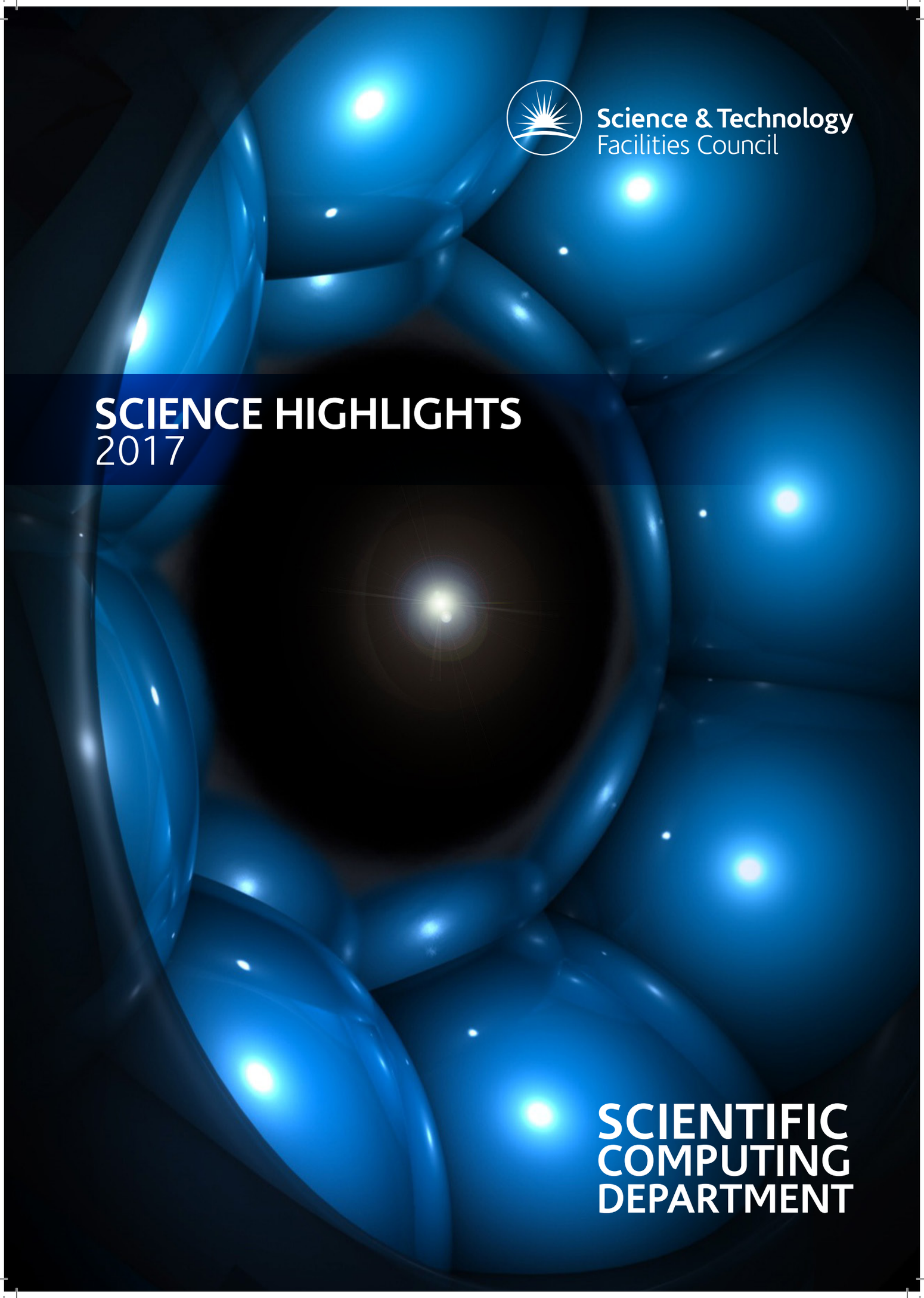

Science & Technology
Facilities Council
SCIENCE HIGHLIGHTS
2017
SCIENTIFIC
COMPUTING
DEPARTMENT

Contents

Design and layout by STFC Graphic Designer Andrew Collins
Produced by RCUK's internal service provider.

Note: All images are courtesy of STFC unless otherwise stated.

Welcome

Welcome to the STFC Scientific Computing Department (SCD) annual report for 2017.

This year the report identifies particular highlights, but within the context of the first year of the SCD 5 year strategic plan (2017 – 2021). As ever, the breadth and depth of the work of SCD to support STFC science, and its national and international collaborations is impressive and I would like to acknowledge and thank the SCD staff for their expertise and dedication in delivering on our many diverse and complex science projects, in what is proving to be a very tough year. It is well understood that lack of computational support has increasingly become a bottleneck on the delivery of science from almost all scientific disciplines, and certainly from those that STFC support – namely Particle Physics, Astronomy, Cosmology and Nuclear physics.

This time last year, we were hopeful of a positive outcome from the RCUK e-Infrastructure proposal to BEIS for the autumn statement. There was, in the end, very little funding from BEIS in response to the RCUK proposal – in total around £10m from the RCUK joint request for ~ £500m. This was disappointing for everyone involved in e-Infrastructure delivery across all Research Councils. Since then I have been involved in some strong and encouraging initiatives developing in response to these circumstances of austerity.

It is particularly encouraging to see an increasingly collaborative response from a wide cross-section of science communities to the austerity agenda. Driven by the recognition that e-Infrastructure resources need to be shared and used more effectively across different scientific disciplines, many of the STFC science communities have become united behind the UKT0 collaboration, which attempts to support all of STFC science, and to share e-Infrastructure and computational expertise wherever it makes sense to do so. This approach is already yielding results. BEIS have recently agreed to provide SCD with £3m this year to help replace ageing e-Infrastructure, and SCD is working with both the STFC facilities (ISIS, CLF, and DLS), and with the PPAN user communities, to optimise this resource to benefit as many of the STFC science users as we can. Meanwhile SCD has also been working with others within STFC and across RCUK to develop a new proposal for BEIS for this year's autumn statement. That proposal is now very well developed and will be delivered to BEIS shortly. We remain cautious after last year's outcome, but ever optimistic – principally because we all recognise the serious impact on UK and STFC science if we have only very limited investment in UK e-Infrastructure for another year.

A further significant development this year has been the increasing emergence of the "STFC e-Infrastructure strategy". The STFC e-Infrastructure strategy has built on the outcome of the STFC Strategic Computing Review (published in Dec 2015), and recognises the need for STFC to improve the coordination of computational resources across all of STFC science. The STFC e-Infrastructure strategy has been through many levels of approval: through National Laboratory Directors, then to the STFC Computing Advisory Panel, and on to STFC Executive Board. It is currently waiting to go to the STFC Council for its final approval, (expected in November 2017), before the strategy becomes a public document. SCD is already working with others within STFC to establish how the strategy will be implemented. This strategy represents a major step forward for STFC. It will also help to position STFC as a major contributor to e-Infrastructure planning for Science within its RCUK partners on the final run in to UKRI.

So enough of the politics. It is the science that counts, but I hope that this introduction has helped to set the context of some of the difficulties that SCD staff and its science user communities have had to cope with this year. From the contents of this report, I am confident that you can continue to understand the impressive and wide ranging contributions that SCD staff are making to STFC science and beyond, despite very difficult circumstances. SCD staff remain dedicated to ensuring the delivery of very high quality science, and it is an honour for me to work with them.

I hope you enjoy our annual report for 2017.

David Corney

Director, Scientific Computing Department

Scientific Computing Department Strategy

During 2016-17 we launched our new 5 year strategy to help us to achieve our mission:

"to maximize the impact of scientific computing through our expertise, leadership and collaboration."

Our mission is fully aligned to support and enable the STFC Corporate Strategy and the STFC e-infrastructure Strategy.

The Scientific Computing Department fulfils its mission by:

- Designing, deploying and operating large and complex computing and data systems
- Supporting the research life-cycle by extracting insights and value from data
- Creating algorithms and software to exploit future research computing infrastructure
- Providing cross-domain expertise to develop, innovate, and sustain software, and related digital assets for research
- Leading and participating in the national and international collaborations working to achieve these aims
- Committing to STFC's corporate theme of Inspiring and Involving to deliver a vibrant programme of public engagement
- Actively supporting and contributing to the developing skills agenda in Scientific Computing

Delivering our mission

SCD is one of the largest scientific computing departments in Europe and comprises around one hundred and fifty staff. We deliver expertise in computational science and professional large-scale scientific and data management and computing systems, services and expertise to our STFC, national, and international scientific user communities, and to our collaborators and stakeholders in other Research Councils – particularly the Engineering and Physical Sciences Research Council (EPSRC), the Natural Environment Research Council (NERC), the Biotechnology and Biological Sciences Research Council (BBSRC), and the Medical Research Council (MRC).

Delivering our mission builds on our core strengths and is described in the following sections.

Theme 1

- SCD will build and strengthen our existing capability of providing enabling computational expertise and e-infrastructure to support STFC in fulfilling its strategic goals: "world class research, world class innovation, and world class skills."

Theme 2

- SCD is committed to deliver and further develop its expertise in scientific analysis, data science, modelling and simulation, to support STFC's commitment to: Establish the Ada Lovelace Centre, an integrated, cross disciplinary, data-intensive science centre, to transform the use of real time data processing, computer simulation and data analytics to deliver more effective research at our national facilities.

Theme 3

- SCD delivers a comprehensive programme of computational and data science services, research and development to underpin STFC's Data Intensive Science ambition: to develop and deliver cutting edge solutions for academia and industry to advance data intensive science and innovation.

Theme 4

- In order to ensure that the UK's e-infrastructure supports the country's leading international research status and delivers the data capabilities essential for academia, industry and STFC's science program, SCD will collaborate with our Research Council UK partners and stakeholders, and with Innovate UK, and JISC, as part of the emerging UK Research and Innovation (UKRI), to shape and build an efficient and effective UK e-infrastructure for UK science and industry.

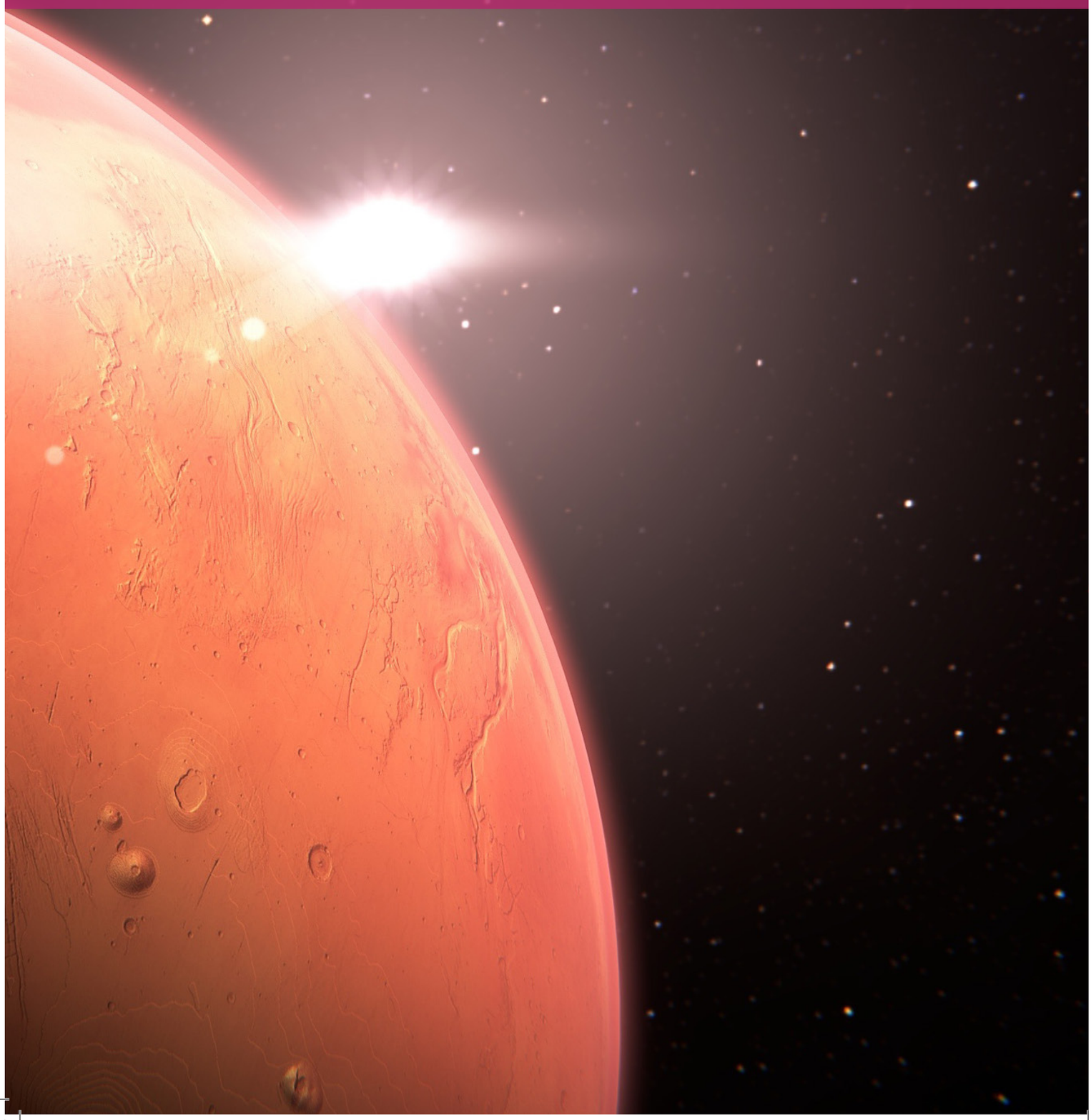
More details can be found in the SCD 5 year Strategy 2017-2021 document which can be downloaded here: https://www.scd.stfc.ac.uk/Pages/SCD-Strategy.aspx

SCD is just starting on the journey to implement the five year plan. The articles in this report give a spotlight on activities so far.

