes and TeraGrid Viz resources are equipped). The size of the datasets we generated challenged most available visualization packages; indeed, the 1024^3 data sets were only visualized for the first time on the show room floor at SC03 using a ray-tracing algorithm developed at the University of Utah running on 100 processors of an SGI Altix. Therefore, access to large scale visualization resources is an important requirement. At the time of the TerGyroid project it was clear that the UK high-end visualization capabilities (and TeraGrid) were not capable of dealing with large scale grid based simulations of order 1 Gpoints.

Scientists and developers collaborated using Access Grid nodes located in Boston, London, Manchester, Martlesham, and Phoenix. The tiled display on the left of Fig. 1 was rendered in real time at the TeraGrid visualization cluster located at ANL, using Chromium – a cluster facility. The display on the right was rendered, also in real time, on an SGI Onyx system in Manchester. The visualizations were written in VTK with patches to refresh automatically whenever a new sample arrives from the simulation. The video streams were multicast to Access Grid using the FLXmitter library. SGI OpenGL VizServer^{TM} was useful to allow a remote collaborator to take control of the visualization. When necessary, the steering client was shared using VNC. An IRC back channel also proved invaluable. All Access Grid and VizServer traffic was routed over the MB-NG production networks.

6. Networking

The event provided a good opportunity to trial the complexity in establishing a collaboration between all of the relevant network bodies in the chain. This is a larger number than one might initially guess:

- University of Manchester Computing Service
- Daresbury Laboratory Networking Group
- MB-NG and UKERNA
- UCL Computing Service
- BT
- SurfNET (NL)
- Starlight (US)
- Internet-2 (US)

for all of whom coordination was required to establish configuration, routing and commissioning. A set of meetings was held between the network managers in the UK, the Netherlands and the US to solve all the technical problems and agree on what the network should pro-
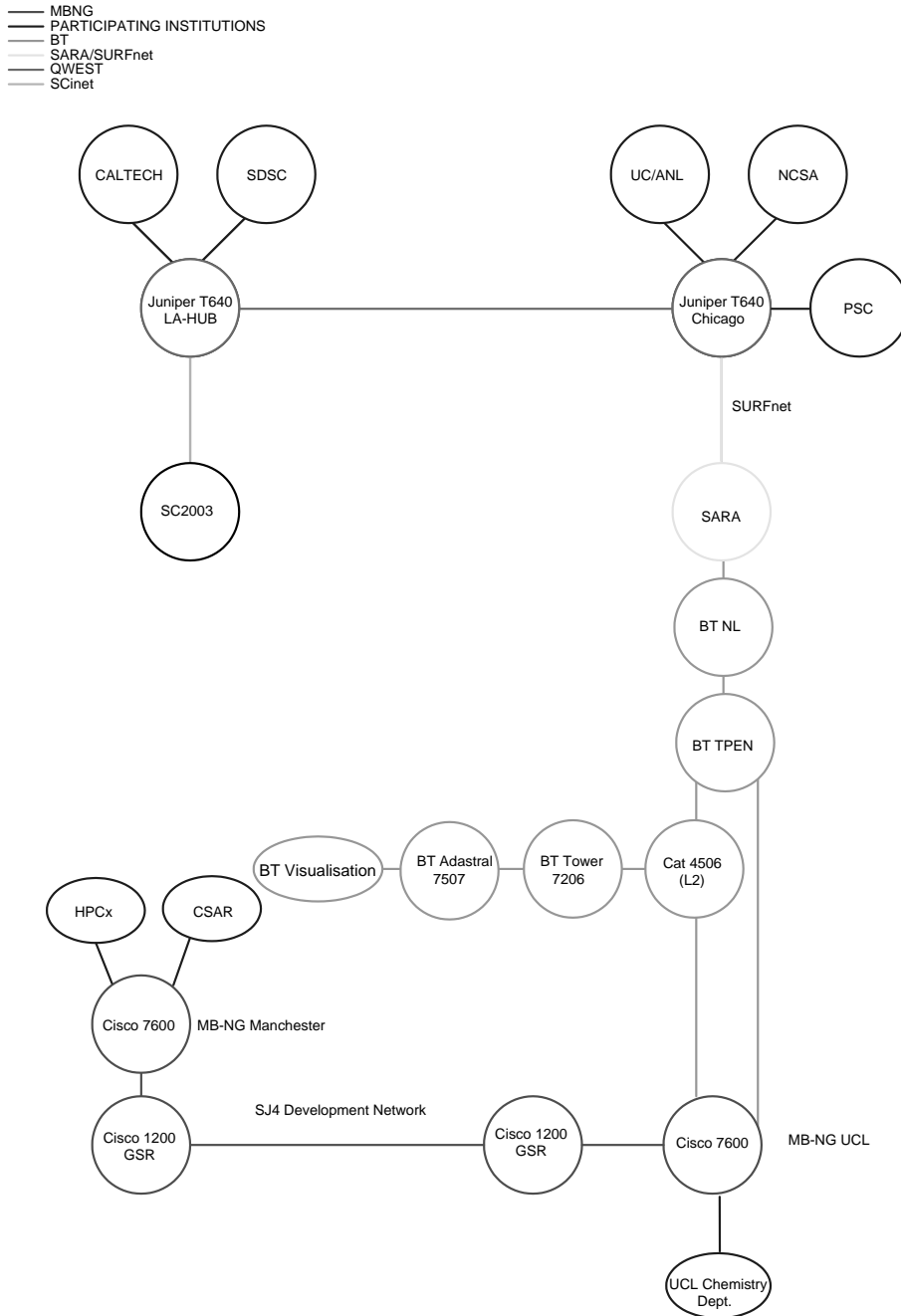
Fig. 6. Network connectivity used for SC'03.

