eScience for molecular-scale simulations and the eMinerals project


1 Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK
2 Cambridge eScience Centre, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, UK
3 Daresbury Laboratory, Daresbury Science and Innovation Campus, Science and Technology Facilities Council, Warrington WA4 4AD, UK
4 Department of Chemistry, University of Bath, Bath BA2 7AY, UK
5 School of Engineering, Computer Science and Mathematics, University of Exeter, North Park Road, Exeter EX4 4QF, UK

We review the work carried out within the eMinerals project to develop eScience solutions that facilitate a new generation of molecular-scale simulation work. Technological developments include integration of compute and data systems, developing of collaborative frameworks and new researcher-friendly tools for grid job submission, XML data representation, information delivery, metadata harvesting and metadata management. A number of diverse science applications will illustrate how these tools are being used for large parameter-sweep studies, an emerging type of study for which the integration of computing, data and collaboration is essential.

Keywords: eScience; eMinerals; molecular simulations

1. Introduction and background

(a) Molecular-scale environmental processes

Although we usually think of environmental science as spanning length scales from the centimetre to the global, and time scales from the second through to the mega-annum, behind many important phenomena are molecular-scale processes that occur over length scales of $10^{-10}$–$10^{-8}$ m and times of the order of $10^{-14}$–$10^{-10}$ s. Examples of environmental science include the adsorption of pollutants within soils, behaviour and properties of rock-forming minerals under extreme
conditions and the weathering of minerals. Technological applications for the environment include materials for nuclear waste encapsulation and materials with properties that can be tailored for applications in environmental sensors. Modern simulation technologies now enable these to be studied using atom-based simulation methods.

The eMinerals project has sought to apply grid computing and eScience techniques to support atomistic simulation studies of environmental processes and materials, and this paper presents a high-level overview of both the technological developments and some of the science outcomes. Following this introduction, in which we discuss some of the general ideas, we devote §2 to describing some of the eScience tools developed within the eMinerals project, and §3 to outlining some exemplar simulation studies that have been enabled by some of the technological developments. In some cases, the science outcomes have already been the subject of prior publication, and thus the focus will be more on the role of eScience than on the science per se. We also note that this collection includes three short papers (Frame et al. 2009; Walker et al. 2009; White et al. 2009) that describe some of the technological aspects in more detail.

(b) Atomistic simulation methods

In our work, we make use of two main types of atomic-scale simulation. The first is where the interactions between atoms are determined using quantum mechanical methods, with the main focus on solving the Schrödinger equation for the electrons. The second approach is to use empirical representations of the interactions between atoms with parameters that are determined from independent quantum mechanical calculations or tuned to fit certain experimental data. Both approaches have relative strengths and weaknesses, which can be broadly summarized by stating that quantum mechanical approaches are more accurate, but this accuracy is only gained at a cost in terms of computational requirements such that there are applications for which empirical methods will be more appropriate.

In modern applications to systems larger than individual molecules, quantum mechanical calculations are carried out using a formulation called ‘density functional theory’ (Payne et al. 1992), in which the basic equations are cast in terms of the electron density rather than the electronic wave functions. The most practical approaches are those in which the electronic wave functions are expanded in terms of basis sets. From the variational principle, the best wave functions for any configuration of atoms are those that give the lowest energy. We use two implementations of this approach as encapsulated in two main codes, CASTEP (Segall et al. 2002) and SIESTA (Soler et al. 2002). The primary difference between these two codes is in how the electronic basis sets are represented: in CASTEP, the wave functions are constructed as a Fourier combination of plane waves, and, in SIESTA, they are represented using atomic orbitals. With its more general representation, CASTEP may prove more accurate, but the approach of using atomic orbitals enables calculations to be run more efficiently. Thus, SIESTA is the only practical option for simulations of large systems (of the order of a few hundred atoms) using quantum mechanical methods. Quantum mechanical simulations are challenging even for the simplest...
systems, but for environmental applications, we need to run simulations with many more atoms than the norm, and repeatedly so for studies with whole families of pollutant molecules.

