Grid computing and molecular simulations: the vision of the *e*Minerals project

Martin T Dove[1,2] and Nora H de Leeuw[3]

1. *Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ*
2. *National Institute for Environmental eScience, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0EW*
3. *School of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX*

Abstract

This paper discusses a number of aspects of using grid computing methods in support of molecular simulations, with examples drawn from the *e*Minerals project. A number of components for a useful grid infrastructure are discussed, including the integration of compute and data grids, automatic metadata capture from simulation studies, interoperability of data between simulation codes, management of data and data accessibility, management of jobs and workflow, and tools to support collaboration. Use of a grid infrastructure also brings certain challenges, which are discussed. These include making use of boundless computing resources, the necessary changes, and the need to be able to manage experimentation.

Keywords

Grid computing, escience, virtual organization

Introduction

Molecular simulation scientists have always needed significant computing resources. In the early days, limitations on available computing power led to a number of constraints, such as the size of the system that could be studied, the length of time for which a simulation could be run, the complexity of the forces that could be used, and the number of state points (e.g. temperature) that could be used within a single study. Over the years, computer power available to computational scientists has grown at an exponential rate, both at the desktop and at high-performance computing facilities, greatly increasing the horizons of the simulation scientists. Growth in computing capacity has been matched by developments of new simulation methods and algorithms. Similarly, desktop visualization tools have become widely available to enable the simulation scientist to view the results of a simulation through graphical manipulation tools and through animations.

The development within the past decade of grid computing and escience methods [1–4] (Foster and Kesselman 1998, 2003; Foster et al 2001; Berman et al 2003) offers the prospect of new evolutionary (even revolutionary) developments in molecular simulations. Grid computing was developed with the idea of linking together computing resources to create large-scale computing infrastructures. Some of the early efforts were based on linking together high-performance resources, but the biggest impact is coming from producing larger grids of more modest resources. One example is the approach of linking together hundreds, or even thousands, of under-used desktop machines. Consider the example of forming a grid across a campus, and linking together 1000 under-used PCs. If each PC operates at around 5 Gflops (a quick test shows that the author's laptop is running at 4.3 Gflops) and has 1 GB memory, a quick multiplication shows that the grid resource has 5 Tflops of power with 1 TB memory. This would equate with computers around the 30th place in the world rankings. Of course, a grid of computers does not make a high-performance computer because connectivity between separate nodes will not be fast enough to support parallel applications, there can be no guarantees against loss-of-service for individual nodes, and the machines on the grid are likely not to be of homogeneous hardware and with common operating systems. Less technical, it is also likely that different machines within a computing grid will be subject to different usage policies. Such a system is more appropriate for Monte Carlo simulations, for example, where each node will gather a subset of the total configurations needed in an individual study, or for high-throughput studies where each node performs an independent calculation. An example of the latter could be a simulation study over many temperatures. An example of the use of a grid environment to accumulate statistics and temperatures is a study of diffusion of ions through interfaces by Monte Carlo methods (Calleja and

Dove, 2004)

Grid computing gives the prospect of what has been described as "boundless computing" in that there is a considerable wealth of underused computing power in many institutes. The default configuration for many PCs is equivalent to the supercomputing capabilities of just a few years ago in terms of processing speed and processor memory, and hence these machines are capable of being used as compute engines for molecular simulation studies. The challenge for the molecular simulation scientist is to be able to harness this computer power.

The original vision of grid computing was characterised by the need to link together computing resources, but there is a much wider vision. The challenge of managing simulations performed over a grid resource has to be met through the use of job submission, monitoring and workflow tools. With increased computing resources is also the need for data management methods. Running a large study with grid resources will necessarily lead to the generation of many output files, too many to be able to manage using conventional approaches. The data management challenge is further compounded by the fact that simulations performed on a distributed compute facility will lead to data being stored in several locations.

With the emergence of the concept of grid computing arose the concept of virtual organisations (VOs). This concept arose as an independent idea, in fact without a necessary dependence on IT, but it is a particularly pertinent idea for escience where organisations come together to share computing resources (Dove et al, 2004, 2005). In fact, collaborative working is increasingly becoming more of the norm in terms of funding. In order to maximise the potential of simulation scientists working within a VO, it is essential for them to have access to collaborative tools, such as desktop videoconference and application-sharing tools. Furthermore, the access to data needs to support easy data sharing, and it is essential for data to be stored in a format that can be shared between different applications.