vide. Once the network was in place networking people were joined by application people running the tests to solve end-to-end problems related to inconsistency between test traffic and real applications traffic. In fact this went exceptionally well, demonstrating both good will and that the right peer relations exist between technical responsibles. The configuration, showing router positioning, is displayed Fig. 6.

BT deserves special mention for project management and for the provision of the link from UCL to SurfNet in Amsterdam. This required interworking differing networking elements, such as native Gigabit Ethernet, CWDM, and concatenated STM16 across two STM-64 SDH rings.
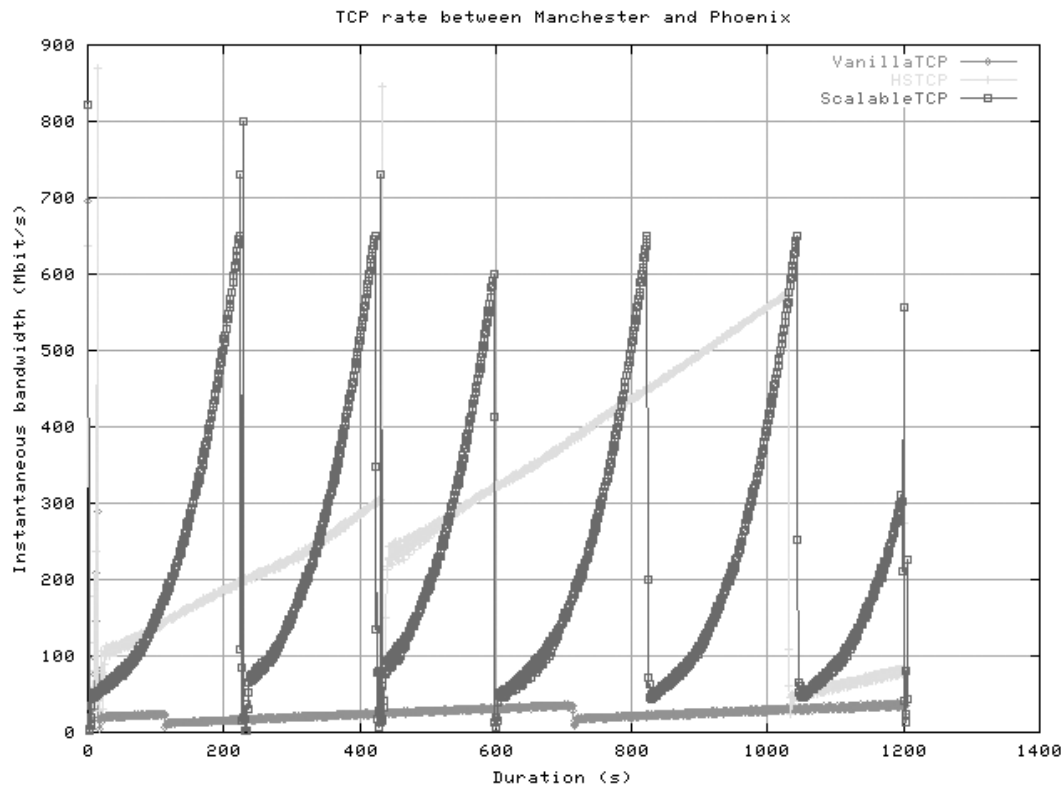
Fig. 7. Measured transatlantic bandwidths during SC2003.