Empirical methods typically use functions of the following form to represent the interaction energy of two atoms labelled $i$ and $j$ at separation $r$:

$$
\varphi(r) = \frac{Q_i Q_j}{4\pi\epsilon_0 r_{ij}} - A_{ij} r_{ij}^{-6} + B_{ij} \exp\left(-r_{ij}/\rho_{ij}\right),
$$

where $Q$ represents the ionic charge, which can be fixed at the formal value, tuned to agree with quantum mechanical calculations, or tuned in an overall fit to experimental data, and $A$, $B$ and $\rho$ are tunable parameters that can be adjusted in the same way as the charge. The first term in this equation is the standard Coulomb interaction, the second term is the lowest order term in the expansion of the dispersive interaction and the last term is known to be a good representation of the repulsive interaction at short distances. More complex functions can be used where demanded by the physics or chemistry, for example to account for atomic polarizability or the effects of covalency. These simple representations of the interatomic interactions enable the simulation of systems as large as a few million atoms, which is several orders of magnitude larger than possible with quantum mechanical methods. Alternatively, when using smaller systems they allow a turnaround time for the calculations that is several orders of magnitude faster than possible with quantum mechanical calculations.

In this paper, we describe the results obtained with two techniques that work with either quantum mechanical or empirical methods. The first technique is lattice energy relaxation, in which the positions of the atoms are adjusted until there are no residual forces. This method gives the configuration of atoms with the lowest overall potential energy. The second technique is molecular dynamics simulation, in which the atoms are moved in a long series of time steps following Newtonian equations of motion, essentially following force = mass $\times$ acceleration, where the force on each is calculated from the interatomic interactions, and the acceleration can be converted into a time-step change in atomic positions and velocities. The molecular dynamics technique can be computationally challenging even when using empirical methods to represent the interatomic interactions. For example, in our work on understanding the damage caused by radioactive decay on potential encapsulation materials, we need to perform simulations containing several millions of atoms. As part of the eMinerals project, we specifically developed the internationally used DL_POLY_3 code (Trachenko et al. 2005a; Todorov et al. 2006) for this work.

(c) eScience for molecular-scale simulations

Grid computing was one of the early areas of focus of eScience. Whereas one of the early ideas was to link together large high-performance computers, the massive growth in the power of commodity computers has led to the development of grids containing very large numbers of independent processors. In 2003, the eMinerals project demonstrated the first UK campus grid at University College London with almost 1000 separate processors—mostly desktop personal computers using the Microsoft Windows operating system—based on the CONDOR
Since then, a number of other universities have developed campus grids using either desktop computers (mostly pools of computers in teaching classrooms) or dedicated processors. An example is CamGrid, a grid comprising both clusters and desktop computers in several departments within the University of Cambridge (Calleja et al. 2004). The National Grid Service (NGS) and the Northwest Grid (NWG) are examples of distributed grids of dedicated processors. Interestingly, both these facilities provide the capability to run parallel jobs across many processors as well as on single processors; these are ideal grid systems for the use of quantum mechanical methods, where each simulation may require approximately 16 processors per job, a number that is considered too small for high-performance facilities.

One of the key findings of a number of eScience projects, including the eMinerals project, is that one of the major applications of grid computing is to enable researchers to run much larger numbers of individual simulations as part of a single study. In climate and Earth systems studies, this might mean running ensembles with slightly different conditions (e.g. Lenton et al. 2007; Massey et al. 2007). In the domain of molecular simulations, the corresponding application is what we call ‘parameter-sweep’ studies, in which many simulations are performed for a wide range of input parameters such as the values of temperature or pressure. To give an idea of how grid computing can change the way in which computational scientists work, we remark that, in the past, researchers would only have been able to perform parameter-sweep studies involving a few tens of state points at most, and would populate the dataset in a linear point-by-point manner, with jobs constructed and submitted by hand in a manner informed by the prior results. By contrast, eScience tools now enable us to submit several hundreds of jobs or more in a single instance, and some of the examples described in this paper take advantage of this to enable more subtle effects in the data to be studied. However, the increase in available computing power provided by grid systems is only one part of the story. Our capacity to undertake increasingly complex studies leads to the significant challenge of handling the huge quantities of data generated; the number of files is now far beyond what the individual researcher can manage, analyse and document without automated tools. Furthermore, this deluge of data militates...
against scientists being able to share their data with collaborators, which is a problem in the era where increasingly large-scale science is carried out within collaborations (Dove et al. 2007).