Strategic Theme 1

SCD will build and strengthen our existing capability of providing enabling computational expertise and e-infrastructure to support STFC in fulfilling its strategic goals: "world class research, world class innovation, and world class skills."

Simulating the Aerodynamic Characteristics of a Novel Airfoil in the Martian Atmosphere



With the increasing interest in deploying aircraft to Mars, our scientists are using advanced computer simulations to test the performance of an airfoil in the Martian atmosphere. Accurate predictions will help researchers in designing the optimum shape needed for successful flight on Mars.

Background

As part of a collaborative research project, scientists from STFC, University of Manchester and China Academy of Aerospace Aerodynamics are investigating the aerodynamic characteristics of an airfoil in the Martian atmosphere to target the future launch of an unmanned aircraft on Mars. It has been proposed that an unmanned aircraft could be deployed to explore the surface of Mars efficiently, as shown in Figure 1. The unique atmospheric conditions of Mars (low density, low temperature, with a gaseous composition consisting primarily of CO_2, as shown in Table 1) require the wing and propeller of any Martian aircraft to operate at a low Reynolds number (Re=103~105) and high Mach number (Mach=0.5~0.8) condition, as shown in Figure 2. The airfoil performance at this condition has some special problems, such as a nonlinear lift curve caused by the formation or burst of laminar separation bubbles (LSB), which are not well understood.

With increasing interest in the deployment of an aircraft on Mars, an accurate prediction of the aerodynamic performance of an airfoil in the Martian atmosphere, with a clear understanding of the flow mechanism involved in LSB, is required. A high-performance direct numerical simulation (DNS), which resolves all details of the flow dynamics, is currently being used to study the aerodynamic characteristics and flow structures of an airfoil designed for flying in the Martian atmosphere.

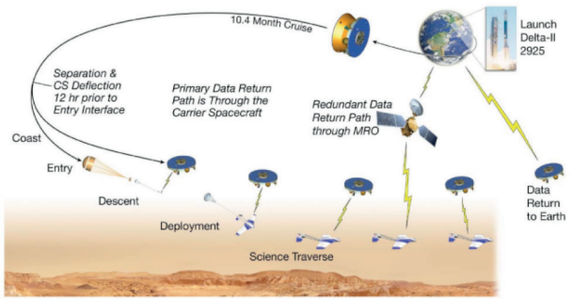


Figure 1. Mission concept overview of Mars airplane [1]

Characteristic		Mars	Earth
Distance from Sun		228M km	150M km
Atmosphere	Pressure	0.6k Pa	101.3 kPa
	Density	0.015 kg/m^3	1.293 kg/m^3
	Composition	95% CO_2 3% N_2 2% Ar	78% N_2 21% O_2 1% Ar
	Temperature	133-293 K Mean: 210K (-63°C)	185-331 K Mean: 287K (14°C)
Length of solar day		1 sol=24h 40m	24h
Length of year		668.6 sols 1.88y=687d	365.25d

Table 1. Comparison between Mars and Earth environment

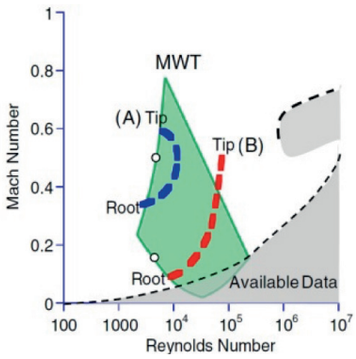


Figure 2. Reynolds and Mach number range of Mars airplane [2]

Computational Fluid Dynamics Software

The compressible high-order finite different CFD code ASTR (Advanced flow Simulator for Turbulence Research) developed and tested in the Scientific Computing Department has been adopted to study flight on Mars. The code has previously been used in the research of high-speed flow dynamics and shock-wave/turbulent boundary layer interactions, and is proven to be efficient in resolving detailed structures in compressible turbulent flow and performs well on high-performance computers, as illustrated in Figure 3.

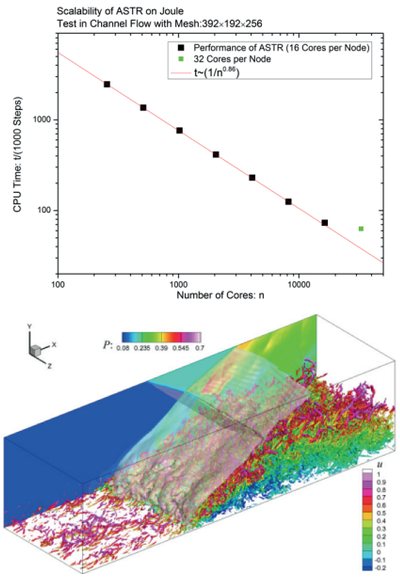


Figure 3. Scalability of ASTR and resolved structures in a supersonic turbulent flow by ASTR.

Aerodynamic characteristics of the Martian Airfoil

A triangular airfoil [2] designed by an American team and tested in Japan is studied here. Figure 4 shows a sketch of the airfoil geometry and the mesh we used to discretise the domain. The lift and drag coefficients agree well, in general, with experimental measurements and DNS of Munday *et al*. [2], although a discrepancy appears when the angle of attack is at a high degree, indicating the complexity of the flow when angle of attack is high, as shown in Figures 5 & 6. There are two flow separation modes identified with the increase of the angle of attack. For a≤6°, a recirculation zone forms behind the apex. At a=9°, a leading-edge vortex forms a recirculation region with reattachment upstream of the trailing edge. For a≥12°, the recirculation zone grows and covers the whole airfoil.

The flow state also depends strongly on the Reynolds number (Re). Figure 7 shows flow structures at two typical Reynolds numbers. At a low Reynolds number (Re=3,000), the flow is basically laminar and three-dimensional structures only appear in the wake region. However, at a high Reynolds number, rich small-scale turbulent structures almost cover the entire upper surface against the airfoil, which enables the flow to be robust against flow separation.

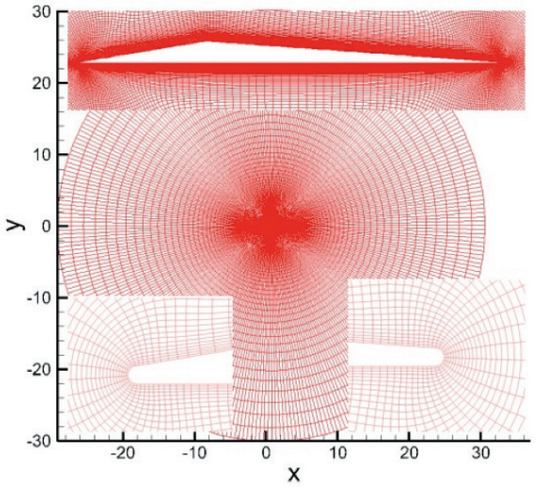


Figure 4. The O-type mesh used in the study.

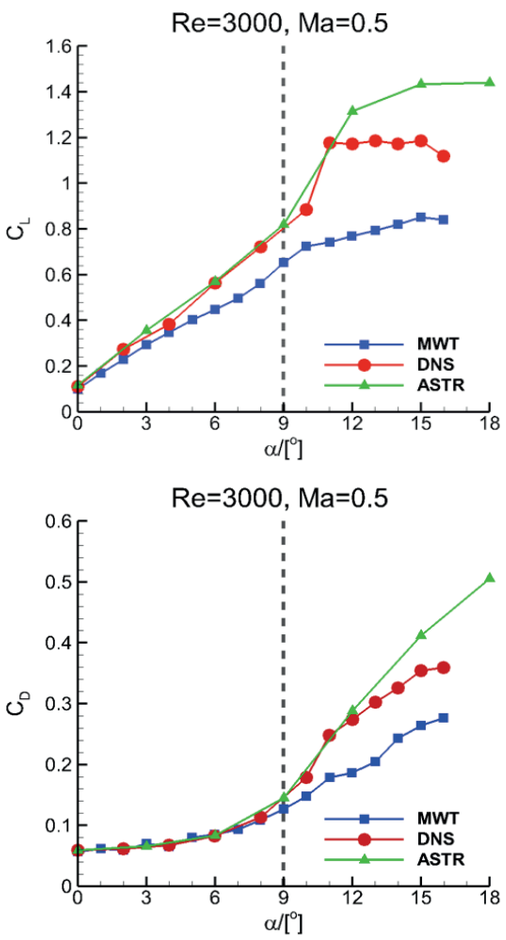


Figure 5. Comparison of time-mean force coefficients with the results from reference.

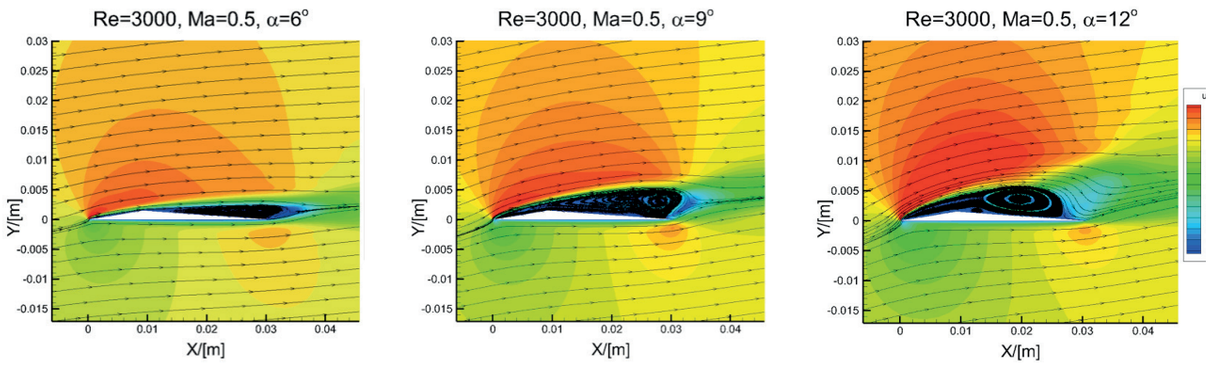


Figure 6. Time-mean flow fields with the contour of normalized velocity magnitude.

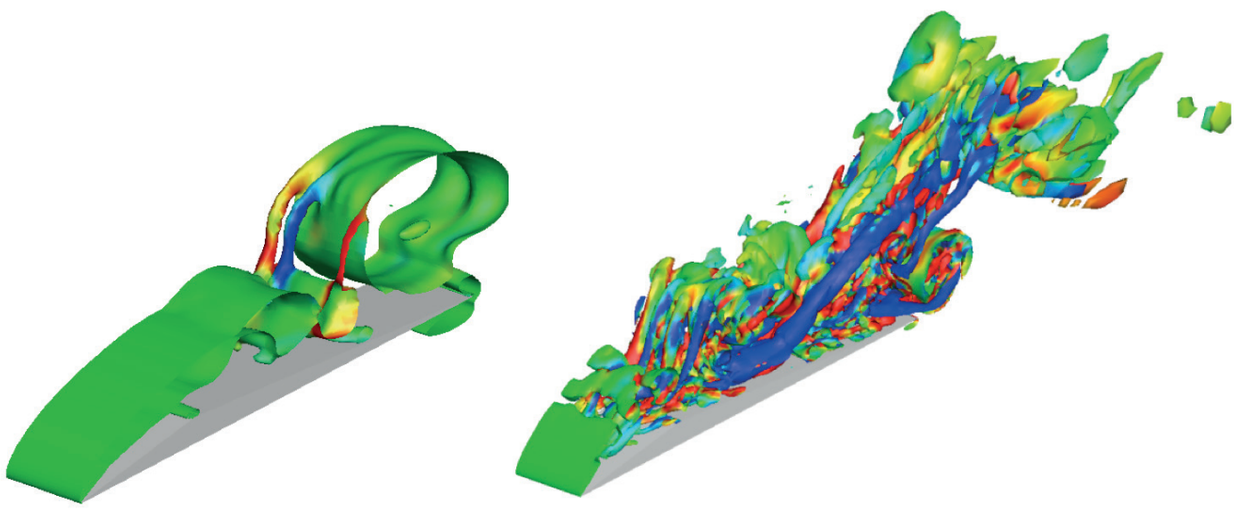


Figure 7. Instantaneous flow structures at Re=3,000 (left) and Re=10,000 (right).

Acknowledgements

The authors thank the UK Turbulence Consortium under EPSRC grant EP/L00030X/1 for access to the UK's national supercomputing facility, ARCHER.

Authors
J.Fang[1], H. Zhan[2,3], A. Revell[2], C. Moulinec[1] and D. R Emerson[1], [1]Scientific Computing Department, STFC Daresbury Laboratory, [2]School of MACE, The University of Manchester, [3]China Academy of Aerospace Aerodynamics, STFC Daresbury Laboratory

References

[1] C. A. Kuhl, Design of a Mars Airplane Propulsion System for the Aerial Regional-scale Environmental Survey (ARES) Mission Concept, 44th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, AIAA Paper 2008-5246, 2008.
[2] P. M. Munday, K. Taira, T. Suwa, D. Numata, and K. Asai, Nonlinear Lift on a Triangular Airfoil in Low-Reynolds-Number Compressible Flow, Journal of Aircraft, 2015, 52(3): 924-931.