HPCx, the CSAR Origin 3800, and the visualization supercomputers in Manchester and UCL were all "dual-homed", connected simultaneously to the dedicated network and to the SuperJanet 4 academic backbone. Complex routing tables were required on the UK end, while route advertisement sufficed on the US end. Dual routing enabled researchers to log in and issue low data rate control and steering commands via the production network whilst enabling the high bandwidth transfer of files via the dedicated experimental network.

It is normally the task of the RealityGrid migration wizard on the user's laptop to initiate the transfer of checkpoint files using the third-party file-transfer capability of GridFTP. This proved problematic when one of the systems involved was dual-homed, and complicated by the fact that the systems were not accessible from everywhere via their MB-NG addresses. The problem would not have occurred if GridFTP provided the ability to specify different addresses for control and data channels. We worked around the problem by setting up a checkpoint transfer service on one of the dual-homed systems for which both networks were visible. This transfer service was aware of which hosts were dual-homed, which interface would provide the best bandwidth for each pair of endpoints, which host certificate to associate with an interface for Globus authentication, which options to globus-url-copy yielded the best performance, and which checkpoint replica was "closest" to the site where it was required. The transfer service was implemented as a Perl script and invoked remotely using globusrun, which introduces a significant start-up overhead.

It was often difficult to identify whether a problem was caused by middleware or networks. We are used to the symptoms caused by the presence of firewalls, and were generally able to resolve these without undue difficulty once the relevant administrators were identified. However, even more baffling problems arose. One problem, which manifested as an authentication failure, was ultimately traced to a router unable to process jumbo frames. Another was eventually traced to inconsistent results of reverse DNS lookups on one of the TeraGrid systems. A third problem manifesting as grossly asymmetric performance between ftp get and put operations was never completely resolved, but MTU settings appear to have been implicated.

This exercise has provided very valuable pre-UKLIGHT experience which will feed into methods of operation which must be in place when we start to exploit the UKLIGHT experimental links, which come online in early 2004. The TeraGyroid project demonstrated that we have the appropriate peer relation chain to enable connection from the UK to the USA. As a result complex dedicated circuits can be provisioned at short notice.

The installed capacity, composed of a mix of layer 1 technologies (OC-192, OC-48, 1 Gigabit and 10 Gigabit Ethernet), had an overall capacity of 2×1 Gigabit per second. Between any two hosts, up to 1 Gbit/s could be achieved. From Manchester to Chicago we were able to achieve a UDP rate of 989 Mbit/s using Jumbo frames and 957 Mbit/s using a standard MTU size; both tests were memory to memory. However during SC2003, the maximum achieved throughput for actual checkpoint files (using TCP) was only 656 Mbit/s. This was primarily due to:

- TCP stacks on end systems: It is well understood that standard TCP has performance problems in high bandwidth high latency environment. Since the machines involved in the tests were SGI with the IRIX operating system, no kernel patches were available for trying other protocols other than the standard vanilla TCP stack.
- MTU mismatch: At the time of the transfers, there were still issues with configuring the hosts with MTUs of 8192, so standard MTUs were used.
- Disks on end systems: There is a performance hit on the end-to-end transfer when standard end-system disks are used. High performance disk systems are required for optimum performance.
- Network interface card: The configuration of the interface card driver is significant in achieving the best end-to-end performance. We did not have time to investigate all the different options for the SGI end hosts.

This is typical, and demonstrated that which is well known – that practical transport rates are limited by the end systems once the protocol limitation is removed. All of these need addressing in the end systems well in advance, in order to take advantage of multi-Gbit/s network links.

The plot in Fig. 7 shows the achieved rate in a test immediately following the SC2003 event. The measurements shown in this plot were conducted between a Linux machine based at Manchester and an SGI IRIX machine. Initially, up to 700 Mbit/s was observed with
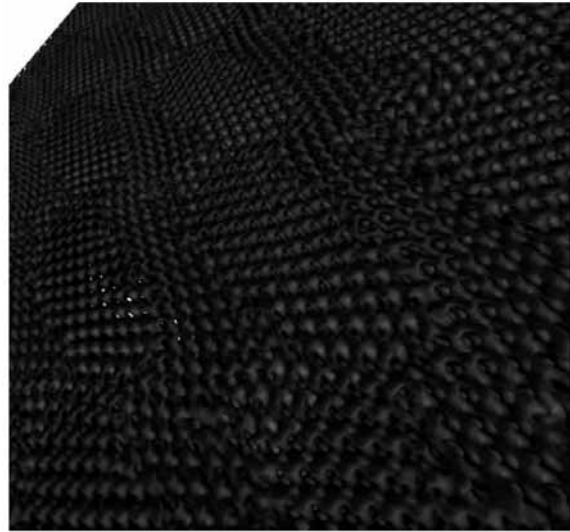

Fig. 8. Shows a cropped plane from a volume rendered example dataset of the upscaled 10243 system. The periodic artefacts introduced due to the upscaling can still be seen here.

out-of-the-box ("vanilla") TCP, but once packets are lost, it remains below 50 Mbit/s for the duration of the test. A better performance is observed when the sender's TCP stack is changed for High Speed TCP (HSTCP) [14] or scalable TCP [15] as shown in the plot.