We now view eScience as the integration of computing, data and collaboration, as illustrated in figure 1. In our opinion, it is this integration that marks out eScience as a qualitatively new approach to science. Furthermore, if there is one application for which this integration will rapidly give a significant step change—the so-called ‘killer app’—it is in the enabling of much larger parameter-sweep, combinatorial or ensemble studies, which are essential for the next generation of environmental applications of atomistic simulations. Without this integration of computing, data and collaboration, progress will increasingly be limited by human rather than machine capabilities. The default view of many providers of computing resources is that data management is the private responsibility of the scientists, who are left to develop their own bespoke management systems. With huge numbers of files, the task of managing data securely, consistently and accurately within this scenario will require considerable effort on the part of the individual researcher, and often a bespoke scheme will not be easy for another researcher (including the team leader) to understand. eScience now enables us to break out of this conundrum, and the work of the eMinerals project has focused on developing practical and general schemes for this integration of computing, data and collaboration. Clearly, the eMinerals project is not the only effort in this regard, but our approach is somewhat different from other efforts (e.g. Coveney et al. 2007; Lenton et al. 2007), in that we have developed systems for the use case in which scientists like to retain control over the simulation codes they use, and wish to use standard unix shells as their working environments (Dove et al. 2007) rather than portals or other interfaces.

2. eScience developments within the eMinerals project

(a) Grid computing and grid job submission tools

The challenge we faced was to develop tools that made submission of many jobs to a grid-computing environment easy for researchers while not sacrificing the available power of the underlying middleware infrastructure. We use the GLOBUS TOOLKIT, with user authentication handled by X.509 certificates. However, GLOBUS commands are not easy for researchers to work with, and so we originally developed a grid job submission tool called my_condor_submit (MCS; Calleja et al. 2005; Bruin et al. 2008), which uses the CONDOR_G interface to GLOBUS and CONDOR’s DAGMAN workflow system. This tool required the user to supply a simple script that closely resembles the scripts used with the CONDOR system, and which our experience showed users were more comfortable with. Not only does MCS make submission of jobs to compute grids much easier, but it also enabled a seamless integration with the use of data grids (§2b), included a powerful metadata harvesting facility (§2d) and provided transparent metascheduling between available compute grid systems. More details on the use of MCS are given elsewhere (e.g. Dove et al. 2007; Bruin et al. 2008).

Not only are GLOBUS tools hard for many researchers to work with, but installing GLOBUS on a user’s own desktop or laptop computers can be problematic. In order to enable researchers to use MCS without a personal installation of GLOBUS, we
have developed a Web service wrapping of MCS called remote MCS (RMCS; Dove et al. 2007; Walker et al. 2009). This enables researchers to use a lightweight set of tools with easy installation (merely copying files onto their computer) that enable them to submit and manage jobs using simple shell commands (such as \texttt{rmcs\_submit}). There is also a Java graphical interface to RMCS that provides the same functionality as the shell commands. Additional tools built on top of the RMCS client tools enable hundreds of jobs that form a parameter-sweep study to be constructed and submitted, with all data files uploaded to the data grid (see below), with a single command. RMCS has been demonstrated to work for a range of different grid systems, including the eMinerals grids, the NGS and the NWG (Thomas et al. 2007); RMCS is now installed as a general user tool on both of these last two facilities.

(b) Data grids

One important component of the RMCS system is the integration of compute and data grids. The work process is for the researcher to upload the data and executable files to the data grid, for RMCS to then download these files to the automatically selected compute grid resource, for the job to run, for metadata to be harvested and for the output files to be archived in the data grid. The researcher can then view the output files directly from the data grid.

This work process has a number of advantages. For the researcher, it means that, with a large parameter-sweep study, the input and output files are automatically archived in a form that is both accurate and reasonably protected. This is much better than the normal process of output files being sent directly to the researcher’s computer for subsequent manual archiving. The use of a data grid also automatically facilitates sharing of data between collaborators with no further action required by the researcher who generated the data. We found that the use of the data grid also makes the essential tasks of data staging for jobs on remote systems somewhat simpler. The main data grid tool used by the eMinerals project was the San Diego Storage Resource Broker (SRB) for which we developed a Web interface that met the primary requirements of simulation scientists (White et al. 2006a). RMCS also works with data grids based on the use of WebDAV technologies.

(c) Data representation and information delivery

Many code developers tend to design the format of output files in a way that requires users to be acquainted with the code and to have a manual at hand. This might once have been acceptable, but it does not lend itself to collaboration (should collaborators be expected to understand such output files?), it does not lend itself to automatic extraction of information and it often does not provide internal important documentation (i.e. it is not a self-describing representation of the data).

Underpinning our attempts to tackle these problems is our use of XML for representing the data generated by our atomistic simulation codes. Specifically, we use the Chemical Markup Language (CML; White et al. 2006b, 2009) in which, briefly, every data item is tagged with a name, dictionary reference, definition of data type, the data value (scalar or array, text or numeric) and the associated units definition.
Most of our simulation codes are written in Fortran, and we have developed
the FoX library to make writing XML from Fortran an easy task for
programmers (White et al. 2006b, 2009). A number of mainstream atomistic
simulation codes, including the aforementioned CASTEP, SIESTA and
DL_POLY_3 codes, as well as a number of less-commonly used atomistic
simulation codes, now routinely generate CML output files. FoX is finding
widespread usage beyond the domain of atomistic simulation, with modules
written for general XML, and now also with specific modules for KML used for
applications with Google Earth (Chiang et al. 2007; White et al. 2009) and
Hydrology Markup Language.