Strategic Theme 1

SCD will build and strengthen our existing capability of providing enabling computational expertise and e-infrastructure to support STFC in fulfilling its strategic goals: "world class research, world class innovation, and world class skills."

DL_POLY/DL_FIELD/DL_ANALYSER – An Integrated Software Engine for Molecular Simulations

Computational scientists at the Daresbury Laboratory have developed and integrated a series of software packages to investigate the behaviour of materials at the atomic scale. Some materials used in food packaging, for example, are composed of miniscule particles, some as small as 1000th the width of a human hair, which may be adsorbed into food and cause major concerns for human health. Other materials, such as those used in spacecraft design, can be badly damaged by tiny meteorites and debris crashing into them as they hurtle through space. Our scientists are using molecular simulations to explore the physical and chemical behaviour of these and other materials. Their work will aid the design of new materials and provide a cost-effective method of testing before materials are widely produced.

Three separately developed software packages at Daresbury Laboratory have been integrated to form an efficient computational infrastructure to investigate a variety of phenomena at atomic scales. This is achieved by inventing a new universal notation called the DL_F Notation to describe the chemical behaviour of atoms that are essential to assign the correct values for computation and enable scientists to carry out analysis that can directly relate to the behaviour of these atoms.

Two test cases have illustrated how the software can be used to answer some important science questions. In test case 1, the effects of extremely tiny particles (with sizes down to a thousandth of the width of a human hair) on human health is a major concern, in particular, how they can stick and adsorb into biological cells. In most cases, foreign particles generally bounced off because of a protective layer of water round the cell membrane. The example illustrates how a tiny polythene particle can glide and slide, making attempts to bypass these protective layers that eventually draws it into the cell layer. In test case 2, the calculations involved smashing a stone on an aluminium surface, mimicking the damage that may be caused to a spacecraft's hull as a result of a tiny meteorite impact in space. The diagrams illustrate bits of aluminium that are ejected violently into space, creating a crater on the hull. They also illustrate such a violent event can occur without making any actual damage to the valuable spacecraft itself.

Introduction

Molecular simulations are essential computational techniques to provide information on the physical and chemical behaviour of the compound materials at atomistic details that are not readily assessable by experimental means. From such, a wealth of valuable information can be extracted to provide the predictive chemical and structural behaviour of a system model. This article will briefly describe how such integration enable users to efficiently produce quality scientific output: from model setup, to simulations and finally the results analysis by using the software packages DL_FIELD, DL_POLY and DL_ANALYSER, respectively. These packages formed part of the DL_Software [1], a collective term for the scientific software packages that are developed at Daresbury Laboratory.

DL_POLY [2] is a versatile and powerful molecular dynamics (MD) program suite that can run efficiently on a variety of computer platforms, from a single-processor PC to massively parallel supercomputers. DL_FIELD [3] is a support application tool for DL_POLY, especially in setting up complex force field models. DL_ANALYSER [4] is a robust tool to carry out post analysis work on the simulation outputs produced by DL_POLY.

DL_POLY is developed to be highly agnostic in nature, whereby it can handle molecular models of varying size and complexities: from inorganic materials such as minerals, zeolites, to complex topological structure models such as graphenes, organic cages, biomolecules and even mixed component systems such as the bio-inorganic models. In case of DL_FIELD, the philosophy behind the software development is to minimise the requirement for users to understand detail knowledge and inner workings of complicated force field descriptions and preparation procedures. It is intended to serve as a user-friendly tool that automatically processes the molecular information with minimum user intervention. On the other hand, DL_ANALYSER contains a collection of unified analysis tools that can produce results in a single read through of a collection of DL_POLY's trajectory files.

Software Integration

To ensure smooth software integration and data transitions between different packages, the standard DL_F Notation [3] is implemented within DL_FIELD. This is a universal notation scheme that standardises the expression of atom types in a molecular system, of which the atom typing procedures are essential to ensure correct setup of simulation models. The DL_F notation is to provide a common solution that harmonises the atom typing which varies wildly from one force field scheme to the other. Quite often, the conversion of one force field scheme to the other is not a trivial task. The notation removes layers of complexities involving data structure conversions and ensures smooth data transition among various force field schemes.

In addition, the notation completely describes the actual chemical identity of an atom within a molecule that can be easily interpreted by both the computational modellers as well as experimentalists. This allows correlation of both cognitive and computational assessment of the simulation results and directly relates the analysis to the actual chemistry of the test structures over a range of different force field schemes. In addition, the ease of interpreting the DL_F Notation means the DL_POLY's trajectory files can be archived for analysis using DL_ANALYSER and to be assessed by other future researchers with only a minimum training requirement.

Figure 1 illustrates the integrated workflow environment of the molecular simulation software engine. In most cases, no further programming or scripting is necessary for the data transfer between packages. For instance, the input files generated by DL_FIELD via a single-step process can be used for simulation runs in DL_POLY without further modification. The dash arrow also indicates future plans to integrate other DL_Software packages by making available to other packages the wide variety of force field schemes generated by DL_FIELD.

Below illustrates two test cases that make use of the molecular simulation software engine.

Test case 1 - study of interactions between polymer nanoparticles and cell membranes.

Nanoparticles are now widely used in areas such as food sciences, materials sciences and common household applications. An understanding of the structure of nanoparticles and the interactions with biological cell membranes is important for understanding nanotoxicological effects on human and animal health and the environment. For instance, it is known that nanoparticles can easily be ingested and adsorbed into living organisms and yet studies regarding their biocompatibility and cytotoxicity effects are still quite limited.

This test case demonstrates the use of the molecular dynamics as a complementary tool to investigate the adsorption behaviour of polymer nanoparticles on cell membranes [4]. Molecular simulation can isolate and investigate systematically specific factors that contribute to the behaviour of particle interactions with the cell membranes, and the underlying atomistic mechanisms.

The simulations were carried out using DL_POLY_4 software package. The DL_FIELD software was used to set up the molecular system, with the all-atom CHARMM36 as the force field model to describe the molecular system. All results analysis was carried out using DL_ANALYSER. The visualisation and graphical outputs were generated using VMD [5].

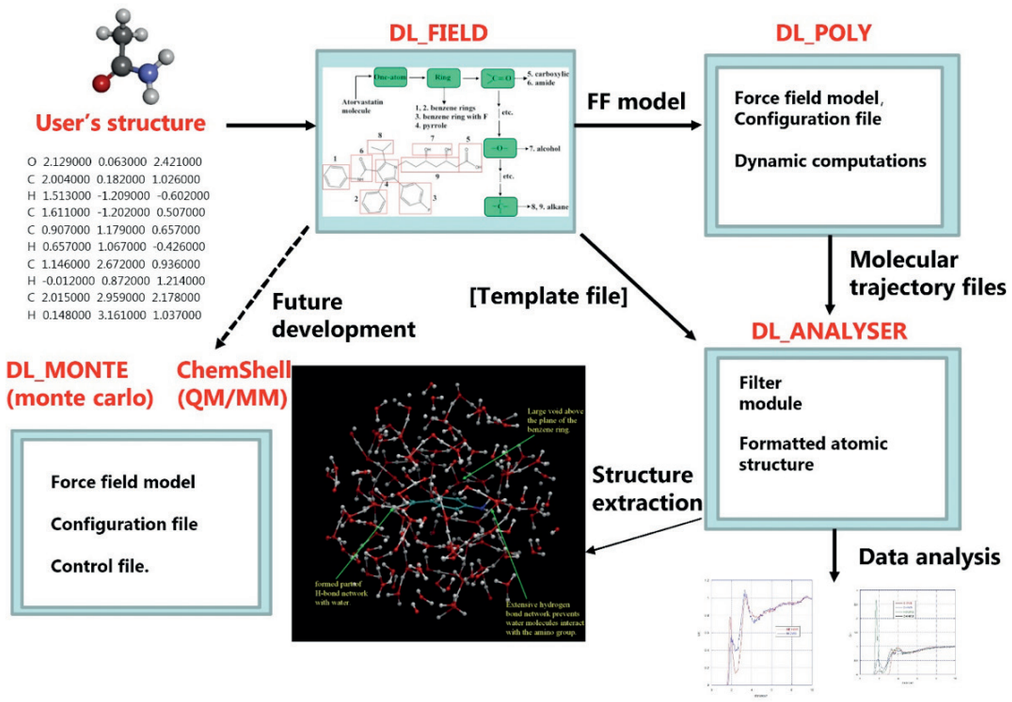


Figure 1. Diagrammatic flow chart of the integrated molecular simulation software engine. The arrows indicate the directions of the data flows.

Figure 2 illustrates the adsorption process of a polyethylene nanoparticle into the POPC biological membrane. It was found that the whole nanoparticle reoriented in such a way that the folded polymer main backbone chain aligned approximately to those of hydrocarbon chains in the membrane, while at the same time the nanoparticle chain slides and glides against each other as the particle is immersed into the membrane. Perhaps such knowledge would be crucial in nanoparticles manufacturing and fabrication by altering the chemical structure in order to prevent these interactions from occurring.

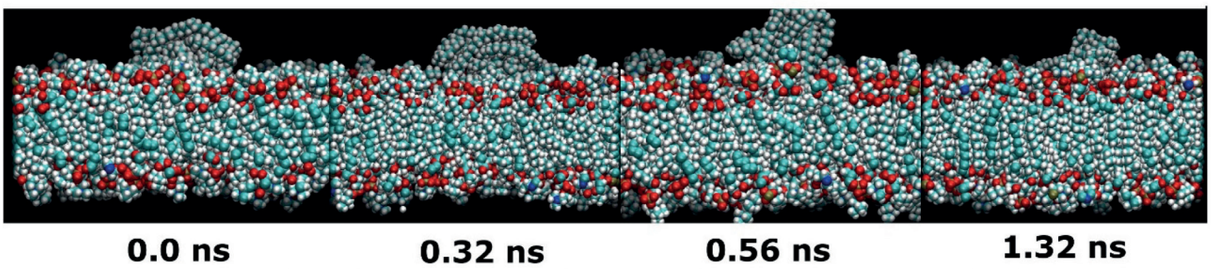


Figure 2. Four movie snapshots of the molecular configurations of polyethylene nanoparticle adsorption on the POPC membrane surface that illustrates the process of particle insertion into the membrane.

Test case 2 – Energetic nanocluster impact on metal surfaces.

We are interested in the damage and corrosion mechanism of surface materials due to orbital debris and micro-meteoritic bombardment in space. The extreme environment of interplanetary space provides an ultimate testing ground to the performance and endurance of materials that are used to construct satellites and space crafts. The particle debris can achieve an impact velocity in the order of 10 km/s. Such collision can cause permanent damage and degrade the material performances in terms of structure and chemical and mechanical behaviour. Testing materials on Earth present a challenge in recreating the appropriate conditions of outer space, as it is very difficult for experiments to capture the dynamics of these highly transient events at atomic scales.

In this test case, an aluminium (Al) surface was bombarded by a magnesium oxide (MgO) nanoparticle with an impact velocity corresponds to 1 keV. Aluminium is the most common construction material in the space industry. MgO represents a typical 'stony' meteorite material that is largely consisted of some mineral oxides and silicates.

The model consists of an Al surface that was large enough to contain the energetic event (in the order of several 105 atoms). A custom-made force field was constructed (Morse splined with Moliere potentials) for the interactions between the metal surface and the meteorite and tabulated in DL_POLY. The large-size trajectory files that subsequently produced were analysed using DL_ANALYSER.

Figure 3 illustrates the sequence of the transient impact process, which occurred in the sub-picosecond time frame. The time = 0.0 ps marked the beginning of the impacting process. During the process at time = 0.23 ps to 2.58 ps after impact, a large amount of energy was imparted into a small area of the surface, which subsequently caused the impact area to implode, while those atoms near to the surface were ejected violently into space as sputtered species. At the final time = 32.0 ps after impact, the surface was eventually cooled as the energy was dissipated to the bulk. Partial healing to the Al surface was also evident with some of the Al atoms re-deposited and solidified as adatoms on the surface. The projectile was also fragmented and embedded near to the impact crater.

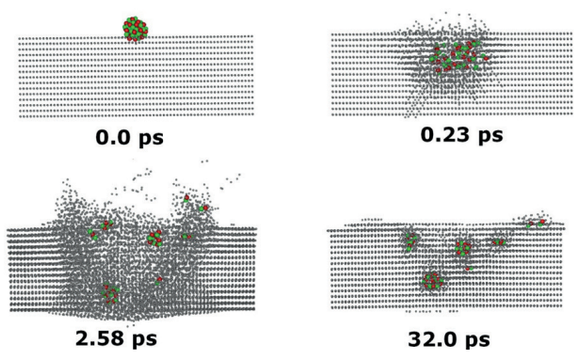


Figure 3. Four movie snapshots of the lateral views of the system configuration. The projectiles have been exaggerated as green and red spheres, representing the magnesium and oxide ions respectively. For clarity purposes, only a small section of the model is shown. The reference time = 0.0 ps indicates the moment when the particle comes into contact with the surface.