Looking towards the future it is clear that in order for UK HPC machines to regularly communicate with remote machines as part of production operation, then deterministic Gbit/s scale transport is required. This should be emphasised in future equipment planning. This is an aspect of HPC largely ignored until now due to most computing being constrained to remain local. In order to achieve Gbit/s scale transport rates into the production network a serious, policy driven effort is required to address the operating system, transport protocol stack, network interfaces, and disk I/O on the HPC machines. This could lend itself to collaborative R&D projects. Related to this is the following: some of the issues (not all, eg disks) would be obviated by being able to advance schedule and use network services such as dedicated circuits. As a result a checkpoint file could be transported in a completely deterministic way and non-TCP transport can be used in this case.

The connection was very reliable. We had tests which ran overnight for 12 hours using a specialist piece of network equipment to generate continuous high rate traffic streams (SmartBits) at line rate. This showed 0% losses on the link up to Chicago.

7. Scientific achievements

During the weeks around and mainly at the time of Supercomputing 2003 in Phoenix, USA, about two terabytes of scientific data were generated. Obviously, the analysis of such huge amounts of data is a very time consuming task and will take much longer than the simulations themselves. In addition, we are grateful to have been given the chance to consume the CPU time we were allocated for the period up to Supercomputing 2003, but could not use up by then. This gives us the opportunity to try to answer questions that are arising during the analysis of the data already generated.

It is important to make sure our simulations are virtually free of finite size effects. Therefore, we simulated different system sizes from 64^3 to 1024^3. Most of the simulations were for about 100,000 timesteps, but some reach up to 750, 000 and at least one case eventually passed the one1 million mark. For 100,000 timesteps we found that 256^3 or even 128^3 simulations do not suffer from finite size effects, but we still have to check our data for reliability after very long simulation times since we are especially interested in the long-term behaviour of the gyroid mesophase. This includes trying to generate a 'perfect' crystal. Even after 600,000 timesteps, differently orientated domains can be found in a 128^3 system. Even though the number of individual domains decreased substantially, we can not find a 'perfect' gyroid. Instead, defects are still moving around and it is of particular interest to study the exact behaviour of the defect movement. Since all generated gyroids are different, this can be done best by gathering statistics within the data of each simulation, i.e. counting the number of defects per data set, the direction of their movement, their speed, their lifetime and other interesting parameters. Being able to gather statistics implies that large numbers of measurements are available which is the reason for the large (512^3 and 1024^3) systems we simulated because only a large system has a sufficient number of defects.

Simulating a 512^3 or 1024^3 system raises some technical problems. First, the memory requirements exceed the amount of RAM available on most of the supercomputers we have access to and limits us to two or three facilities. Second, achieving a sufficient simulation length requires a lot of CPU time. For example, running on 2048 CPUs of Lemieux we are only able to do about 100 timesteps of a 1024^3 system per hour (including IO). Third, such large simulations generate data files which are 4 GB each and checkpoints which are 0.5TB each. It is quite awkward to move around such amounts of data using the infrastructure available today. Due to the amount of manual intervention required and restrictions in machine availability, these huge simulations, probably the biggest lattice-Boltzmann simulations ever performed, were still ongoing many months later.

For a 64^3 or 128^3 system it is very easy to run a simulation from scratch for 100, 000 timesteps, but for a 1024^3 system this would require more than two million CPU hours or 42 days on Lemieux. Obviously, it is very hard to justify the use of such an amount of resources for a single simulation. Therefore, we decided to scale up a smaller system after a gyroid was formed by filling up the 1024^3 lattice with 512 identical 128^3 systems. This was possible because we use periodic boundary conditions. In order to reduce effects introduced due to the periodic upscaling, we perturbed the system and let it evolve. We anticipated that the unphysical effects introduced by the upscaling process would decay after a corresponding small number of timesteps, thus resulting in a system that is comparable to one that started from a random mixture of fluids. This has to be justified by comparing to data achieved from test runs performed on smaller systems.