We have built a number of tools and capabilities on top of our use of XML,
which enable researchers and their collaborators to easily understand the
information contained within the data files (Frame et al. 2009). Our SRB
interface (White et al. 2006a) will automatically convert CML files into XHTML
for viewing in a Web browser, with tables and graphs generated on the fly, and
with all data items marked with a dictionary definition to aid understanding.
RMCS is able to harvest metadata directly from the output XML files using the
AGENTX tool (Tyer et al. 2007). We also have tools to enable researchers to easily
collate specific data items from many XML files generated within parameter-
sweep studies. For example, in the case where a researcher has performed several
hundred simulations at different temperatures, these tools will quickly and
accurately construct a table of temperature values and other associated data
items in a form suitable for plotting. Moreover, it is no longer necessary for
researchers to understand the formats of the output files because that task is
handled by these tools. It should be clear that this approach makes it much easier
for collaborators to use the data generated by their colleagues.

(d) Metadata

We have developed a new approach to tackling the ubiquitous problem of
attaching metadata to data, which is based in part on the use of XML for data
representations, and in part through the development of the RCOMMANDS
metadata system (Tyer et al. 2007). The approach is based on automatically
harvesting metadata from the XML output files during the job workflow
immediately after the simulation has completed and before RMCS uploads the
output files to the data grid. The metadata are then uploaded to the metadata
database using the tools of the RCOMMANDS system. We collect code-specific
information, values of input parameters and user-selected output values, such as
average or final values, as metadata, from the output file, augmented by various
types of run-time information. Although not traditionally viewed as metadata,
output quantities can fulfil the same role as traditional metadata in describing
and searching for data (e.g. when searching for the simulation with the lowest
calculated energy from a set of many simulations).

The RCOMMANDS system provides a set of interfaces to the metadata database
that can be used by lightweight client command-line tools and Web browsers.
These give researchers direct access to the metadata, including the ability to
search for data from the metadata, and from the metadata the researchers have
direct access to the associated data files within the data grid.
Having a rich set of metadata provides an unexpected function, namely as one of the primary interfaces to the core data values. The metadata harvesting process acts as a first pass across the data, and the researcher who has run several hundred jobs within a single parameter-sweep study can extract a table of output values using a single command without having to download and analyse all the output files.

The RC OMMANDS client tools are available for general use on the NGS. We believe that the approach developed by the eMinerals project should have more general applicability, and have demonstrated this for other types of environmental simulations. The key to fully exploiting this approach is to use XML data representation, because it makes the automatic extraction of metadata much easier. Thus, for this approach to work for other domains requires that their simulation codes generate XML output. With the FoX library, this task is made relatively easy for Fortran simulations, and other languages already contain native XML capability. The specific choice of XML language need not be critical; CML, for example, contains many elements that are not specific to chemistry or atomistic simulations (e.g. metadata, parameter and property tags) and which could be re-used for other scientific domains.

\textit{(e) Collaboration}

We briefly remark here that the use of a data grid, XML data representation, associated tools for information delivery and metadata tools, as described previously in this section, greatly facilitate collaborative working. They provide enormous added value over the ‘traditional’ approach of collaborators exchanging files by email, a process that is virtually impossible with large parameter-sweep studies.

Collaborators require more than exchange of information. We have explored the use of interactive technologies such as instant messaging and video conferencing (including the ACCESS GRID on the desktop). We are currently running experiments that demonstrate the value of social networking technologies to support the documentation of collaborative working: www.SciSpace.net is the resultant product and is available for any researcher to join.

These issues are described in more detail in the accompanying paper by Frame et al. (2009).

3. Science outcomes

In this section, we outline a number of case studies of the science that have been enabled through the use of our eScience tools. Several of these show the power of grid technologies for parameter-sweep studies. The examples that only required a small number of calculations have nevertheless been enabled by using our local CONDOR pools or RMCS to access external resources. Most of these examples have been performed collaboratively, enabled by the information-sharing tools and other collaborative tools described above.