This paper serves as an introduction to a collection of papers written by members of the eMinerals project, which represent an attempt to develop an integrated compute and data grid infrastructure together with support to enable the project team to work as a VO. The rest of this paper gives a broad outline of the vision that drives the project.

The *e*Minerals project

The eMinerals project (formal name, "Environment from the Molecular Level") is one of the escience testbed projects funded by NERC. (Dove et al, 2003) The scientific goal of the project is primarily to use grid-computing methods to facilitate simulations of environmental processes at a molecular level

with increased levels of realism, achieved through being able to use larger samples, more complex systems, or through increased accuracy to better handle changes in chemical bonding. The science applications concern issues such as nuclear waste encapsulation (Pruneda et al, 2005; Trachenko et al, 2005), adsorption of organic pollutants on soil particles, and weathering and precipitation of minerals, as described in the cited references. The wider aim of the project is to develop a cross-institute collaborative infrastructure that can be used by the scientists to accomplish the scientific objectives.

Based on the discussion in the introduction, the *e*Minerals project team has needed to take a three-track approach, based on the science to be accomplished (including the development of the simulation codes), the grid infrastructure that has been needed to support the science, and the virtual organization tools required to facilitate the team working together. The project team consists of a mixture of scientists, code developers, computer scientists and grid specialists. To some extent the project has been an experiment in enabling such a diverse team to work together towards a common objective.

The simulation work in the *e*Minerals project involves a wide range of methodologies. These include both models with classical empirical potentials and quantum mechanics, the latter employing both plane wave and localized basis sets. The quantum mechanical methods include both density functional and Quantum Monte Carlo approaches. Both methods are used for static energy and lattice dynamics methods (zero-temperature methods) and molecular dynamics methods (for non-zero temperatures). With respect to the molecular dynamics method, one of our challenges has been to develop a molecular dynamics code that can handle the necessarily large samples required for radiation damage studies in systems with long-range electrostatic interactions. This has led to the development of the DL_POLY3 molecular dynamics code as part of the work of the *e*Minerals project. One interesting aspect of the work of the eMinerals project has been to enable code developers to work with the end users (i.e. the scientists) and to take advantage of a heterogeneous computing environment within which to test the codes.

In addition to enabling scientists, code developers and grid specialists to work together in ways that have not happened before, the eMinerals team also recognizes the capacity within many team members to multitask outside their notional area of expertise. Thus, for example, some of the scientists have become members of the grid development team, other scientists have joined with grid specialists to take a leadership role in developing XML applications,

The vision of the eMinerals project

The broad vision is to develop an infrastructure that allows scientists to do the science they want to do,

free of the limitations of managing resources and data, free of the need to learn new tricks, free of the need to convert data between different formats, free to collaborate, free to share data, and free to share resources. This vision can be broken down into several components.

1. *Compute grid infrastructure*. Shared computing resources, including the use of resources that may otherwise be barely used, can be combined to produce an infrastructure of some considerable computing power. In fact, given the large amount on untapped desktop resources that can be found in a typical institute, there is the prospect of collaborations having access to what has been described as "boundless computing resources". As noted below, the possibility of providing boundless resources to simulation scientists will provide new challenges.
2. *Integrated data grid infrastructure*. Computing and data go hand in hand, and the vision for the eMinerals project has at its core the integration of computing resources and data management. When jobs are run on distributed computing resources, the data generated will need to be managed in ways that hide the distribution from the user. Moreover, partners within any collaboration will need to be able to access data without needing to be told specific details of the location of files. The vision for the eMinerals project includes the need to provide an infrastructure to support collaborative and grid data management, and also, as discussed in the following points, enable data to be understood by partners and by the codes run by partners.
3. *Automatic metadata capture*. Files of data require information about the data, such as when the data were obtained, from which simulation program, the person who ran the simulation, and the computer that the simulations were run on. The file system will provide a date and file owner, but that information may be corrupted through moving the data file between file systems. Ideally simulation programs will provide headers within the data files to provide information about the program (such as version number), but frequently these headers will not be propagated into all output files. Some of this information may be stored in written documentation by the scientist, but ideally such information should be stored with the data, not in a physically different medium. One type of metadata that is not built into many systems concerns the context of the data, covering issues such as the motivation for collecting the data (for example, is the file a result of a production or test run, and is it part of a larger series of data files, and in any case why did the scientist want to run this particular simulation), and the quality of the data (is this the best simulation that was possible at the time, or do the results look suspect). Again, some of this information may be captured in a notebook, but the vision is for such metadata to be capture automatically, whether at the time of the submission of the job or when the data are analysed, and kept with the data. The