Fig. 4 shows the centre of mass velocity component along y-direction (perpendicular to the surface orientation), for the atoms located in the respective Regions as a function of time. The Regions are alphabetically labelled from A to E, as shown in Fig. 4 on the right, which are defined by a series of hemispheres each with the center O, which is the point of impact. Larger spheres are successively drawn out, each with an increase of the radius by a value Ri. Negative velocity component means the average Al atoms in a given region are moving into the surface, whereas positive velocity component means they are moving away from the surface. It can be seen that the successive peaks labelled from B to E reveal the nature of the energy deposition as the energy front travels from one Region into another. The magnitude of the peaks become smaller as the energy is deposited to a greater number of Al atoms. The initial speed of propagation was estimated to be 9000 ms_{-1} and reduced to 6500 ms_{-1} at region E. Assuming the speed of sound in Al is 6400 ms_{-1}, this means that shock waves were generated during the energy propagation.

Conclusions

This article highlights the capabilities of the molecular simulation software engine developed at Daresbury Laboratory by making references to two test cases that are very different in nature. The first test case refers to the complex biological systems specifically in the areas of nanotoxicology, while the second test case refers to the large system models that are relevant to the areas of materials science. The article also illustrates how a variety of system models can be efficiently investigated to produce scientific results in a cost effective way.

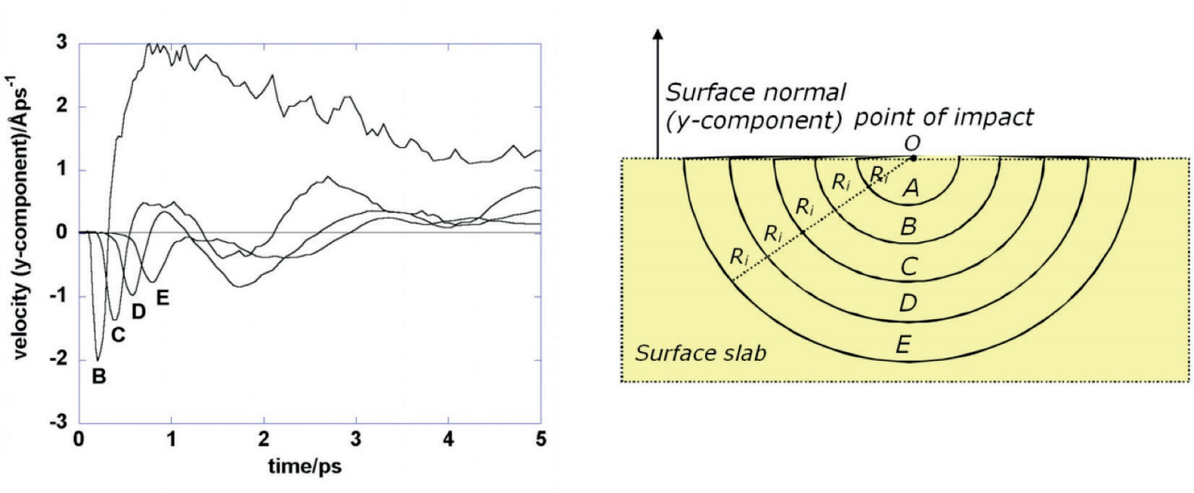


Figure 4. Centre of mass velocity along y-component as a function of time. Each profile corresponds to the velocity component in different Regions as marked by the alphabets. Profiles for Region A are omitted for clarity purposes.

Authors
C. W. Yong[1,2] , [1]Computational Chemistry Group, Scientific computing Department, STFC, Daresbury Laboratory, [2]Manchester Pharmacy School, Faculty of Medical and Human Sciences, Manchester Academic Health Science Centre,

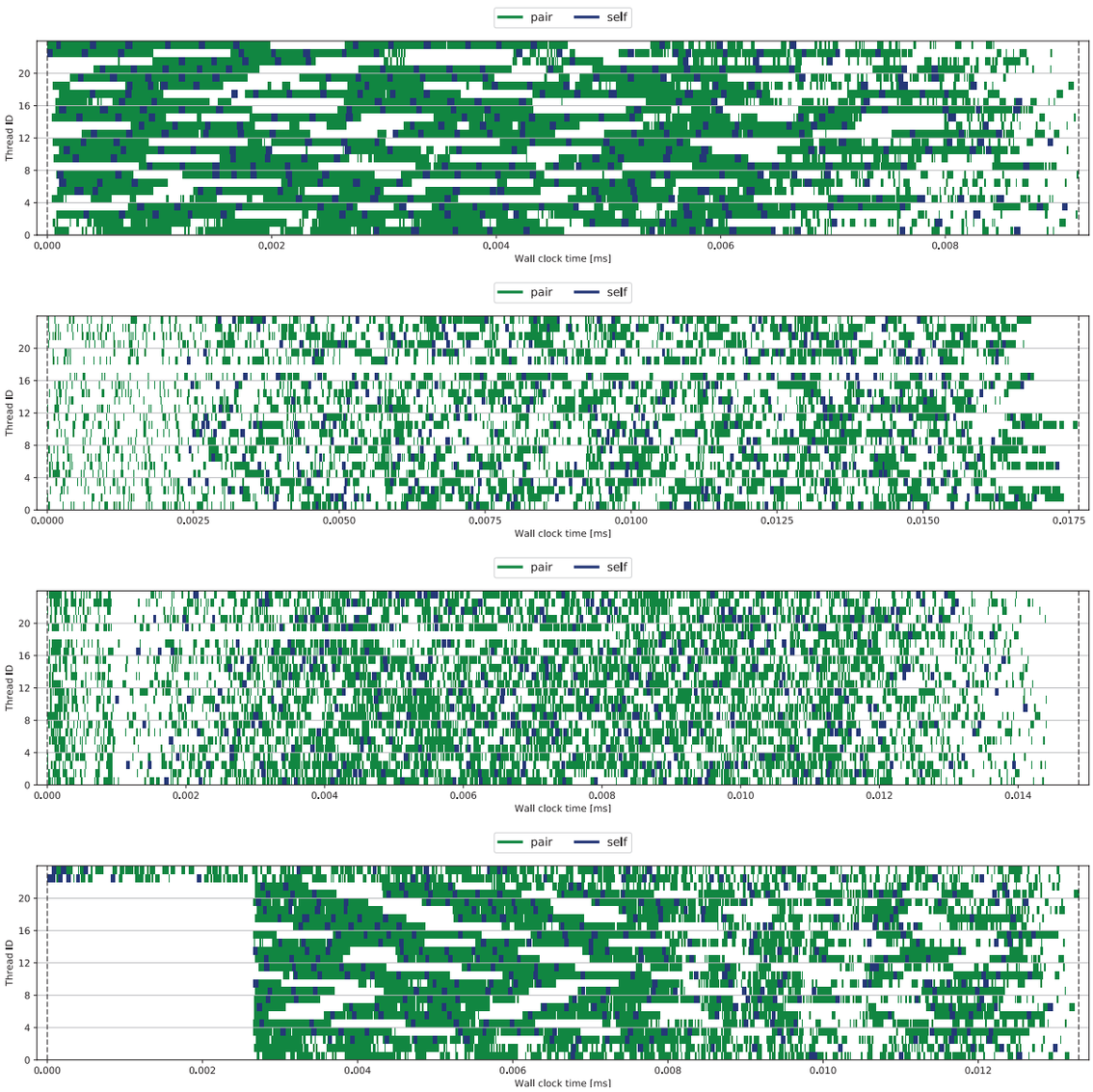
References

[1] DL_Software is the collective term for a range of scientific software suites developed at the Computational Chemistry Group at Daresbury Laboratory, spanning across multi-length and time scales, from quantum mechanics to mesoscopic scales. (http://www.ccp5.ac.uk/software).
[2] I. T. Todorov, W. Smith, K. Trachenko and M. T. Dove, J. Mater. Chem. 16, 1911 (2006).
[3] C. W. Yong, J. Chem. Inf. Model. 56, 1405 (2016).
[4] C. W. Yong, Phil. Trans R. Soc. B. 370: 20140036.
[5] W. Humphrey, A. Dalke and K. Schulten J. Molec. Graphics 14, 33 (1996)

Strategic Theme 1

SCD will build and strengthen our existing capability of providing enabling computational expertise and e-infrastructure to support STFC in fulfilling its strategic goals: "world class research, world class innovation, and world class skills."

Task Based Parallelism with OpenMP: A Case Study with DL_POLY_4



Our scientists are developing ways to speed-up codes and produce faster and more efficient processing. One of these is Task-Based Parallelism, a method for passing data between processes in which a program is split into a series of tasks, which are then assigned to multiple cores (processing units) all working in parallel until the calculations are completed. This method reduces the amount of computing time needed, and increases the time scientists can spend on research.

Introduction

When performing computations where load balancing is complex, dynamic load balancing is becoming increasingly necessary. In this paper we examine one of these methods, Task-Based Parallelism. Many libraries implement Task-Based Parallelism, however in this paper we examine the OpenMP standard and implementations, and apply it to the Classical Molecular Dynamics code, DL_POLY_4, focusing on the two body force calculations that make up a large percentage of the compute in many simulation runs. Our results show reasonable performance using Open MP tasks, however some of the extensions available in other libraries such as OmpSs or StarPU may help with performance for problems similar to Molecular Dynamics, where avoiding race conditions between tasks can have a substantial scheduling overhead.

Task-Based Parallelism is a method for shared memory parallel programming in which a program is split into a series of tasks, which are picked up and executed by a set of cores in parallel until the computation is completed. To avoid concurrency issues, the *dependencies* between tasks need to be considered, i.e. if there are 2 tasks, A and B, where task A produces a result needed to compute task B, we say B is dependent on A, or A *unlocks* B. The tasks and their dependencies form a Directed Acyclic Graph (DAG). Task-Based Parallelism has two major advantages over traditional parallel processing techniques.
Firstly, since the tasks are assigned to the cores dynamically, the work is automatically load balanced. Secondly, the task dependencies avoid the need for explicit synchronisation between the cores. In the OpenMP 3.0, 4.0 and 4.5 specifications, Task-Based Paralellism has been added and extended, allowing a standardised way to use task-based parallelism.

DL_POLY_4 [1] is a classical molecular dynamics code developed at Daresbury Laboratory, containing a large number of time integrators and statistical ensembles. The code is MPI parallelised, and uses a simple domain decomposition to distribute work between the processors. In this paper we show how we used OpenMP tasks inside DL_POLY_4, and show the performance of the resulting code.

OpenMP added explicit tasks in the OpenMP 3.0 specification, and expanded upon it in the 4.0 and 4.5 specifications. As with many OpenMP directives, additional options should be added to the task directive to control the data sharing between threads. Tasks can be executed by any thread from the same (innermost) parallel region as the thread that encounters the task directive (known as the *spawning thread*). Each task's execution is usually (but not always) deferred, but the standard specifies a variety of points at which tasks may be executed, known as task scheduling points. The most important of these for this work are: i) Immediately after generation of an explicit task; ii) Immediately after completion of a task region; iii) When encountering a taskyield directive; iv) Inside any implicit or explicit barrier. When a thread encounters a taskyield directive while inside a task region, the current task may be suspended and another task be executed instead. The depend clause allows the programmer to specify the data requirements of tasks, as either *in, out,* or *inout*. The priority clause allows the programmer to give each task an integer priority, and allows the scheduler to prioritise tasks of high priority.

OpenMP tasks in DL_POLY_4

To use Task-Based Parallelism in a scientific code, the code needs to be able to be broken down into chunks of work which can be executed in parallel. To help achieve this in DL_POLY_4, we rewrote the algorithm that performs the pairwise force loops. DL_POLY_4 uses the linked-cell method [2] [3] to build Verlet lists to compute pairwise interactions. While this method has been used with tasks in [4] it is not expected to give the best performance, as global Verlet lists have been shown to perform poorly on modern architectures [5]. Instead, we use the sorted cell-list algorithm [6], with the aim of improving to the pseudo-Verlet algorithm if successful [7].

The sorted cell-list algorithm involves organising the particles into cells of size equal to or greater than the cutoff radius (r_c) in each dimension. Since DL_POLY_4 divides the domain into equal sized domains with its domain decomposition, we split these domains into equal sized cells. The cells size must be bigger than r_c in each direction but not too large as this will affect performance. The cells are then sorted in each of the axes between the centres of pairs of adjacent cells (in 3 dimensions there are 26 axes). Gonnet et al. in [6] recommends sorting in 13 axes and then reversing the indices for the other 13 axes.

In DL_POLY_4 however, halo cells may have different dimensions to the domain cells, so boundary cells (those adjacent to halo cells) need to be sorted in all 26 axes. The cells can be sorted in parallel, though we did not use OpenMP tasks to parallelise this section of the code (due to issues with dependencies). The particles themselves are not sorted, but the indices of the sorted particle arrays are stored for each direction. Figure 1 shows how the particles are sorted on the cell pair axis.

The traditional neighbour list method can be intensive with respect to memory usage, as the neighbour access may not be cache-friendly. The addition of the link-cell does not help with this in DL_POLY_4, as the cells are just stored using a linked-list, as are the neighbour lists. The sorted-cell lists however involve a reordering of some of each particle's data, such as the position and forces upon the particle, resulting in this information being located adjacent in memory, which may improve cache performance. Additionally, since particles in the same cell are likely to interact with similar particles from an adjacent cell, performing interactions by cell can improve cache reuse.