\textit{(a) Adsorption of pollutants on mineral surfaces}

One of the environmental problems affecting health is the widespread occurrence of pollutants within soils. It is an open question as to how such pollutants are held in soils and groundwater, and we have looked in detail at
persistent chlorinated organic pollutants such as dioxins, \( \text{C}_{12}\text{O}_2\text{H}_x\text{Cl}_{8-x} \), and polychlorobiphenyls, \( \text{C}_{12}\text{H}_x\text{Cl}_{10-x} \). Our approach has been to use a combination of quantum mechanical and empirical potential methods to determine the energies of adsorption onto soil mineral surfaces. This is a big challenge, because soils are complex environments containing minerals, organic matter, free ions and water. Thus, we are taking a systematic approach by looking at individual components first in order to properly understand the nature of the various interactions. Moreover, this is a new application area for these methods, and there is a sparsity of experimental data against which we can calibrate the tools; thus, care is needed to understand how the two methods compare.

The calculation of the energies of dioxin molecules on the (001) surface of pyrophyllite, a representative clay mineral (figure 2), shows that the energy of adsorption varies most strongly with the number of Cl atoms, but not significantly with their location within the molecule. For example, for two dioxin molecules with \( x=4 \) in the chemical formula, and where there is the largest electric dipole moment possible in one case and no dipole moment in the other, both have adsorption energies that differ only by a few thousandths of an electronvolt in the quantum mechanical calculations, both being roughly \( 0.47 \text{ eV} \). The variation in dioxin adsorption energies from \( x=0 \) to 12 is calculated to be \( 0.1 \text{ eV} \) in the quantum mechanical calculations and \( 0.2 \text{ eV} \) in the empirical simulations, the difference being associated with the lack of dispersion interactions in the quantum mechanical calculations and an overbinding in the empirical models. We have carried out calculations using higher level theory for quantum chemistry (known as Møller–Plesset methods), which suggest that the true adsorption energies actually lie between those given by these two approaches. We have also found that contact electrostatic interactions are more important than the longer range electrostatic...
contributions. Induction effects were found to be almost negligible in spite of the substantial polarizability of these aromatic pollutant molecules (Austen et al. 2008).

We also cite results for the adsorption of dioxin molecules on the (211)R surface of calcite. The adsorption energies found using quantum mechanical methods are larger than that for the (001) surface of pyrophyllite, and have a greater variation, from $-0.89$ to $-0.74$ eV for $x=0$–10.

There is a reversal of the trend in increasing adsorption energy with Cl number when studying the PCB molecules on smectite surfaces. In this case, the steric hindrance to the flattening of the fully chlorinated PCB molecule is greater than that for the fully protonated molecule, such that the latter can approach more closely to the surface and thus bind more strongly, $-0.8$ eV, compared with $-0.5$ eV for the fully chlorinated case.

We have also carried out a number of simulations with water. For example, we have calculated the change in free energy as a dioxin molecule is transported to the pyrophyllite surface. The results show a free-energy barrier as the molecule is immersed in water, not surprising given that it is insoluble, and a free energy of adsorption that is larger than the calculated adsorption energy. Despite the pyrophyllite having a hydrophobic surface, the water density above the surface is layered, and the dioxin molecule adsorbs within the lowest density layer just above the surface.

This example shows the integration of the three components of computation, data and collaboration, as illustrated in figure 1. The large parameter-sweep study performed using empirical potentials was carried out using the NGS grid resources. The quantum mechanical calculations were performed using the NWG resources. The calculations reported here were carried out using 32 processors each, a scale of parallelism that is not optimal for high-capability computers that are specifically designed to support individual tasks that require many more processors and fast global interprocessor connections, but is well matched for grid systems such as NGS and NWG that support parallel computing. Behind the headline results were many calibration runs that were also performed as parameter sweeps (Austen et al. 2006). The work was shared between two teams of researchers, namely Bath and Cambridge, with data sharing using the data grid, with the XML tools permitting the sharing of information, the use of regular desktop ACCESS GRID discussions and shared wiki tools (including the use of www.SciSpace.net) for documenting results.

(b) Simulation studies of radiation damage and nuclear waste encapsulation

The safe immobilization of highly radioactive nuclear waste and decommissioned plutonium requires us to understand the physical processes taking place in encapsulation matrices (waste forms) at the microscopic level. In particular, structural amorphization induced by $\alpha$-radioactive decay will have detrimental effects on the ability of the waste form to immobilize radioactive atoms (Trachenko et al. 2004). The challenge to be able to study systems containing millions of atoms with electrostatic interactions, as required to simulate the recoil of atoms from nuclear disintegration (cascade events) with realistic energies, has been met by the development of the DL_POLY_3 code as a key part of the eMinerals project (Todorov et al. 2006). Two examples are shown in
Not only have we been able to simulate decay events involving realistic recoil energies of approximately 100 keV, but we have also been able to study ion bombardment with energies of up to 1 MeV in order to provide a link with experiments (Trachenko et al. 2005a, 2006).