approach within the eMinerals project is to build this in to the portal infrastructure currently being developed.

4. *Interoperability*. One constant problem is that programs generate data in formats that prevent easy re-use of the data in other programs. Although there are many instances where data can simply be parsed from one format to another, it is also possible that the formats of files contain exceptions (such as when programs throw out a helpful line of information when a certain condition has been encountered). Thus central to the vision is the need to be able to handle data in a general form that enables the data to be understood by other programs. Implicit in this is the need for data to be self-described, which links to the use of metadata. For many of our applications we are finding that XML provides the functionality and interoperability we require. XML files can readily be converted into a form that can be view via a standard web browser, including assembling data into graphical form. The vision for the eMinerals project encompasses support for the simulation scientists to obtain an immediate view of the results of simulation studies, including studies in which many separate jobs are run, without the tedious process of pasting numbers into separate data analysis programs.
5. *Management of data and data accessibility.* This point follows from the previous points. Collaborators need to be able to share data easily. Appropriate use of metadata and data markup will make the experience of handling shared data easier, but on top of this is the need for a data archiving infrastructure that enables rapid access to the data. The same metadata should be used to enable collaborating scientists to search through the archives in order to obtain exactly the sets of data (which may involve many files) that are required.
6. *Management of jobs, including workflow*. Simulation scientists tend to manage their tasks manually or through bespoke scripts. A typical study will involve a number of discrete stages, including setting up the files, running the jobs, monitoring their progress and intervening if necessary, gathering together the output data, extracting the key numbers, performing subsequent analysis and data presentation tasks, and archiving the data generated in the study. This frequently involves the scientist in carrying out a number of tedious tasks, and the management of complex workflow patterns may often be the bottleneck in the scientist's productivity. Many of the difficulties in handling the job management and workflow tasks can be solved using escience methods.
7. *Collaborative infrastructure.* eScience is beginning to provide the tools to enable collaborators based in geographically distributed locations to be able to work together as if sharing adjacent offices within a single building. Among the tools being developed are desktop audiovisual

conferencing tools and tools for sharing applications. Ideally such tools should not have significant overheads for the user, and should require no more effort than that required to knock on the door of the person in the next office along the corridor. The vision is to be able to reproduce the experience of geographical proximity for distributed collaborating scientists. This goes hand in hand with the tools for sharing resources and data.

The challenge of the vision

The main problem with the vision outlined above is that it will require significant changes to the way that scientists carry out their work. In particular we identify several challenges that face individual scientists as well as the project team.

1. *Challenge of access to boundless computing*. At face value, this challenge is counterintuitive, since boundless computing is every simulation scientist's dream. However, most scientists' working practices mitigate against making use of boundless computing. Scientists often prefer to handle one study at a time, and to develop the study incrementally through running simulations in order. Access to boundless computing enables scientists to launch many more simulations that they could normally manage, which requires several components of the vision outlined in the previous section. What access to boundless computing gives the simulation scientist is the ability to run several projects in parallel, to run many speculative simulations, to obtain many more data points than previously possible, and to gain many more data for statistical analysis. The tools enable the simulation scientist to think ahead and launch jobs for a study to which he/she would like to devote time only at some point in the future.
2. *Challenge of change*. It is a fact of life that most of us get used to a certain way of doing things, and find it hard to make the time required to change to a new way of working. The vision of escience for simulation science is a significant change to the normal way that most simulation scientists will operate, and there is some effort required to learn the new ways of doing things. It is not just a case of learning new techniques; it must not be underestimated that the challenge is to the very way in which simulation scientists normally operate.
3. *Challenge of experimentation*. Grid computing is a long way from maturity. Many of the components required to properly implement the vision are simply not yet available in a robust form. Thus scientists engaged in a grid computing environment at the present time need to understand that they will most likely suffer some level of inconvenience through tools not quite doing what they are required to do!