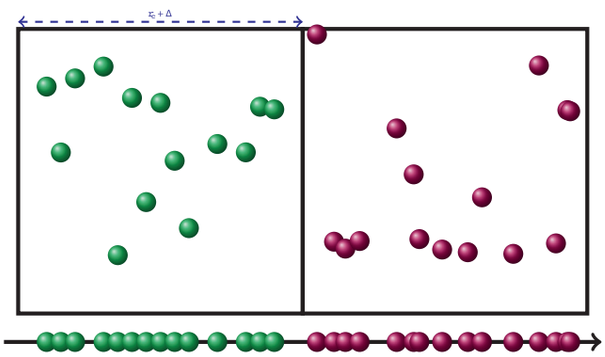


Figure 1: The figure shows two cells and the particles contained within, and the particles shown sorted along the axis between the cells. rc is the cutoff radius between the cells and Δ is a positive constant to enforce that each cell is at least r_c.

As the cells are at least as large as r_c in each dimension, we know all the neighbours of any particle in a cell are contained either in the same cell or an adjacent cell. In DL_POLY_4 this means we have 3 types of tasks: Cellself tasks. These tasks compute the pairwise interactions between all particles within a single cell; Local cell-pair tasks. These tasks compute the pairwise interactions between all particles within a pair of non-halo cells, and update the force on the particles in both cells; Nonlocal cell-pair tasks. These tasks compute the pairwise interactions between all particles within a non-halo cell and a halo cell, and update the force on the particles in the non-halo cell. In addition to specifying the tasks, we need to ensure that we avoid race conditions between tasks. Each task writes to the particle data associated with either one or two cells, so we need to avoid multiple tasks that write to the same cells being executed simultaneously. In OpenMP tasks, this can be done either manually, or automatically using the depend keyword. Avoiding race conditions using the depend keyword is straightforward. We add a depend (inout:cells), where cells is a list of the cell objects written to during each task. For a cell-self or non-local cell-pair task, this will be a single cell, whereas for the local cell-pair tasks it will contain both the cells required for the task. When using dependencies with OpenMP, any tasks that depend must be spawned by the same thread. The depend keyword adds dependencies between any tasks that share a variable (or array section) contained in any dependency clause (subject to certain rules). In our system this means dependencies between any tasks that write to the same cell. The first task to be spawned must be executed before any dependent tasks can be executed. This can lead to serialisation of the work (and thus poor parallel performance) for particle methods, as shown in [8].

In molecular dynamics the order in which the tasks are executed does not matter, at least within the pairwise interactions, but one wants to avoid race conditions. As such, the dependencies currently available in OpenMP are too constraining. Instead, we want to use conflicts, as described in [9], also known as commutative dependencies in OmpSs [10] and StarPU [11]. It is possible to implement something similar using OpenMP 4.5 using taskyield and locks. We extend the cell type to contain an OpenMP lock. When executing a task, we attempt to lock any cells that are written to by the task. If successful, then the task executes as normal. If one of the locks can't be obtained, we unlock any locks obtained and yield the task using the taskyield OpenMP directive:

```
do while(.not. locked)
 if(omp_test_lock(cell_i%cell_lock)) then
  if(omp_test_lock(cell_j%cell_lock)) then
   locked = .true.
  else
   Call omp_unset_lock(cell_i%cell_lock))
  end if
 end if
 if(.not. locked) then
  !$omp taskyield
 end if
end do
```

If we just lock cell_i then cell_j, it is possible for deadlock to occur, where three threads are attemping to lock cells k, l, m and each locking a different cell first. This problem is known as the Dining Philosophers problem. Since each cell is assigned a unique integer ID, we attempt to lock the lower numbered cell first, which is the most straightforward solution to the problem. Removing dependencies also allows us to spawn the tasks in parallel, which may lead to a performance improvement.

Results

We implemented 4 variants of the code for DL_POLY_4: i) A method using inout dependencies to control race conditions, with a single thread spawning all of the tasks; ii) A method using locks plus taskyield, with a single thread spawning all of the tasks; iii) A method using locks plus taskyield, using all threads to spawn tasks; iv) A method using locks plus taskyield with task priorities, using all threads to spawn tasks. We ran these variants on Intel Xeon Ivy Bridge (E5-2697 v2) and Intel Xeon Phi Knight's Landing (KNL 7210), with the Intel 17.2.050 and gcc 6.3 runtimes. We used a simple testcase involving only van der Waals forces, though tests with additional short-range electrostatic forces showed similar results.

The scaling and parallel efficiency of the parallel region containing the tasks is shown in Figure 2 with the Intel runtimes on both processors. The parallel region achieves roughly 80% parallel efficiency when using the entirety of both processors, meaning we lose a significant amount of performance due to additional task overheads when compared to the serial version. On the KNL, we compare the performance with different numbers of threads per core, relative to a single core with a single thread. At low core counts, more threads perform better, however when using the entire KNL best performance 2 is achieved with two threads per core, and one thread per core performs better than using three or four.

To be able to see the performance of the parallel region in more detail, we manually instrumented the code. At the start of each task (or after the locks are obtained if using taskyield) the code records the system time using omp_get_wtime(), and records it again at the end of the task region. This allows us to create a timeline showing how the tasks are executed. These are shown on the title page, the upper three graphs for taskyield with parallel spawning, taskyield with serial spawning and dependencies with the Intel runtime.

When using the OpenMP inout dependencies, the amount of possible parallelism is reduced. Firstly, the tasks have to be spawned by a single thread, however most task frameworks struggle with parallel task creation. Secondly, the dependencies enforce an ordering on the tasks that need to be executed. This serialises certain sections of the work. Its possible that by changing the order in which tasks are created that the amount of parallelism will increase. However, we spent minimal time investigating this as the taskyield method performed so much better, and the commutative dependencies suggested in other runtimes seemed more ideal.

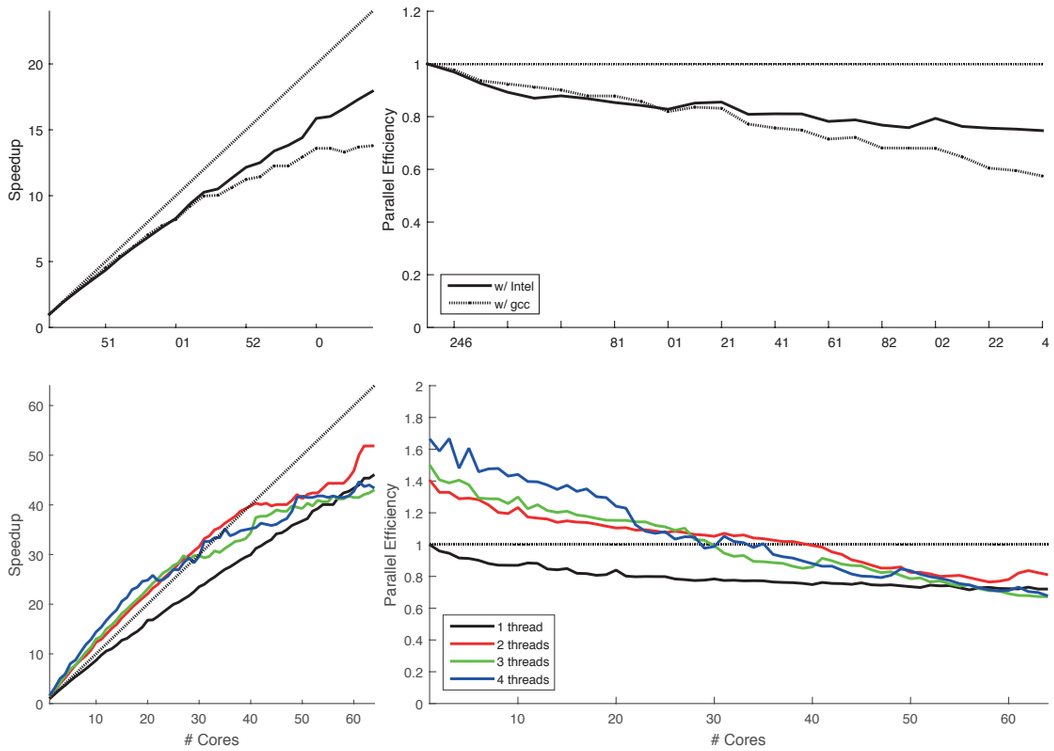


Figure 2: Speedup and Parallel Efficiency of the parallel region on Xeon Ivy Bridge and Knight's Landing. Both lose significant performance due to the overheads associated with the OpenMP task framework. The results with Intel KNL are all shown relative to a single thread on a single core, which means when running with multiple threads per core we can achieve a parallel efficiency of above 1.

With taskyield, the intel runtime significantly outperforms the gcc-6 runtime, shown on the title page last panel. This is due to the way that taskyield is implemented in the two runtimes. In the gcc runtime, the taskyield keyword is ignored, leading to tasks potentially being stuck in a while loop attempting to lock cells, rather than executing other available work. The intel runtime searches for a new task when finding a taskyield, while keeping the yielded task in the executing thread's stack.

In all examples there is significant "white" space throughout the computation, which cannot be entirely due to the methods used to avoid race conditions. When tasks are spawned by a single thread, that thread spends a large amount of time only spawning tasks to be executed. When tasks are spawned in parallel, we can assume that the overall time required to spawn the tasks must be at least as large, which may explain many of the white gaps throughout the computation. It is unclear what causes the threads to spawn tasks vs execute tasks, however this is implementation-defined. From the parallel spawning task plots with Intel and gcc it appears that the Intel runtime spawns relatively few tasks before threads begin executing them, while gcc waits until most of the tasks have been spawned before the majority of threads begin execution. Due to the difference in the taskyield implementation between the two runtimes it is not possible to argue which approach is better from these results, however if the task creation involves memory allocation or other serial operations it may suggest having some threads performing work while others generate tasks may help.