We have found that cascade events can evolve in several ways. For some materials (such as zircon), the cascade leaves a large damaged region that is stable over long time periods. For other systems, a significant proportion of initial damage anneals out, leaving either a relatively small amorphous pocket inside a crystalline matrix or point defects. We also find a significant dependence of the damage formation on the density of cascade events. Some materials will show annealing of single cascade events, but multiple events in one location will leave permanent amorphous damage. From our studies of a number of materials, we have concluded that recrystallization processes take place over surprisingly short time scales, namely of the order of several picoseconds. We have identified the mechanisms of recovery of crystallinity, and coupled with quantum mechanical calculations, we have been able to correlate the way that damage recovers with features of the interatomic potentials (Trachenko et al. 2005b). Quantum mechanical calculations have also been used to understand the local structural relaxation associated with point defects generated by cascades (Pruneda et al. 2004).

The very large cascades performed using DL_POLY_3 were necessarily performed using high-performance facilities rather than grid computing. On the other hand, the supporting quantum mechanical calculations were carried out using CONDOR pools set up in Cambridge (a forerunner to CamGrid). This work is a collaboration between researchers in Cambridge and the Daresbury Laboratory, for which we required the use of the collaborative tools discussed in this paper.

Figure 3. (a) Radiation damage created by 40 keV U recoil in rutile TiO$_2$. The direction of recoil is from top right to bottom left. This snapshot shows atoms displaced from equilibrium positions by more than 0.75 Å. The damage is not recoverable and leads to permanent amorphization. There is also shear distortion around the damaged core that is elastic and reversible. (b) Zirconolite (CaZrTi$_2$O$_7$) following two U cascades at 70 keV. The fast recoil atoms moved from bottom left to top right, and the illustration shows the resulting damage after 10 ps. For clarity, only the calcium (grey) and zirconium (black) atoms are shown.
Diopside, CaMgSi$_2$O$_6$, is one of the important rock-forming minerals. In order to understand the composition of the Earth under conditions that are not accessible to experiment directly, it is important to have a good knowledge of the elasticity of its component minerals. Increasingly, this information is being sought from quantum mechanical calculations, and this example considers a study of the crystal structure and compressibility of diopside to high pressure, performed using the resources of the NWG (Thomas et al. 2007). This study is another example of how grid resources are ideal for parameter-sweep studies, with all jobs being launched simultaneously and data collected from the data grid.

The computed equation of state (pressure versus volume relationship) is shown in figure 4; it is in reasonable agreement with experimental data, and near-perfect agreement is obtained by applying a shift to the pressure scale to account for the known underbinding of the method (density functional theory with generalized gradient approximation). The results were fitted to a standard third-order Birch–Murnaghan equation ($P$ is the pressure, $V$ is the volume and $V_0$ is the volume at $P=0$),

$$P = \frac{3K_0}{2} \left[ \left( \frac{V_0}{V} \right)^{7/3} \right] \left( \frac{V_0}{V} \right)^{5/3} \times \left\{ 1 + \frac{3(K'_0 - 4)}{4} \left[ \left( \frac{V_0}{V} \right)^{2/3} - 1 \right] \right\},$$

obtaining values of $K_0$ (the bulk modulus at $P=0$) and $K'_0$ (the derivative of the bulk modulus with pressure) of 99.8 GPa and 4.9, respectively, from the raw data, and 122.0 GPa and 4.7, respectively, after applying a small pressure correction to account for the overbinding of the calculations.

Using simulations, we were able to extend the range of pressures studies beyond the range for which the crystal structure had been measured experimentally, and we were also able to investigate in detail the changes in the crystal structure under pressure in order to identify the main mechanisms of compression. A key result was to understand how the CaO$_8$ polyhedron develops a more regular shape under compression, with an equalization of Ca–O bond lengths.

![Equation of state of diopside](image)

**Figure 4.** (a) Computed equation of state of diopside, CaMgSi$_2$O$_6$ (Walker et al. 2008). (b) Ca and O atoms in the diopside structure, highlighting the two longest Ca–O bonds at ambient pressure. By 20 GPa, these bond lengths are almost identical in the simulation results.
lengths that are very different at ambient pressure (Walker et al. 2008). We were also able to study the relative stiffness of the structural polyhedral SiO$_4$ tetrahedra, the MgO$_6$ octahedra and the irregular CaO$_8$ polyhedra. By fitting the variation of the polyhedral volumes to the same Birch–Murnaghan equation, we obtained values of $K_0$ of 334.4, 88.9 and 75.3 GPa for the SiO$_4$, MgO$_6$ and CaO$_8$ polyhedra, respectively, showing that (as expected) the SiO$_4$ tetrahedra are the stiffest components.