Another experiment, performed in order to speed up the generation of a well-established gyroid, and which was still ongoing at the time of writing, uses the influence of the surfactant temperature on the mobility of the surfactant particles. After differently oriented gyroid domains have been formed, the system will be evolving very slowly towards a perfect crystal. Therefore, we increase the surfactant temperature, which will start to break up the gyroid. After a short while, we push the temperature back to its original value, hoping that the system will anneal into a state which contains fewer different domains. This is a typical application where computational steering is useful, since we can immediately see how the system behaves and use this information to adapt the change of the temperature.

Other experiments which are still ongoing are related to the stability of the gyroid mesophase. We are interested in the influence of perturbation on a gyroid. Using computational steering we are able to tune the strength of the perturbation and find out if the crystal stays intact or breaks up. Similar experiments are being done by applying constant or oscillatory shear. Here, we study the gyroid stability dependence on the shear rate, and expect to find evidence of the non-Newtonian properties of the fluid.

The RealityGrid environment provides not only a very convenient way to combine job handling on HPC

resources, but the use of high performance visualization facilities is naturally included. And without this feature, computational steering would be of very limited use for us. Without our new platform independent checkpointing and migration routines, we would not have been able to adapt our scientific requirements to the available compute resources as easily. We also benefitted substantially from the code optimizations and IO improvements that have been introduced by various people within the RealityGrid project.

None of the experiments we have done or are still doing would have been possible without the benefits obtained from software development within the RealityGrid project. Computational steering allowed us to monitor our simulations and to use the available resources as efficiently as possible. We rehearsed and staged several distributed, collaborative sessions between the geographically-dispersed participating scientists, making use of Access Grid. These collaborators were based at UCL, Manchester, Boston University and BTexact (Ipswich) (as well as Phoenix during SC03), who were each able to run and steer simulations. Indeed, Access Grid proved an indispensable tool [17] for co-ordinating the project, having been used successfully to conduct both regular multi-site management meetings and small ad hoc conferences that might otherwise have required a telephone conference or physical meeting. IRC channels proved very effective for user communications during assigned supercomputing slots before and during SC2003.

Even the simplest computational steering scenarios require co-allocation of computational and visualization resources [13]. For a scheduled collaborative steering session, it is also necessary to reserve these resources in advance at a time that suits the people involved. When Access Grid is used to provide the collaborative environment, one usually needs to book the physical rooms as well. Although some batch queuing systems support advance reservation, such support is far from universal, and standard protocols for reservation are still lacking. Making the necessary arrangements will continue to be problematic for some time to come.

8. Management/Promotion

The TeraGrid project emerged from discussions between the UK High-end Computing (HEC) and the TeraGrid communities about how to demonstrate the potential of international Grids in advancing real scientific applications. For this experiment to be genuinely useful, it would be necessary to ensure that it was not just another transatlantic metacomputing demo for Supercomputing. Rather, it should be an experiment that demonstrated the new opportunities available through the Grid, for example new scientific capabilities and international collaborative working, and that it should seek to establish technical collaborations to support the development of national and international Grids.

Given the UK HEC focus on capability class applications the project would be successful if it undertook a simulation exploiting a community code that runs on the integrated resources of the TeraGrid and UK facilities which amounts to some 6TB of memory and some 5 K processors. An application that could exploit this scale of computational resource would be well on the way to the class of application that would be targeted at the early phase of a next generation next supercomputing facility in the UK. It is worth noting that the world Grid capabilities are changing rapidly with a doubling in the HPCx performance (6TF Linpack) 2004 and the installation of the ETF facilities – 10TF at NCSA and 4TF at SDSC.

The indicative plan (with estimated resource requirements) to undertake a scientific experiment on the gyroid phases of ternary fluids on the integrated UK-TeraGrid high-end computing resource was discussed with Rick Stephens and Peter Beckmann at the 'All Hands' meeting in Nottinghom?? in early September 2003. It is a small miracle that an award winning demonstration was undertaken some 2 months later at Supercomputing in November 2003. The following diary illustrates some of the activity:

- Approval from ETF for project 12 September 2003
- Richard Blake asked to coordinate UK Input – 16 September 2003
- Establishment of project teams in UK:
  RealityGrid: Stephen Pickles – technical e-science
  RealityGrid: Peter Coveney – science
  HPCx: Stephen Booth – technical
  CSAR: Kevin Roy – technical
  UK Grid Support centre: Robert Allan – Daresbury Laboratory
  Networking: Peter Clarke – University College London
  Coordination: Richard Blake -Daresbury Laboratory
- 22 September 2003: first joint meeting of UK and US sites via Access Grid. Discussed background, potential scenarios and requirements, networking, Globus infrastructure, securing ids on TeraGrid fa-

cilities and applications porting. Established peer group interactions.

