
Mathematics and e-Science

Report on a meeting held at the London Mathematical Society, 18 March 2003

Ursula Martin, Queen Mary University of London
Chair, LMS Computer Science Committee

Mike Dewar, NAG Ltd

Philippa Gardner, ICSTM

Nicholas Higham, University of Manchester

Ian Roulstone, Met Office

Iain Stewart, University of Durham

This report presents the outcomes of an informal discussion meeting of mathematicians and theoretical computer scientists, organised by the Computer Science Committee of the London Mathematical Society, the UK learned society for mathematics.

E-Science offers a vision of how the scientist and engineer can generate, analyse, share and discuss insights, data, experiments and results, enabled by a computing infrastructure commonly called the Grid. This vision of a globally connected community has broader application than science, with the same technologies being used to support e-Commerce and e-Government. E-Science has stimulated a challenging research agenda for building a future e-Science infrastructure and understanding how best to exploit it.

Mathematics is at the heart of representing and reasoning about scientific and engineering data and knowledge, and the role of mathematics in e-Science is potentially profound:

- mathematics as the language of science underpins almost all scientific and business applications of the Grid. Hence new ways of doing **mathematics enabled by the Grid**, and the identification of how the Grid can best handle mathematical computation and data, whether numeric or symbolic, have the potential to impact mathematics itself, and mathematical modelling, such as epidemiology or weather forecasting.
- The vision of the semantic Grid conceives e-Science as a set of core knowledge services which allow the user seamless access to multiple streams of data and computation. Such **mathematical Grid services** will require descriptions of problems and services as ontologies in languages such as MathML/OpenMath, and techniques for giving the user increased assurance of results thus obtained.
- mathematics, in particular the discrete mathematics and logic which underpin computer science, provides the tool for **understanding and modelling the Grid** itself, for example new techniques for resource allocation and modelling, handling and mining data or modelling network infrastructure. Logical techniques based on applied semantics and computational logic provide profound new ways of understanding resource management, distributed data, space and mobility.

The rest of this report gives more detail, and identifies e-Science research opportunities in: mathematical modelling, scientific computation, numerical mathematics, analysis, linear algebra, inverse methods, control theory, variational methods, symbolic computation, computational logic, discrete mathematics, graph theory, operations research, economic mathematics, stochastic analysis, algorithms and applied semantics.

1 Computational mathematics and e-Science

1.1 New challenges in mathematical modelling

E-Science offers massive computational power, which by offering finer resolution models and the ability to access massive distributed datasets is able to transform many applications. The scope of e-Science is very wide: present and future developments in science and engineering will generate a variety of codes and datasets with mathematical content: from protein databases to engineering designs to inputs from massive networks of environmental sensors. As experience develops, fundamental mathematical challenges are emerging which are common to various e-Science testbed projects whose focus is on services within a particular application domain -- such as Gravitational Waves (GridOneD), BioInformatics (myGrid), engineering optimisation/design (Geodise), medical image analysis (eDiamond) or weather forecasting (Climateprediction.net).

The integration of multiple evolving distributed data sources, for applications such as epidemiology, weather forecasting or marine ecosystem modelling, offers challenges that are fundamentally new. For example it is predicted that new satellite data sources will triple the amount of data available for weather forecasting over the next few years. The challenges include data reduction, that is computationally efficient ways of turning raw observations into appropriate model variables, and data assimilation, in other words handling the input of massive and heterogeneous data into predictive models in a consistent and stable manner. This raises new challenges in inverse methods, control theory, variational methods and Bayesian statistics. There is an increasing need for the academic community to become involved in the research and development of models being developed in operational centres such as the Met Office and CLCRC.

1.2 Numerical computation

Scientific computation continues to be dominated by problems for which there is no convenient "exact solution" and for which numerical methods must be applied in floating point arithmetic. The cornerstones of scientific computing are numerical solution of differential equations and - often as a subproblem of the latter - numerical solution of algebraic eigenvalue problems and linear systems of equations. Efficient, reliable solution of these problems is a prerequisite for e-Scientists working in fundamental application areas such as structural mechanics, fluid dynamics and electromagnetism. Key numerical challenges for e-Science include estimating the accuracy of solutions and then refining them if necessary, developing Grid-enabled solvers, and developing helper services that automatically choose solvers and their parameters (e.g., convergence tolerances and preconditioners) given just the user's problem specification.

In differential equations, a fundamental challenge is the development of reliable and efficient self-adaptive refinement techniques for spatial and temporal discretization. The key issue is computation of tight bounds on the discretization error, without solving an infinite dimensional problem, so as to give a "confidence interval" on the accuracy of the computed solution. Achieving this goal is currently possible for ODE problems and linear elliptic PDEs, but to make progress in the case of nonlinear PDEs will require a concerted effort across the entire applied mathematics community. A generic methodology will require tools from asymptotics, reduced order modelling, and will build on hierarchical approximation techniques such as wavelets and radial basis algorithms. The need to underpin this research effort by rigorous analysis represents a considerable challenge to pure and applied mathematics within the UK.