(d) Amorphous silica under pressure: anomalous compressibility

Amorphous silica shows the remarkable and counter-intuitive property that it gets softer on compression, up to a maximum of the compressibility, defined as $k = V^{-1} \partial V/\partial P$, at approximately 2 GPa. Increasingly, it is being recognized that a number of crystalline materials have a similar property. Not only is it interesting to understand such anomalous behaviour, but it is also important to understand compression mechanisms in general because of their role in both geological and materials sciences.

To study the case of amorphous silica, we performed large-scale parameter-sweep constant-pressure molecular dynamics simulations of silica glass under both positive and negative pressures, using three different configurations and two different empirical models for the interatomic interactions to ensure that the simulated phenomena were independent of the details of the method. The choice of pressure range was made in line with a working hypothesis that the compressibility maximum arises due to a flexibility in the structure that is reduced when the system is either stretched or compacted, and the large number of points was required for the detailed analysis (discussed below). The job workflow included an equilibration run of the simulation, a production run and two analysis calculations, with results deposited in the data grid for subsequent collation (Walker et al. 2006).

The calculated curves of volume versus pressure are shown in figure 5. All the curves show a point of inflection at approximately 1–2 GPa, which corresponds to a maximum in compressibility. To obtain a quantitative measure of the compressibility, we fitted polynomials to the $V$ versus $P$ curves, which could be differentiated. This approach minimizes the effects of statistical errors on the individual points, but requires the large number of points shown in the graph. The analysis of the configurations showed clearly that the enhanced compressibility arises from the flexibility of the silica network that is lost under extreme compression or expansion, in line with the original hypothesis (Dove et al. 2006; Walker et al. 2007).

(e) Colossal thermal expansion materials

High-performance materials will have a role in developing environmental technologies, particularly for operating under extreme conditions. We have recently discovered a material, Ag$_3$[Co(CN)$_6$], that undergoes the largest positive thermal expansion in two directions and the largest negative thermal expansion in the orthogonal direction (Goodwin et al. 2008). Materials with such properties might find application, for example, in coatings for optical devices in satellites because they can be tailored to minimize thermal strains.

Phil. Trans. R. Soc. A (2009)
The origin of the colossal thermal expansion has been shown to arise from the behaviour of argentophilic Ag...Ag interactions within an inherently flexible framework structure. The atomic structure is shown in figure 6a; the representation highlights the existence of CoC₆ octahedra that are linked across the diagonals of the unit cell through C–N...Ag...N–C connections. The connections provide the flexibility, in particular forcing an expansion in the vertical direction to be accompanied by a similar shrinkage in the perpendicular directions (and vice versa). The evidence for this flexibility and the role of the argentophilic interactions has been accumulated using a combination of the reverse Monte Carlo method—a technique to generate atomic configurations from neutron diffraction data (Tucker et al. 2007)—and quantum mechanical methods.

Figure 6. (a) Crystal structure of Ag₃[Co(CN)₆], showing the framework of CoC₆ octahedra linked by C...C chains, and Ag layers. (b) Energy surface computed by quantum mechanical methods; the thick grey curve is the locus of experimental data (Calleja et al. 2008).

The origin of the colossal thermal expansion has been shown to arise from the behaviour of argentophilic Ag...Ag interactions within an inherently flexible framework structure. The atomic structure is shown in figure 6a; the representation highlights the existence of CoC₆ octahedra that are linked across the diagonals of the unit cell through C–N...Ag...N–C connections. The connections provide the flexibility, in particular forcing an expansion in the vertical direction to be accompanied by a similar shrinkage in the perpendicular directions (and vice versa). The evidence for this flexibility and the role of the argentophilic interactions has been accumulated using a combination of the reverse Monte Carlo method—a technique to generate atomic configurations from neutron diffraction data (Tucker et al. 2007)—and quantum mechanical methods.
simulations (Calleja et al. 2008; Conterio et al. 2008). The results of the quantum mechanical simulations are shown as a contour plot in figure 6b; the highly correlated expansion and shrinkage along perpendicular directions is shown as the flat valley with very steep sides. The flatness of the floor of the valley reflects a very soft Ag...Ag interaction, which leads to a low elastic stiffness and the colossal values of the coefficients of thermal expansion (positive and negative).