– 1 October 2003: Review of UK progress – discussion on where to visualize largest datasets, size of calculations that could be run on each facility. Scientific programme of runs to be developed and placed on WIKI. Detailed resource request plan and schedule of outages to be prepared. Draft flier and advance publicity to be prepared.

– 10 October 2003: Progress review – BT to provide network to NetherLight by 31 October. Acceptance of UK certificates by TeraGrid required. Resource requests submitted to EPSRC and TeraGrid sites. DNS names of all systems required. Need to identify which visualization servers could cope with largest datasets. Visibility of production nodes to outside world highlighted as an issue at PSC and HPCx.

– 20 October 2003: Scientific project plan published. 64 bit Globus version for AIX – problems in porting this. Resources and schedule of test slots agreed. 1 Gpoint simulation not able to fit into HPCx memory.

– 31 October – application built on all systems, visualization for 768^3 system sizes demonstrated on Chromium and SGIs. Flier completed. Network in place – being tested.

– 30 October, 4 November, 6 November, 11 November, 13 November, and 15 November – 6 six-hour development slots on all facilities. 18 November – 21 November – experiment and demonstration at SC'03.

All of these meetings were conducted via Access Grid with 20+ participants at many of them. The meetings could more sensibly have been split into separate management and technical sessions but time did not permit. The split into project teams worked well with the team leaders progressing actions vigorously between meetings.

With a longer time to plan the project (formal go ahead on an indicative plan was only granted in early September) the project could have developed more precise estimates of resources and a more relaxed schedule of slots for development work with more warning. Whilst the experiment was disruptive to users there were few complaints.

In terms of effort the TeraGyroid project involved of order 40 staff days at each of Daresbury Laboratory (including EPCC), Manchester and University College London with effort focussed on coordination, Globus porting, systems support and development, port forwarding, applications porting and optimisation and networking.

The project was allocated significant resources on the UK and USA TeraGrid Resources in particular:

HPCx: 351K Allocation Units

CSAR: 55, 000 CPU hours on Green: 25, 000 C-P-U hours on Newton.

TeraGrid: 215K CPU hours

The following demonstrations of the TeraGyroid experiment were presented at Supercomputing:

– TeraGyroid on the PSC Booth
Tue 18, 10:00–11:00
Thu 20, 10:00–11:00
– RealityGrid and TeraGyroid on UK e-Science Booth
Tue 18, 16:00–16:30
Wed 19, 15:30–16:00
– RealityGrid during the SC'03 poster session:
Tue 18, 17:00–19:00
– HPC-Challenge presentations:
Wed 19 10:30–12:00
– SC Global session on steering:
Thu 20, 10:30–12:00

Demonstrations and real-time output were available at the University of Manchester and HPCx booths.

Fliers describing the project were available from the booths of the participating sites. A rolling powerpoint presentation summarizing the project was made available to booths from all of the participating sites.

We are aware of the following press releases and references to the TeraGyroid project in the press.

– EPSRC achieves a world first in high performance computing: Issue date: 21st November 2003.
– http://www.realitygrid.org/tg-pr.html
– www.psc.edu/science/2004/teragyroid
– www.britainusa.com/science/articles-show asp? a=7648

Explicit publications from the project included:

The TeraGyroid Experiment: S.M. Pickles, R.J. Blake, B. Boghosian, J.M. Brooke, J. Chin, P.E.L. Clarke, P.V. Coveney, R. Haines, J. Harting, M. Harvey, S. Jha, M.A.S. Jones, M. Mc Keown, R.L. Pinning, A.R. Porter, K. Roy, and M. Riding. Proceedings GGF10.

The TeraGyroid Experiment, S.M. Pickles, R.J. Blake, B.M. Boghosian, J.M. Brooke, J. Chin, P.E.L. Clarke, P.V. Coveney, R. Haines, J. Harting, M. Harvey, S. Jha, M.A.S. Jones, M. McKeown, R.L. Pinning, A.R. Porter, K. Roy and M. Riding, Proceedings ISC 04.

The second of these publications won an award at ISC 04 in the category Integrated Data and Information Management.

The success of the project was also influential in EPSRC issuing a Call for Proposals in collaboration with the NSF for new experiments to be demonstrated at SC2005.