In linear algebra, error estimation and refinement are well established for dense problems, but need much more development for the sparse problems that typically arise in very large-scale Grid applications. In particular, establishing procedures consistent with the overarching nonlinear problem is crucial. In both linear algebra and differential equations much work is needed on Grid-enabled solvers, with first-generation meta-systems such as NetSolve for linear algebra, and NEOS for nonlinear optimization, providing ideas on which to build.

1.3 Symbolic computation

Scientific computation using exact, or symbolic, quantities, is a growth industry, typified by the increasing use of packages such as MAPLE, MATHEMATICA or GAP in both research and education. The use of networked resources such as Maple.net allows algorithms to have access to evolving libraries of solutions and methods. On-line numeric-symbolic differential equation solvers will extend the techniques above with symbolic methods based on symmetries of differential equations, formal (series) solution generators including asymptotic expansions, and the Risch algorithm. Groebner basis concepts for the solution and analysis of polynomial systems have now been extended to systems of differential equations with many potential applications. Symmetries have also been exploited in the solution of constraints, with a wide variety of applications in scheduling and the like, and in the theory of moving frames with a whole new sphere of applications to computer vision, differential geometry and so forth. Moreover, there is now considerable interest in combining symmetry with numerical integration, in the hot new field of "geometric integration". The main bottleneck to wider uptake is the so-called "memory gap": management of distributed memory when CPU speed is not matched by speed of memory access, also a fundamental issue for the Grid, and one whose solution offers great potential in the greater exploitation of symbolic computation.

1.4 Mathematics and the "Semantic Grid"

To move towards the e-Science vision from the present Grid infrastructure requires the development of a Semantic Grid, a conceptual framework for a service-oriented architecture structured around data/computation, information and knowledge layers. This deepens Berners-Lee's vision of the semantic web, where information is given a well-defined meaning so we can apply inference rules to it in a consistent way. Ontologies allow different information sources to share meaning, and software agents to process data on behalf of human users or exchange results with each other. Agents which understand the connections between related ontologies can also act as "translators" mapping a concept in one domain to one in another, thus allowing the web to behave as one large, evolving knowledge base.

The worldwide web consortium (W3C), supported by key players including IBM, Microsoft, Sun and HP, is developing a suite of technologies to support web services and integrate them with the semantic web. The Math WG of W3C and the Openmath Society, supported by market leading software vendors such as NAG Ltd, Wolfram Research Inc. and Waterloo Maple, have an ambitious agenda for leveraging these technologies to build web-based mathematical services to support applications throughout science and engineering. This is being carried forward in Europe by NAG Ltd, under the EU MONET initiative, ensuring that UK and European strategic interests are addressed at a global level by engagement with the development and exploitation of

standards such as the Open Grid Service Architecture from the Global Grid Forum, and the Web Ontology Language (OWL) from W3C.

A number of experimental services are under development: numerical algorithms at NAG in Oxford, closed-form integration at the University of Bath and symbolic solutions of differential equations at INRIA in Nice. Making a wide variety of mathematical services available on the Grid will help application developers build software solutions rapidly and effectively. Software agents, based on computational logic techniques for reasoning with ontologies and service descriptions, will help end-users select the most appropriate algorithms for their problem and interpret their results.

1.5 Security and Trust

A fundamental assumption of the Semantic Grid is that users will use external Grid services rather than hand-crafting their own codes. This will require that the service provides them with sufficient information and assurance that they trust the results it produces. Such assurance is particularly important where IP issues prevent disclosure of the source codes or the methods used: in a different domain it has already proved a barrier to the use of third-party components in chip design. This assurance may take the form of the return of numeric results with the guarantees sketched above, or with more general arguments that a remote service has used the correct method: techniques of computational logic allow a much more general symbolic analysis of arguments of this kind. For example, recent work in collaboration with Qinetiq concerned the nature of such assurance arguments in avionics design, brokered between a “design service”, control engineering via analysis and numerical simulation in the continuous domain, and an “implementation service”, code generation for in the discrete domain.

This complements more general techniques for ensuring safe code, which consider both the communications protocols between entities, and the codes that they communicate. Model-checkers and symbolic techniques based on logical proof, bisimulation or game semantics are used to verify or falsify security protocols, and safe languages with robust typing and static analysis protect against buffer overruns. In the proof-carrying code project, programs are distributed with proofs of their type safety, which can be checked once they reach their target destination. The next step is to provide formal policies that also guarantee resource-bound and access control properties of certified code, plus certifying compilers that can propagate such source properties and evidence for their validity to low-level code. Such a project has recently been funded by NSF as part of the USA-focus on the Grid.

2 Modelling the Grid

The Grid problem has been described as one of providing “flexible, coordinated resource sharing among dynamic collections of individuals, institutions and resources”. As such, it is a distributed computing problem, with new twists arising from its focus on developing common communication and computational infrastructure. The UK has a strong international presence in these research areas, but we stress that the selection is by no means exhaustive. A more comprehensive account can be found in the forthcoming documentation associated with the UK Computing Grand Challenge on the “Science for Ubiquitous Computing. The current EU Global Computing initiative is also relevant.