Summary

The vision we have outlined above and in this collection of papers has been developed specifically for a collaboration between molecular simulation scientists. There are several areas in which the vision can be developed, but the challenges have yet to be explored. To mention two, testing the scalability for larger virtual organisations, and inclusion of experimental data. At the time of writing, the vision of the eMinerals project is still being implemented, and much progress is still awaited. The collection of papers following this introduction provides a snapshot of some of the technological developments that have been made and implemented, and gives an impression of the science that is being carried out by the *e*Minerals project team.

Acknowledgement

We are grateful to NERC for financial support.

References

[1] Foster I, and Kesselman C, ed (1998) The Grid: Blueprint for a New Computing Infrastructure, 1st edition (Morgan-Kaufman)

[2] Foster I, and Kesselman C, ed (2003) The Grid: Blueprint for a New Computing Infrastructure, 2nd edition (Morgan-Kaufman)

[3] Foster I, Kesselman C and Tuecke S (2001) The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal or Supercomputer Applications, **15**, 200–222

[4] Berman F, Hey AJG and Fox G, ed (2003) *Grid Computing: Making The Global Infrastructure a Reality*, (John Wiley),

[5] M Alfredsson, JP Brodholt, PB Wilson, GD Price, F Corà, M Calleja, R Bruin, LJ Blanshard and RP Tyer (2005) Structural and magnetic phase transitions in simple oxides using hybrid functionals. Molecular Simulations **XX**, 1234–5678

[6] M Calleja, R Bruin, MG Tucker, MT Dove, RP Tyer, L Blanshard, K Kleese van Dam, RJ Allan, C Chapman, W Emmerich, PB Wilson, J Brodholt, A Thandavan, VN Alexandrov (2005) Collaborative grid infrastructure for molecular simulations: The *e*Minerals minigrid as a prototype integrated compute and data grid. Molecular Simulations **XX**, 1234–5678

[7] M Calleja and MT Dove (2004)

[8] C Chapman, J Wakelin, E Artacho, MT Dove, M Calleja, R Bruin and W Emmerich (2005) Workflow issues in atomistic simulations. Molecular Simulations **XX**, 1234–5678

[9] Z Du, NH. de Leeuw, R Grau-Crespo, PB Wilson, JP Brodholt, M Calleja and MT Dove (2005) A computational study of the effect of Li-K solid solutions on the structures and stabilities of layered silicate materials – an application of the use of Condor pools in molecular simulation. Molecular Simulations **XX**, 1234–5678

[10] MV Fernández-Serra, G Ferlat and E Artacho (2005) Two exchange-correlation functionals compared for first-principles liquid water. Molecular Simulations **XX**, 1234–5678

[11] A Marmier, H Spohr, DJ Cooke, S Kerisit, J Brodholt, PB Wilson and SC Parker (2005) Self diffusion of argon in flexible, single wall, carbon nanotubes. Molecular Simulations **XX**, 1234–5678

[12] JM Pruneda, L Le Polles, I Farnan, K Trachenko, MT Dove and E Artacho (2005) Calculation of the effect of intrinsic point defects and volume swelling in the nuclear magnetic resonance spectra of $ZrSiO_4$. Molecular Simulations **XX**, 1234–5678

[13] K Trachenko, MT Dove, EKH Salje, I Todorov, William Smith, M Pruneda and E Artacho (2005) Radiation damage in the bulk and at the surface. Molecular Simulations **XX**, 1234–5678

[14] J Wakelin, P Murray-Rust, S Tyrrell, Y Zhang, HS Rzepa and A García (2005) CML tools and information flow in atomic scale simulations. Molecular Simulations **XX**, 1234–5678

[15] S Wells, D Alfe, L Blanchard, JP Brodholt, M Calleja, CRA Catlow, GD Price, R Tyer and K Wright (2004) *Ab initio* simulations of magnetic iron sulphides. Molecular Simulations **XX**, 1234–5678