This study is an example illustrating how we have used eScience tools for the analysis of both experimental data and simulation. Most of the computations were carried out using the CamGrid campus grid. The data shown in figure 6b were obtained as a fine grid of individual calculations as a parameter sweep across the dimensions of the crystal. An interesting technical note is that many of the nodes on CamGrid contain four cores, which we set up to run as small parallel shared-memory computers under CONDOR to support the quantum mechanical simulations. The SRB was used for data sharing, and the collaboration tools (particularly www.SciSpace.net) were used to document discussion between teams in Cambridge and the Rutherford Appleton Laboratory.

(f) Phase transitions and thermal expansion in carbonates

It has been known for a long time that there is a large volume change in the mineral calcite (CaCO₃, the principal component of chalk, marble and limestone) on heating. It was originally assumed that this is associated with the orientational phase transition at 1260 K, above which the planar carbonate groups are rotationally disordered about their threefold axes (Dove et al. 2005). However, we have recently considered the analogy with the isomorphous material BaCO₃, which is known to undergo a phase transition to the cubic rock-salt structure with complete three-dimensional orientational disorder of the carbonate groups. In this case, the lattice dimensions would need to change considerably, particularly along the axis that is normal to the molecular threefold axes at low temperatures.

In order to investigate this possibility, we have used the molecular dynamics method with empirical potentials (Archer et al. 2003), run with many temperatures as a parameter sweep. The results have confirmed the hypothesis that the large volume change on heating is indeed associated with the transition into the cubic rock-salt structure with complete three-dimensional orientational disorder of the carbonate groups (figure 7). The data show the known transition in calcite at a temperature of approximately 900 K, and the transition to the rock-salt structure at approximately 1700 K. This latter transition has been observed (but not characterized) in BaCO₃, and, by analogy, we propose that the same processes give rise to the large thermal expansion in calcite even though the rock-salt structure is inaccessible to experiment.

All calculations represented in figure 7 were run on the NGS. A large number of data points were collected in order to capture the details associated with the two phase transitions. It is interesting to note that the researcher (MTD) was able to set up all data files, upload them to the data grid and submit all jobs, with just two shell commands. Using the core data captured as metadata, once the jobs had been completed, the tables of data for plotting were collated with a single shell command. This final example highlights how the eMinerals toolkit makes submission, management and analysis of large parameter-sweep studies.
much easier than before. We believe that the challenges associated with the data management required for studies such as this one and some of the other examples presented in this paper would make them impractical without a considerable overhead of bespoke programming effort.

4. Summary

We believe that a new type of simulation science is emerging from eScience, one in which the integration of computing, data and collaboration is absolutely essential. Grid computing methods enable us to run far too many simulations than can be managed or documented by hand, leading to researchers being able to generate data at a rate that cannot be managed or analysed using traditional approaches. Researchers who are competent programmers may well be able to create bespoke solutions, but it is time-consuming to set these up and ensure that they are accurate and robust. Moreover, very few bespoke solutions will facilitate collaborative exchange of data and information. Thus, the provision of grid computing requires the commensurate provision of integrated computing, data and collaborative systems that are easy to use, widely applicable and not disruptive to the researcher’s work patterns; this is what the eMinerals toolkit is aimed at. The core tools are the RMCS job submission system and the RCOMMANDS metadata tools, coupled with the XML tools and libraries. The examples we have described here focus on applications to molecular simulation sciences, and the breadth of these examples (different scientific issues, different methods and different resource demands) demonstrates the generality of the approach. We have also exploited this work for quite different applications within environmental sciences, including applications in remote sensing and hydrology (Chiang et al. (2006, 2007), and work to be published).
Inevitably, it has been impossible to include all details in this paper. Some of the cited references will provide more details on specific tools. Alternatively, the eMinerals website (http://www.eMinerals.org/) provides up-to-date links to the tools described here.

We acknowledge financial support from NERC under the eScience thematic programme. We acknowledge help and support from the National Institute for Environmental eScience, the DTI-funded MaterialsGrid project, the Cambridge eScience Centre and other colleagues in the eMinerals project, the University of Cambridge, and the STFC eScience Centre (Daresbury and Rutherford Laboratories). Additional financial support from EPSRC and Trinity College Cambridge is gratefully acknowledged. We are particularly grateful to Jonathan Churchill and colleagues in the STFC eScience Centre for porting the grid tools developed by the eMinerals project onto the NGS.

**References**


*Phil. Trans. R. Soc. A* (2009)