2.1 Space and Mobility

It is now accepted that **boundaries** (whether physical or virtual) and **movement** of components (whether data, processes or devices) are properties to be modelled directly

in understanding widely distributed systems. Calculi and logics are now emerging for modelling, for example, boundary access, process and boundary mobility, and message and host failure. Such concepts are fundamental to the notion of virtual organisation which underpins the Grid. Within a few years we can expect these calculi and logics, together with associated programming languages, verification and performance tools, to be rigorously assessed in experimental systems such as a sentient building, business-to-business transactions and scientific databases. E-home, e-Business and e-Science provide an invaluable source of such examples, which would inevitably reveal further challenges to the fundamental theory.

2.2 Resource Management

Mobile agents must be able to acquire necessary resources from the environments that they visit. Methods based upon logics and types now exist to **describe** allocation and deallocation of resources, such as memory space and CPU timeslices, in a dynamically changing environment where hosts (as well as users) can join or leave the system at unpredictable times; but methods to **control** resource usage and to **analyse** resource descriptions are in their infancy. The importance of resource specification has already been acknowledged in the Grid community. For example, the Globus toolkit includes a Resource Specification Language, used to communicate requests and replies between different components of the resource management architecture. No canonical resource specification language has emerged from the Grid community, so logical studies could have significant impact on future Grid infrastructures. Equally, the diversity of resource on the Grid provides a stern scientific challenge to resource logics: which have not yet been applied to the resource management problems of distributed systems.

2.3 Resource allocation and modelling

Consider a computational Grid in which resources can be provided to users, possibly at a cost to be agreed by a user and a resource provider. In order to allocate resources to users, a distributed allocation mechanism must be employed. This should be able to: schedule jobs amongst chosen resources (respecting priorities, resource availability and quality of service demands); ascertain the sorts of services offered by resource providers and broker the allocation of a resource to a user (whilst maintaining a market which behaves in a fashion acceptable to all participants who may have very different personal strategies and goals); and adapt to changes in user demands and fluctuations in resource provision (whilst maintaining global operation respecting the goals and demands of the participants). All of these tasks must be undertaken in a computationally efficient manner where the computational costs include time, memory and communication bandwidth: the problem provides a rich field for applications of mathematical economics, network modelling and probability.

2.4 Algorithms for handling and mining data

E-Science enables data-gathering on a scale completely unprecedented in the history of computing (and, indeed, in the history of science). It is essential that the processing and handling of this data is done efficiently, and a number of algorithmic challenges arise in this context. How can we apply or modify the theory and practice of algorithms for storing data on distributed and hierarchical memory, which has been very influential for storing medium-scale data sets such as GIS, to the much larger and much more heterogeneous Grid? Which kinds of data compression techniques are appropriate for scientific data, and what are the theoretical and algorithmic foundations of such

techniques? How can synopsis data structures and algorithms on data streams facilitate the warehousing of scientific data? A wide variety of large-scale algorithmic data-mining techniques have been developed (and successfully applied) in the context of web searching. Can similar techniques be applied to scientific data? If not, what are the algorithmic foundations of mining scientific data? What are the algorithmic foundations of visualising large-scale scientific data sets?

2.5 Distributed data

Semi-structured data, such as XML, plays a fundamental role in the exchange of information between globally distributed applications. Which data structuring techniques are needed to support the efficient processing of scientific data held in such semi-structured formats? What are the underlying principles behind efficient query evaluation for such semi-structured formats? And extending the previous questions: how can we answer complex queries on (highly) compressed representations of scientific data organised in semi-structured formats without uncompressing them?

The analysis techniques and languages associated with such data may be drastically different from conventional ones, and emphasise the significant potential for collaboration between researchers in semi-structured databases, applied semantics, algorithms and Grid services.

For example: declarative query languages with smooth integration between data model, logic and language, as found for the relational model; pattern-matching languages arising from the ML community, with powerful types to analyse the structure of data; and mobile query languages, incorporating queries with process mobility, such as a hyperlink and a light-weight trusted process for retrieving data associated with the link.

2.6 Modelling network infrastructure

The increasing use of the internet for a diverse range of distributed applications has posed many challenges in the area of mathematical modelling, and growth in the use of the Grid will do the same. One example is that in contrast to traditional telephone networks, the load imposed by end-systems is much less predictable and the resulting service achieved by a user fluctuates as a consequence of the behaviour of other users with entirely separate needs and incentives. This has led to more refined notions of quality of service which differ across users and applications, and the challenge is to find simple and robust mechanisms which enable differential quality of service. This has also led to the need for more accurate forms of measurement and monitoring in very large scale network infrastructures which will allow network providers to make informed management decisions affecting the provisioning of resources over the long term which are also economically efficient with respect to the short term traffic characteristics of the service level agreements that are in place with the users.

The modelling of the internet graph as an evolving random graph has led to new insights and effective algorithm design: there is similar scope in modelling the Grid at both hardware and software levels. The many, often competing or conflicting, visions of future networks increases the importance of insightful mathematical models, to explore concepts and identify fundamental trade-offs. And conversely the increased ability to study such networks, like the Grid, offers an enormous opportunity to study a new and significant emerging mathematical phenomenon.