Authors
A. B.G Chalk, A. M.Elena, STFC Daresbury Laboratory

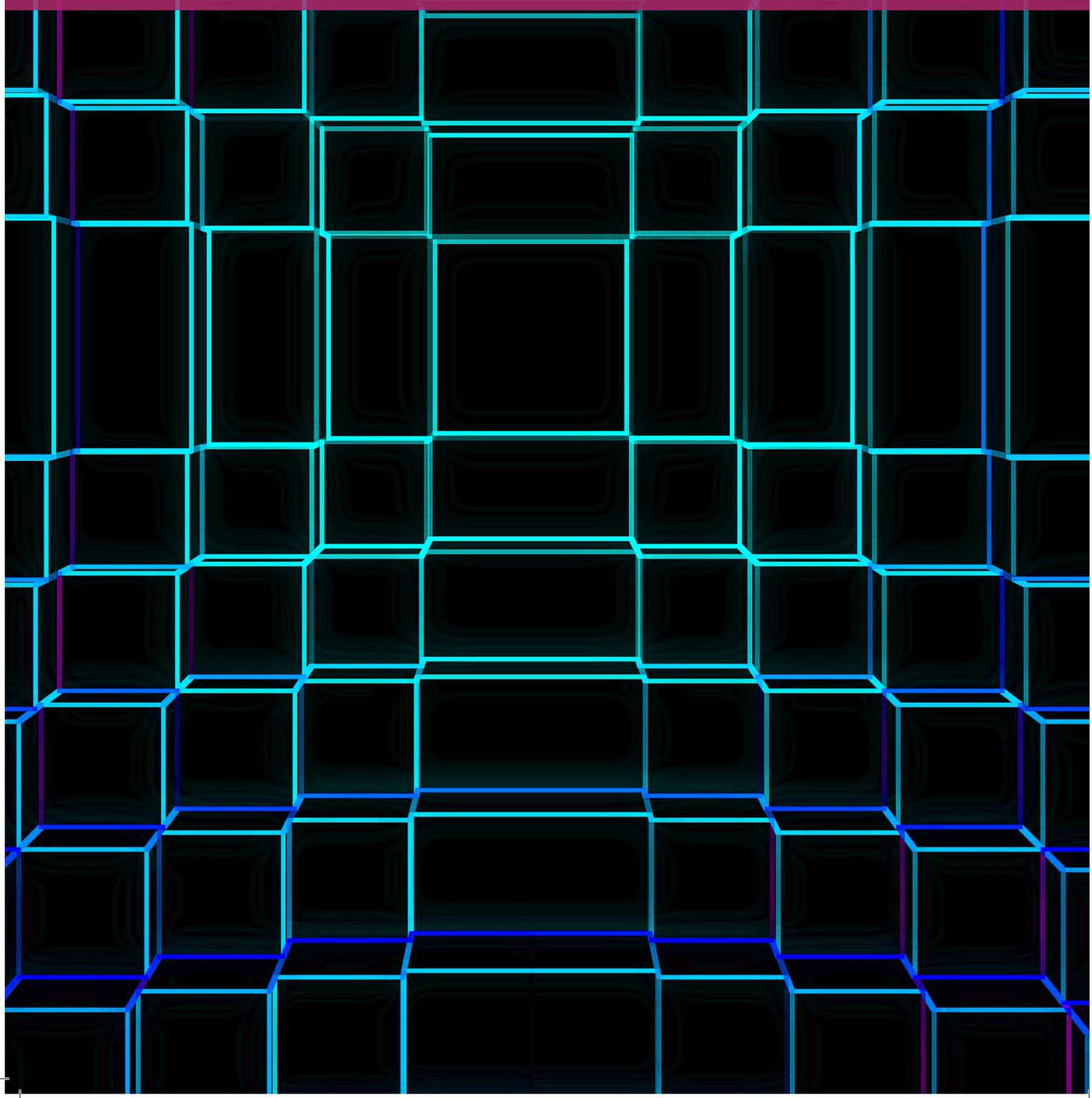
References

[1] Ilian T Todorov, William Smith, Kostya Trachenko, and Martin T Dove. DL_POLY_3: new dimensions in molecular dynamics simulations via massive parallelism. Journal of Materials Chemistry, 16(20):1911–1918, 2006.
[2] B Quentrec and C Brot. New method for searching for neighbors in molecular dynamics computations. Journal of Computational Physics, 13(3):430–432, 1973.
[3] Gary S Grest, Burkhard Dünweg, and Kurt Kremer. Vectorized link cell fortran code for molecular dynamics simulations for a large number of particles. Computer Physics Communications, 55(3):269–285, 1989.
[4] Ralf Meyer. Efficient parallelization of molecular dynamics simulations with short-ranged forces. Journal of Physics: Conference Series, 540(1):012006, 2014.
[5] Pedro Gonnet. Pairwise verlet lists: Combining cell lists and verlet lists to improve memory locality and parallelism. Journal of computational chemistry, 33(1):76–81, 2012.
[6] Pedro Gonnet. A simple algorithm to accelerate the computation of non-bonded interactions in cell-based molecular dynamics simulations. Journal of Computational Chemistry, 28(2):570–573, 2007.
[7] Pedro Gonnet. Pseudo-verlet lists: a new, compact neighbour list representation. Molecular Simulation, 39(9):721–727, 2013.
[8] Hatem Ltaief and Rio Yokota. Data-driven execution of fast multipole methods. Concurrency and Computation: Practice and Experience, 26(11):1935–1946, 2014.
[9] Pedro Gonnet, Matthieu Schaller, Tom Theuns, and Aidan BG Chalk. Swift: Fast algorithms for multi-resolution sph on multi-core architectures. arXiv preprint arXiv:1309.3783, 2013.
[10] Alejandro Duran, Eduard Ayguadé, Rosa M Badia, Jesús Labarta, Luis Martinell, Xavier Martorell, and Judit Planas. OmpSs: a proposal for programming heterogeneous multi-core architectures. Parallel Processing Letters, 21(02):173–193, 2011.
[11] Cédric Augonnet, Samuel Thibault, Raymond Namyst, and Pierre-André Wacrenier. StarPU: a unified platform for task scheduling on heterogeneous multicore architectures. Concurrency and Computation: Practice and Experience, 23(2):187–198, 2011.

Strategic Theme 1

SCD will build and strengthen our existing capability of providing enabling computational expertise and e-infrastructure to support STFC in fulfilling its strategic goals: "world class research, world class innovation, and world class skills."

Exploiting Parallelism in the Solution of Sparse Linear Systems



Large-scale computational problems that have an underlying mathematical structure often arise in science, engineering and industry. Finding the best solution commonly involves constructing a mathematical model to describe the problem and then fitting this model to observed data. Such problems can be difficult to solve but scientists collaborating on the NLAFET Project (Parallel Numerical Linear Algebra for Future Extreme Scale Systems) aim to develop novel algorithms and software that are an improvement, both in speed and accuracy, on what is currently available. They intend to scale these up to work efficiently on extremely-high-performance computers, and will aim to release an open source library of the software implemented during the project.

The solution of large sparse linear systems underpins most large-scale scientific and engineering challenges. Horizon 2020 FET-HPC is supporting work on this through an award for the project NLAFET (Parallel Numerical Linear Algebra for Future Extreme Scale Systems). The project is coordinated by Umeå University in Sweden, with partners at Manchester University, INRIA Paris, and the Computational Mathematics Group (CMG) here at RAL.

The key goals are:

- The development of novel algorithms that expose as much parallelism as possible, exploit heterogeneity, avoid communication bottlenecks, respond to escalating fault rates, and help meet emerging power constraints.
- The exploration of advanced scheduling strategies and runtime systems, focusing on the extreme scale and strong scalability in multi/many-core and heterogeneous environments.
- The design and evaluation of novel strategies and software support for both offline and online autotuning.
- The release of an open source NLAFET library including all the software implemented during the project.

The project covers many aspects of numerical linear algebra, including the solution of equations with dense or full matrices, symmetric and unsymmetric eigenproblems and generalized eigenproblems, and the solution of sparse linear systems using direct, iterative, or hybrid methods. In each case, the intention is to develop algorithms and software that have the potential to scale to extreme scale computers with a peak performance in the high Petaflop or even Exaflop range. The main vehicle for obtaining portability, flexibility, and high performance over a range of architectures is to use runtime systems.

The CMG is focusing on the solution of sparse systems by direct or hybrid methods (methods that combine direct and iterative schemes).

In this article, we briefly illustrate how we can exploit parallelism when solving large sparse linear systems of equations by focusing on symmetric positive-definite systems, and then we discuss numerical pivoting for preserving accuracy of the computed solution in the context of indefinite systems.

Direct methods solve the linear system

$$Ax = b \quad (1)$$

by factorizing A into a product of simpler matrices. For symmetric positive definite problems, it is possible to compute a triangular matrix L using the Cholesky factorization algorithm, such that

$$PAP^T = LL^T \quad (2)$$

where P is a permutation matrix.

Sparse factorization methods divide the work into operations on small dense matrices and we design energy efficient, low-communication dense tasks for use within our sparse factorizations. In our algorithms, we follow the approach taken by earlier work by Hogg, Reid, and Scott [4] for obtaining fine-grained parallelism by using directed acyclic graphs (DAGs). We divide the factorization into subblocks (called tiles) as shown in Figure 1 and then define the dependency of the resulting task-based approach using a DAG. We illustrate such a DAG in Figure 2 where the letters in the circles correspond to the tasks involved in the factorization of the nodes of the tree. The tasks f correspond to the Cholesky factorization of a block corresponding to a tile on the diagonal, the tasks s represent a triangular solve on a subdiagonal block using a factor computed by a task f, the tasks u perform an update of a block within the node, using the blocks created by the kernels s, and the tasks a represent an update between nodes with factorizations at one node updating a block in an ancestor node.

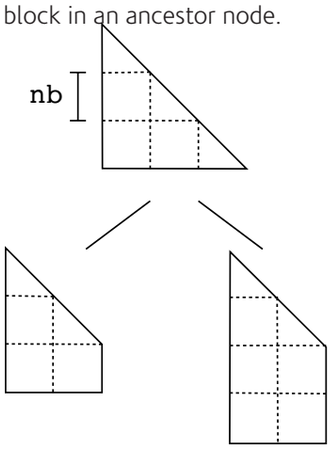


Figure 1: Simple elimination tree with three nodes partitioned into square blocks of size nb.

In order to address the ever-increasing complexity of modern computing platforms equipped with multicore processors, deep memory hierarchy and multiple accelerators such as Graphical Processing Units (GPUs), we use a runtime system to express and execute the DAG on the target architectures. The runtime system can be seen as a software layer between the application and the architecture and

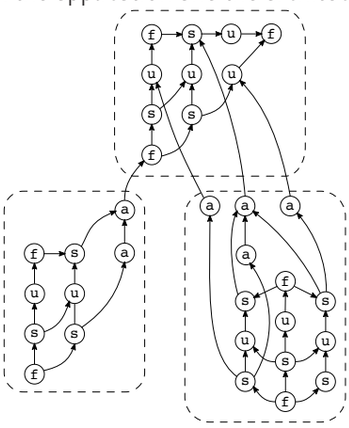


Figure 2: Directed acyclic graph resulting from the block partitioning of the simple elimination tree in Figure 1.

handles the management of task dependencies as well as task scheduling and maintaining data coherency. Although runtime systems are widely used in dense linear algebra, this approach is challenging for sparse algorithms because of the irregularity and variable granularity of the DAGs arising in these systems. Within the framework of NLAFET, we are primarily concerned with the runtime systems StarPU, OpenMP (using task features in Version 4.0 or above) and PaRSEC. Using these runtime systems, we have been investigating two different programming models for implementing the factorization algorithms presented above. In the first model, called a Sequential Task Flow (STF) model, the tasks in the DAG are submitted to the runtime system following the sequential algorithm. It is therefore a very intuitive programming model for implementing parallel codes. However, when the tasks are small, the overhead in setting them up in the runtime system can be high and because the DAG is traversed sequentially some parallelism may be missed. For this reason we also considered the Parametrized Task Graph (PTG) model which is a data-flow programming model that offers a more scalable approach for deploying the DAG on the target architectures. In our implementation we use the StarPU and OpenMP runtime system for implementing STF versions of our solver SpLLT and we use PaRSEC for implementing the PTG variant. In Figure 3 we show that our solver offers competitive performance results compared to the state-of-the-art solver HSL_MA87.

In the case where the matrix A in equation (1) is indefinite then numerical pivoting is needed. A simple example is the matrix

$$\begin{bmatrix} 0 & \times \\ \times & 0 \end{bmatrix}$$

where it is clear that, whatever the value of the entries, a Cholesky factorization will fail because of the zeros on the diagonal. We note that, if x is nonzero, the matrix is nonsingular. The good news is that we can stably factorize an indefinite matrix using only 1x1 and 2x2 pivots [1], [2]. However, pivoting creates major problems for exploiting parallelism. If simple methods are employed then the matrix needs to be updated as pivots are chosen and used. Some parallelism is achievable and is exploited by the HSL code HSL_MA97. Other codes, notably PARDISO [5], capture the parallelism of the positive definite case by not implementing a stable factorization and using an iterative scheme to compensate. Hogg has designed an algorithm that enables us to use block factorization with greater parallelism while maintaining stability [3]. This is used in the SSIDS and NLAFET software. We show the performance of the SSIDS code in Figure 4 on so-called easy problems where little numerical pivoting is required. We see that the performance is about half of that achieved by dense matrix-matrix multiplication on our local Haswell 28-node machine.

In Table I, we examine the performance of three different pivoting strategies on some matrices, characterized as hard indefinite systems, that require significant numerical pivoting.

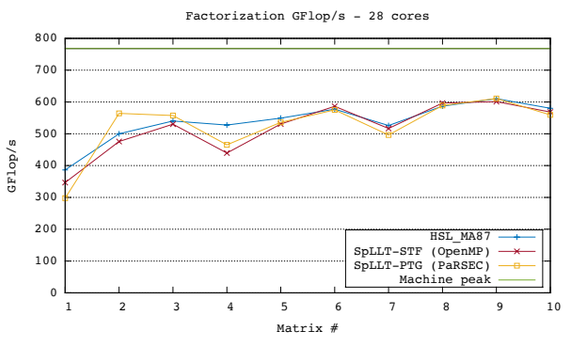


Figure 3: Performance of the sparse Cholesky code SpLLT using OpenMP (STF) and PaRSEC (PTG). These codes are being compared to the state-of-the-art solver MA87 from the HSL library. The practical machine performance peak is 768 GFlop/s.

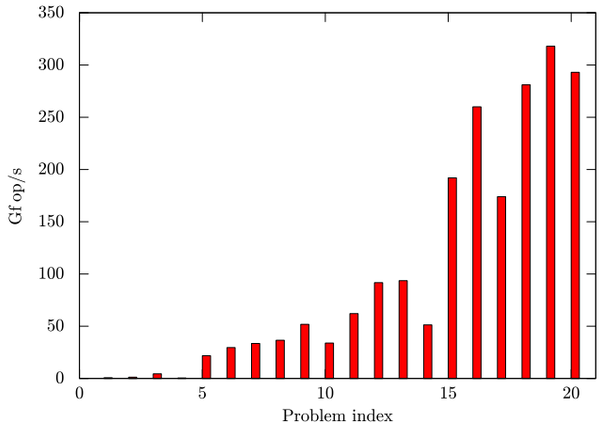


Figure 4: Performance of SSIDS software for easy indefinite matrices on the 28-node Haswell machine.

Matrix	stokes128	cvxqp3	ncvxqp7
Order $\times 10^3$	49.7	17.5	87.5
Entries $\times 10^6$	0.30	0.07	0.31
	Factor time		
HSL_MA97	0.15	1.52	8.18
PARDISO	0.12	0.33	1.50
SSIDS V2	0.11	0.29	1.67
	Backward error		
HSL_MA97	$1.6\ 10^{-15}$	$3.1\ 10^{-11}$	$4.4\ 10^{-9}$
PARDISO	$3.9\ 10^{-3}$	$1.1\ 10^{-6}$	$1.4\ 10^{-7}$
SSIDS V2	$1.4\ 10^{-15}$	$2.0\ 10^{-11}$	$7.3\ 10^{-9}$

TABLE I
HARD INDEFINITE SYSTEMS ON THE 28-CORE HASWELL MACHINE.

Here we see that both SSIDS and PARDISO are significantly faster than HSL_MA97 and, in addition, SSIDS has a much smaller backward error than PARDISO, even though iterative refinement is used with PARDISO. Indeed SSIDS is comparable with the numerical performance of the provably stable pivoting algorithm implemented in HSL_MA97. We parodize these codes as: with HSL_MA97 we have to *pay*, with PARDISO we have to *pray*, while with SSIDS we can *play*.

In summary, there are several levels of parallelism in solving sparse systems that we seamlessly exploit by using DAG-based algorithms. Ensuring the accuracy of the computed solution often impacts the parallelism in the case of indefinite systems and we have utilized a new algorithm for trying to bridge the gap with the positive-definite case. In addition, one reason why we have used runtime systems is so that our software can be easily ported to different architectures and, in the last year of our project, we will port our codes to GPU-based heterogeneous architectures and distributed memory systems.