9. Conclusion

In the following we summarise the extent to which the Project met its stated objectives:

- *securing high-level recognition for UK computational science on the world stage:* the demonstration of the experiment at SC2003 resulted in the project winning the HPC Challenge competition in the category of "Most Innovative Data Intensive Application" and winning an award at ISC 04 in the category "Integrated Data and Information Management".
- Grid aspects of the project have been written up for a paper presented at GGF10 and scientific publications are in preparation.
- *enhanced expertise and experience to benefit the UK HEC and e-Science communities:* the project resulted in the first transatlantic federation of major high performance computing facilities through the use of Grid technology. The project pointed to the need to develop robust applications that can exploit a range of different high-end systems architectures.
- *added value to the construction/operation of the UK e-Science grid:* the technical teams developed valuable interworking relationships, as well as experience of the practical issues in respect of federation of such facilities by a "high performance backplane". The results feed directly into the proposed ESLEA project to exploit the UKLIGHT infrastructure.
- *data to inform strategic decisions about the future UK involvement in global grids:* remote visualisation, checkpointing and steering require access to very high bandwidths which need to be dedicated and reservable.
- *long-term strategic technical collaboration with ETF sites.* The project established and strengthened relationships in particular between the visualisation and Globus communities in the UK and the USA. The project gave an insight into what is happening in the USA and established strong technical relationships between operational HEC Grid centres in both countries.
- *long-term scientific collaborations with users of the ETF sites.* The project built on a strong and fruitful existing collaboration and such a relationship should be the cornerstone of future projects.
- *experiments with clear scientific deliverables* – an explicit science plan was developed, approved and executed. Data analysis from the experiment is still ongoing.
- *choice of applications to ensure that benefits are returned to computational science and e-Science community* – the discretisation grid based technology is key to a number of other applications in particular in the area of computational engineering and the experiences have already been reported through publications accepted by GGF10.
- *a reciprocal arrangement that encourages ETF participants to contribute complementary demonstrators* – it was not possible given the late approval date to put effort into supporting a separate US driven project.

This was a significant achievement, which helped to confirm that the UK has a world-class presence in the area of High-end Computing and e-Science. It is gratifying to see that this initial experiment will be followed by a second series of projects targeted at SC2005, currently the subject of a Joint Call for Proposals from EPSRC and NSF. Lessons learned from the initial experiment include:

- The proposal was developed without peer review although the experiment built on existing peer-reviewed projects. Collaborative ventures like this are valuable to explore how to work with counterparts, and tease out technical issues. They gain enormous "intangible" benefits as noted above. The projects may not always be appropriate for peer review if such concentrates purely on the scientific output.
- The original intent was to have two experiments, one from the UK and one from the USA. In the event, there wasn't one from the USA. Future experiments should be bilateral – given that these experiments will be planned and approved much further in advance of the target date this should not be a problem in the future.
- The cost of carrying out the experiment was higher than initially estimated, as computer time was not budgeted for explicitly in the original proposal. A

considerable amount of work had to be undertaken to port, optimise and validate the simulations on large numbers of processors. There does need to be some flexibility in terms of resource allocation as it is the nature of these exploratory experiments that it is not always possible to quantify the exact resource requirements in advance of undertaking a considerable amount of work.

– There was some disruption to the other users of the HPCx and CSAR services. Planning in future years should be improved to a) reduce the disruption to the minimum, and b) give the users plenty of warning of any scheduled down time. Few complaints were received from the users about the level of disruption, indeed a number of positive messages were received.

– HPCx and CSAR worked well together – a good example of collaboration between the two UK national academic supercomputing services. The project also provided a good example of collaboration between the UK High-end Computing and e-Science Programmes;

– Site configuration issues are important as most applications assume a specific system performance model. It became clear during the project that applications need to characterise performance on a range of target systems in terms of CPU, memory bandwidth, local I/O bandwidth, parallel file system bandwidth and network interconnect speeds between the file system and the outside world.

– Visualisation facilities within the UK are not capable of meeting the needs of the largest-scale capability simulations undertaken within this project.

– Effort needs to be put into developing a scalable solution for Certificate Authorities to accept each other's certificates. Work needs to be put into managing communications between systems with multiple access addresses effort needs to be put into defining the external visibility of high-end computing systems during the specification of their requirements.

– Technical interconnection by a high speed link in even this "simplest scenario" way still required significant planning due to the large amount of equipment en-route. Unless high end machines are purchased with this concept in mind, they will most likely not have the required high performance network interfaces or transport stacks in place to utilise the available bandwidth.

– The TeraGyroid experiment has generated a wealth of data that has taken a deal of time to analyse

and produce new science [16]. This should not be forgotten in planning these experiments.