Authors
I. Duff and F. Lopez, STFC Rutherford Appleton Laboratory

References

[1] J. R. BUNCH AND L. KAUFMAN, Some stable methods for calculating inertia and solving symmetric linear systems, Mathematics of Computation, 31 (1977), pp. 162–179.
[2] J. R. BUNCH AND B. N. PARLETT, Direct methods for solving symmetric indefinite systems of linear equations, SIAM J. Numerical Analysis, 8 (1971), pp. 639–655.
[3] J. HOGG, A new sparse LDLT solver using a posteriori threshold pivoting, Technical Report RAL-TR-2016-017, RutherfordAppleton Laboratory, 2017. NLAFET Working Note 6.
[4] J. HOGG, J. REID, AND J. SCOTT, Design of a multicore sparse Cholesky factorization using DAGs, SIAM J. ScientificComputing, 32 (2010), pp. 3627–3649.
[5] O. SCHENK AND K. GÄRTNER, On fast factorization pivoting methods for sparse symmetric indefinite systems, Electronic Transactions on Numerical Analysis, 23 (2006), pp. 158–179.

Strategic Theme 1

SCD will build and strengthen our existing capability of providing enabling computational expertise and e-infrastructure to support STFC in fulfilling its strategic goals: “world class research, world class innovation, and world class skills.”

SCD Public Engagement Committee

The SCD Public Engagement Committee has continued to meet regularly over the financial year 2016-7 and has made a key contribution to the National Laboratories Public Engagement Programme. In addition to the Daresbury Open Week, which involved 70 members of SCD and saw 8,600 students, teachers and members of the public visit the Daresbury Laboratory, we have also reached over 1600 people over the course of the year at smaller events. New resources and workshops have also been developed, and are now a core part of the public engagement programme.

- SCD played an important role in the very successful Daresbury Open Week in July 2016, which saw 7,500 people visit the campus on a public day and 1,100 school children visit as part of a KS2/3 and 4/5 days. The SCD offerings included demonstrations of our visualisation capabilities to corporate guests and schoolchildren in both age groups, the BeeBot/'Computing without Computers' activity for the KS2/3 day, Arduino/'Build a Temperature Sensor' activity and a masterclass in the chemistry of solar cell materials for the KS4/5 day. On the public day, SCD delivered activities including data centre tours, computer coding for beginners, Lego Mindstorms 'Battle Bots', as well as different versions of the BeeBot and Arduino activities from the schools days. All told, the department ran a total of eleven activities for the general public, all of which proved very popular, such that one of the busiest buildings on campus was A Block, which houses SCD. SCD also provided the NFC scanning stations and the visualization of the tracking information which allowed for the Public Engagement team at DL to track which activities were visited and when.

Children participating in the BeeBot activity on the Daresbury Open Week Public Day

- We once again ran a successful Summer Coding week at RAL. This saw 20 children, aged 8-16, choose a challenge related to data and solve it using the technology of their choice. This year was the third year of running this activity at RAL and was very successful, with excellent feedback from the participants, parents and SCD mentors. Since being bought three years ago, the Mindstorms have become a key part of SCDs outreach programme. A working group is being set up to launch a similar event at DL next summer.

- A Saturday Coding club has continued to run at Rutherford Appleton Laboratory over the 2016-7 year. Once a month, a four hour session is held on site where participants work on projects of their own choosing with help and support from SCD mentors and the NLPE team. The focus of these events is on peer learning, rather than working from a lesson plan. This gives the participants a chance to explore computing for themselves, using the Arduino, Mindstorms and other resources, and builds their confidence. At the end of each club a peer presentation session is run, to which parents and family are invited, giving the participants a chance to show what they have achieved. This also demonstrates the 'soft skills' side of working in the sciences, improving the participants' communication and presentation skills. The club has become steadily more popular, and there is now a waiting list, despite the fact that no direct marketing has taken place. The long-term format of the club also allows SCD mentors (and by extension RAL) to build relationships with the participants and several of them have gone on to do work experience at the lab as a direct result of attending the club. These events take place once a month and reach approximately 20 people per event.

- SCD once again played an important role in the popular work experience programmes at RAL and DL. This year we organised two workshops for Work Experience supervisors to share best practice and help new supervisors develop suitable projects. The RAL workshop was limited to SCD, but following its success the DL workshop was opened up to the whole site. We also ran the three-day computing workshop at RAL again for 15 placement students to give them an introduction to coding. SCD also supervised 15 placements for work experience at RAL and hosted two work experience students at DL this summer from the Social Mobility Foundation. This is a charity which aims to make a practical improvement in social mobility for young people from low-income backgrounds.

- In March 2017, SCD staff gave a presentation at a careers fair at Lymm High School, Warrington. The event was for Key Stage 5 students. As part of that event, there was a focused careers talked aimed at A-level students. Of those in attendance, 40% were girls. A SuperSTEM / SCD Masterclass was run at Daresbury for Key Stage 5 students, which was organised by SCD staff.

In addition to the above activities, the previously established events and workshops carried on as an integral part of the public engagement programme. Throughout the year, there were regular tours of the machine rooms and visualisation suites at RAL, an Ada Lovelace Day using Arduinos, teacher training sessions, Ada Lovelace workshops at school events, and Bee-Bot activities at school and public events, reaching over 700 participants.

The department will be improving on their Public Engagement work in the coming year. For example, to mark international Ada Lovelace Day 2017, both RAL and DL will be co-hosting a joint event with IBM. The event will be the first joint computing outreach event between RAL and DL, the first IBM outreach event hosted at RAL, and will have a focus on encouraging more girls into computing.

Children building Lego MindStorm Robots at Summer Coding 2016

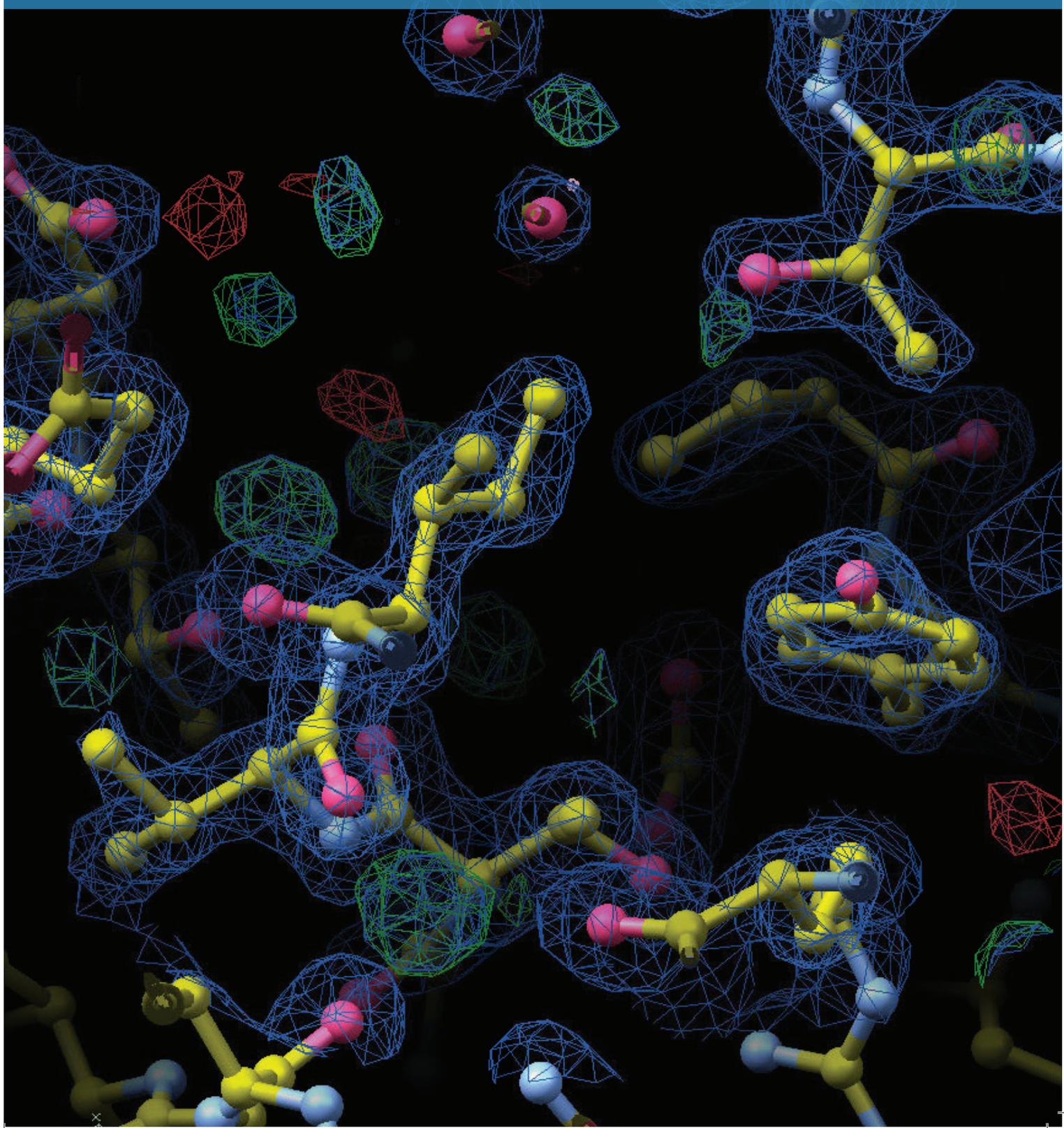
The Summer Coding 2016 Participants

Authors
G.Corbett, STFC Rutherford Appleton Laboratory

Strategic Theme 2

SCD is committed to deliver and further develop its expertise in scientific analysis, data science, modelling and simulation, to support STFC's commitment to: Establish the Ada Lovelace Centre, an integrated, cross disciplinary, data-intensive science centre, to transform the use of real time data processing, computer simulation and data analytics to deliver more effective research at our national facilities.

Data Analysis as a Service



Software underpins the research process: the Software Sustainability Institute 2014 survey [1] of Russell Group [2] Universities showed that 92% of the respondents use research software and 56% of those write their own software. Providing services to enable researchers to use research software and tools to enable better quality software to be written are part of the remit of the Software Engineering Group. The group also aims to reduce the effort required by the facilities in supporting scientists for the analysis process and help increase the scientific output of the facilities.

Data Analysis as a Service

Analysing the scientific data produced by the STFC facilities is becoming more and more complex due to the advancements of the instruments and the scientific techniques. In some cases the volumes of data have grown so large that it is no longer practical for scientists to transport their data back to their home institution. In other cases, the analysis requires access to high performance computing and complex chains of software where these resources and access to the necessary expertise may not be available. All of this technical complexity is being exposed to the scientists, who are quite often not computing experts, and is resulting in the analysis process becoming a bottleneck.

The Scientific Computing Department is working with the facilities based at the Rutherford Appleton Laboratory (ISIS Neutron and Muon Source, the Central Laser Facility and the Diamond Light Source) to implement and deploy a 'Data Analysis as a Service' system (DAaaS) to support scientists during the analysis process. This service provides scientists with easy to use access to compute resources, collocated with the experimental data archives, to efficiently and easily process their data, within a managed, secure virtual environment. Commonly used software packages are systematically made available via a deployment and configuration system, and the environment offered to users is customised according to the nature of the experiment and requirements of the experimental team.

This service aims to reduce the technical complexity of the analysis process by simplifying user workflows and automating complex processes – for instance automating the movement of large volumes of data between computing resources. From this, it will reduce the effort required by the facilities in supporting scientists for the analysis process and help increase the scientific output of the facilities.

One of the leading use cases for DAaaS is the ISIS Data Analysis as a Service project which provides users from multiple communities with all the necessary software, data and compute resources they require to complete their analysis process.

Automated testing of research software

The EPSRC funded Software Engineering Support Centre provides an automated testing facility to the academic community with a focus on the Collaborative Computational Projects. We undertook a survey at the end of 2016 to establish the Continuous Integration practices in the academic community and potential appetite to use a central service. It underpins our planning for the SESC Build Service beyond the current funding and provides an interesting and useful view on the usage of CI and the interest in CI for HPC testing. Full report available from http://purl.org/net/epubs/work/33360356 .

Key points from the survey were:

- 75% of all respondents are already using CI systems
- 81% of those who identified themselves as software developers/RSEs are already using CI systems
- A significant minority of respondents use more than one CI system
- All of the seven UK Research Councils are represented, with EPSRC being the biggest funder
- Python, Fortran, C and C++ are the most common languages reported
- Travis and Jenkins are the most common CI frameworks in use
- Compilers and specific libraries are the most important tools needed for automated testing
- There is an interest in HPC testing, and a smaller interest in novel architectures
- Shibboleth and OpenID are of equal importance as authentication mechanisms

Bringing the two together

One of our aims for the coming year is to enable research software using the automated testing service to be automatically deployed to the DAaaS analysis environments . This will allow developers to release their up to date and well tested software to a large number of users and allow scientists to access the latest software with minimal effort.