It is difficult at Supercomputing for anyone to make an impact, but the success of the experiment did attract a great deal of interest with a number of Press Releases and media articles. This helped the UK to have a much stronger performance at SC2003 than previously.

Acknowledgements

References

[1] N. González-Segredo, M. Nekovee and P.V. Coveney, *Phys. Rev. E* **67**(046304) (2003); M. Nekovee and P.V. Coveney, *J. Am. Chem. Soc.* **123** (2001), 12380.

[2] N. González-Segredo and P.V. Coveney, Self-assembly of the gyroid cubic mesophase: lattice-Boltzmann simulations, *Europhys. Lett.* **65** (2004), 795.

[3] H. Chen, B.M. Boghosian, P.V. Coveney and M. Nekovee, *Proc. R. Soc. London A* **456** (2000), 2043.

[4] P.J. Love, M. Nekovee, P.V. Coveney, J. Chin, N. González-Segredo and J.M.R. Martin, *Comp. Phys. Commun.* **153**(3) (2003), 340–358.

[5] J.M. Brooke, P.V. Coveney, J. Harting, S. Jha, S.M. Pickles, R.L. Pinning and A.R. Porter, *Computational Steering in RealityGrid*, Proceedings of the UK e-Science All Hands Meeting, September 2–4, 2003, http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/179.pdf.

[6] J. Chin, J. Harting, S. Jha, P.V. Coveney, A.R. Porter and S.M. Pickles, Steering in computational science: mesoscale modelling and simulation, *Contemporary Physics* **44** (2003), 417–434, http://taylorandfrancis.metapress.com/openurl.asp?genre=article&eissn=1366-5812&volume=44&issue=5&spage=417.

[7] Stephen M. Pickles, Peter V. Coveney and Bruce M. Boghosian, *Transcontinental RealityGrids for Interactive Collaborative Exploration of Parameter Space* (*TRICEPS*), Winner of SC'03 HPC Challenge competition in the category Most Innovative Data-Intensive Application, http://www.sc-conference.org/sc2003/inter cal/inter cal detail.php?eventid=10701#5.

[8] M. Romberg, J. Brooke, T. Eickermann, W. Woessner, B. Boghosian, M. Nekovee and P. Coveney, *Application Steering in a Collaborative Environment*, SC Global conference, November, 2003, http://www.sc-conference.org/sc2003/inter cal/inter cal detail.php?eventid=10719. A Windows Media Stream archive of this session is available at mms://winmedia.internet2.edu/VB-on-Demand/AppSteering.asf.

[9] S.M. Pickles, J.M. Brooke, F. Costen, E. Gabriel, M. Mueller and M. Resch, Metacomputing across intercontinental networks, *Future Generation Computer Systems* **17** (2001), 911–918.

[10] W.M. Gelbart, D. Roux and A. Ben-Shaul, *Modern ideas and problems in amphiphilic science*, Springer: Berlin, 1993; D. Chapman and M.N. Jones: Micelles, monolayers and biomembranes, Wiley: Chichester, 1995.

[11] K.W. Brodlie, L.A. Brankin, G.A. Banecki, A. Gay, A. Poon and H. Wright, GRASPARC: A problem solving environment integrating computation and visualization, in: *Proceedings of IEEE Visualization 93 Conference*, G.M. Nielson and D. Bergeron, eds, IEEE Computer Society Press, 1993, p. 102.

[12] J. Chin and P.V. Coveney, *Towards tractable toolkits for the Grid: a plea for lightweight, usable middleware*, http://www.realitygrid.org/lgpaper21.htm.

[13] C. Karl, P. Stephen and P. Jim, *Volker Sander, Usage Scenarios for a Grid Resource Allocation Agreement Protocol*, GGF Memo, February 2003.

[14] S. Floyd, *High Speed TCP for Large Congestion Windows*, IETF Internet Drafts, http://www.ietf.org/internet-drafts/draft-ietftsvwg-highspeed-01.txt.

[15] T. Kelly, *Scalable TCP: Improving Performance in High Speed Wide Area Networks*, Proceedings of PFLDNet-1, Feb. 2003.

[16] J Harting, M.J. Harvey, J. Chin and P.V. Coveney, Detection and tracking of defects in the gyroid mesophase, *Comp. Phys. Commun.* **165**(2) (15 January, 2005), 97–109.

[17] P.V. Coveney and M.J. Harvey, Europe: AG aids RealityGrid and the TeraGyroid Project, *AG Focus* **2**(1) (2004), 4.