Authors
C. M. Jones, F. Barnsley (DAaaS), S. Lamerton (SESC), STFC Rutherford Appleton Laboratory

References

[1] https://www.software.ac.uk/blog/2016-09-12-its-impossible-conduct-research-without-software-say-7-out-10-uk-researchers The data collected during this survey is available for download from Zenodo ("S.J. Hettrick et al, UK Research Software Survey 2014", DOI:10.5281/zenodo.14809). It is licensed under a Creative Commons by Attribution licence (attribution to The University of Edinburgh on behalf of the Software Sustainability Institute

[2] The Russell Group comprises the following universities – Birmingham, Bristol, Cambridge, Cardiff, Durham, Edinburgh, Exeter, Glasgow, Imperial College London, King's College London, Leeds, Liverpool, London School of Economics and Political Science, Manchester, Newcastle, Nottingham, Oxford, Queen Mary University of London, Queen's University Belfast, Sheffield, Southampton, University College London, Warwick and York.

Strategic Theme 2

SCD is committed to deliver and further develop its expertise in scientific analysis, data science, modelling and simulation, to support STFC's commitment to: Establish the Ada Lovelace Centre, an integrated, cross disciplinary, data-intensive science centre, to transform the use of real time data processing, computer simulation and data analytics to deliver more effective research at our national facilities.

Building Better Tools for Tomography

The Visual Analytics and Imaging Systems group supports two Collaborative Computational Projects (CCPs) for tomography (imaging by sections) in material science and medical imaging. Through the work at these CCPs we are building capacity in advanced software infrastructure for the material science and medical tomography communities.

The Collaborative Computational Project in Tomographic Imaging (CCPi) aims to provide the UK tomography community with a toolbox of algorithms that increases the quality and level of information that can be extracted by computer tomography (neutron or x-ray tomography).

The Collaborative Computational Project in PET/MR aims to exploit the recent integration of positron emission tomography (PET) and magnetic resonance (MR) imaging into a single simultaneous imaging system. The aim is to provide enabling infrastructure for research in PET-MR by development and promotion of a common software framework to tackle the specific challenges of this imaging modality. This will be achieved by standardisation of data formats, creating tools to export data in these data formats and development of a software platform for integrated PET-MR image reconstruction.

Core Imaging Library

The CCPi Core Imaging Library (CIL) is a set of modules for each process involved in the data analysis workflow for Computed Tomography datasets: image pre-processing, reconstruction, quantification, segmentation and visualization.

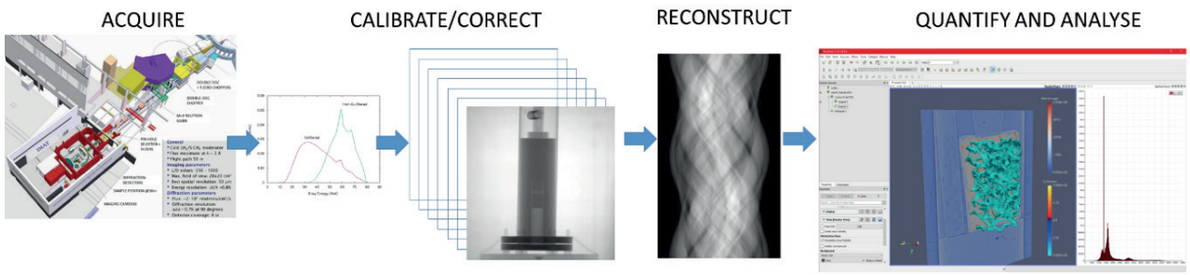


Figure 1. A Computed Tomography pipeline

The algorithms are contributed by the CCPi community and are engineered to make them scalable, optimized, accessible, maintainable and documented. While the core of the library is developed in C++, our aim is to give programmers and scientists more flexible ways to explore and integrate these algorithms into high level languages, such as Python.

Currently there are 5 available modules:

- Pre-processing: Beam hardening[1]
- Iterative reconstruction algorithms[2] (CGLS, SIRT, MLEM and CGLS with 3 regularization methods) for parallel beam
- Topological Segmentation based on Contour Tree [3]
- Quantification: Accessible Volume[4] and Label Quantification[5] algorithms
- Interactive Viewer for 3D volumes and surfaces

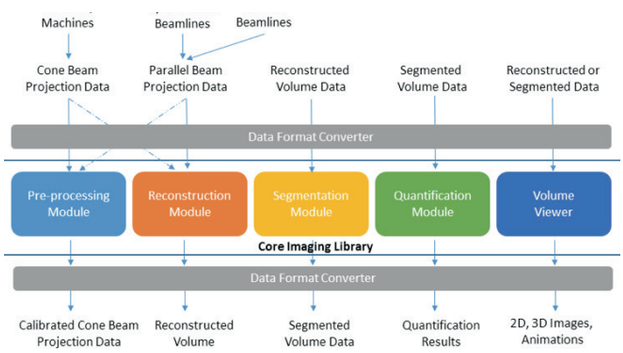


Figure 2. Available modules in the Core Imaging Library and how they fit the CT pipeline

The CCPi Core Imaging Library is distributed as a Python module via the CCPi conda channel https://anaconda.org/ccpi. To increase software usage and user engagement there is a mailing list and a documentation website with explanation and examples. Additionally, plugins for a number of image analysis software platforms such as Paraview, Avizo and Fiji are distributed.

The Core Imaging Library was launched in June 2017 with a bi-yearly stable release plan.

Synergistic Image Reconstruction Framework

The Synergistic Image Reconstruction Framework (SIRF) software is an Open Source toolkit for the reconstruction of PET and MRI raw data. The aim is to provide code simple enough to easily perform a reconstruction, yet powerful enough to be able to handle real, full-size datasets. Our strategy in achieving this aim is to employ available Open Source reconstruction software written in advanced programming languages such as C++ and provide basic-user-friendly interfaces to it written in script languages, primarily Matlab and Python. The code builds upon existing Open Source software packages for medical image reconstruction. Currently, SIRF uses STIR for PET reconstruction and Gadgetron for MRI.

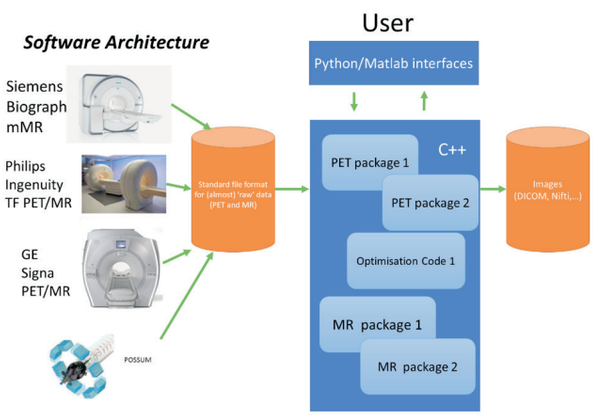


Figure 3. Synergistic Image Reconstruction Software architecture.

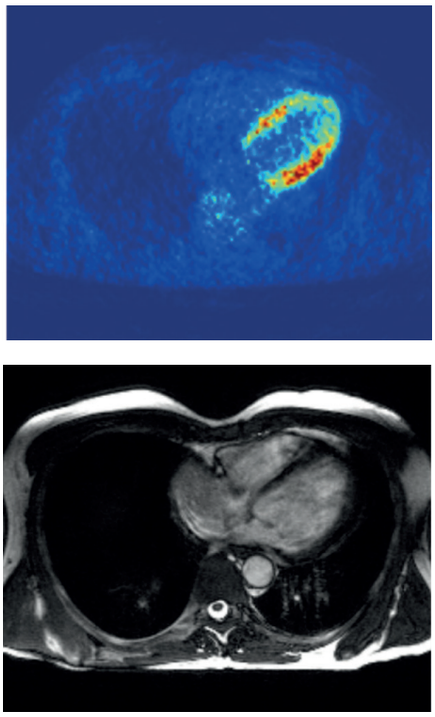


Figure 4. Example images of cardiac patient data acquired with Siemens mMR reconstructed with SIRF. Data: Kolbitsch et al., Fully integrated 3D High-Resolution Multicontrast Abdominal PET-MR with High Scan Efficiency, MRM 2017.

A pre-release of the Synergistic Image Reconstruction Framework was launched at the PSMR2017 conference followed by the first release 0.9.0 on the 14th of June 2017.

Summary

The Visual Analytics and Imaging Systems (VAIS) group has contributed to the development and distribution of two major advanced software frameworks for the analysis, quantification and visualization of tomographic data in the field of material science and medical imaging. With the release of the Core Imaging Library and Synergistic Image Reconstruction Framework VAIS delivered and is committed to further develop advanced software solutions for the handling, analysis, visualization and interpretation of experimental data for more effective research at our national facilities.

Authors
E. Yang, STFC Rutherford Appleton Laboratory

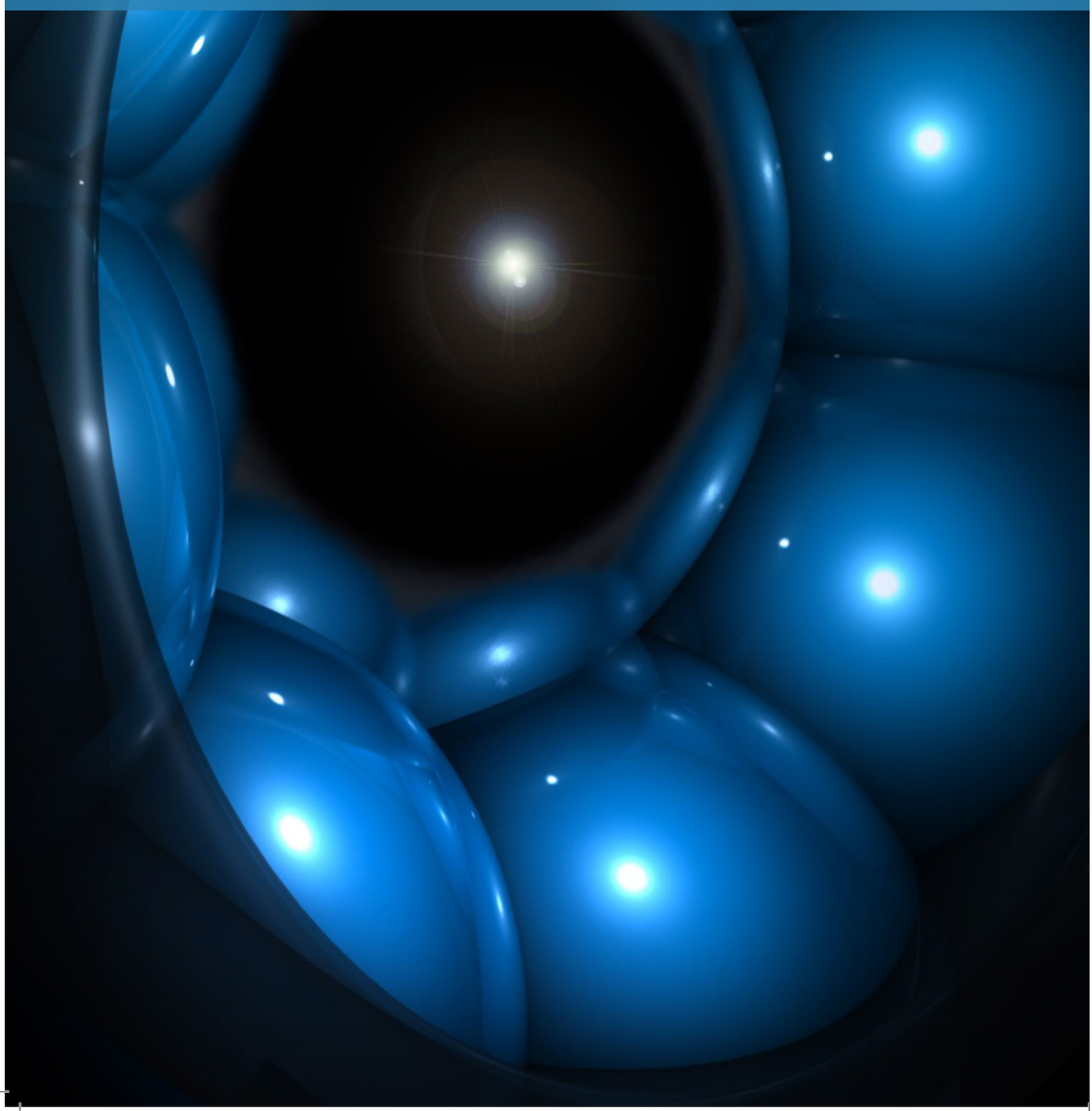
References

[1] "Quantitative high contrast X-ray microtomography for dental research," J. Dent., vol. 41, no. 5, pp. 475–482, May 2013.
[2] M. Soleimani and T. Pengpen, "Introduction: a brief overview of iterative algorithms in X-ray computed tomography.," Philos. Trans. A. Math. Phys. Eng. Sci., vol. 373, no. 2043, Jun. 2015.
[3] H. Carr, J. Snoeyink, and U. Axen, "Computing contour trees in all dimensions," Comput. Geom. Theory Appl., vol. 24, no. 2, pp. 75–94, 2003.
[4] "Additive manufactured porous titanium structures: Through-process quantification of pore and strut networks," J. Mater. Process. Technol., vol. 214, no. 11, pp. 2706–2715, Nov. 2014.
[5] "Evaluation of 3-D bioactive glass scaffolds dissolution in a perfusion flow system with X-ray microtomography," Acta Biomater., vol. 7, no. 6, pp. 2637–2643, Jun. 2011.

Strategic Theme 2

SCD is committed to deliver and further develop its expertise in scientific analysis, data science, modelling and simulation, to support STFC's commitment to: Establish the Ada Lovelace Centre, an integrated, cross disciplinary, data-intensive science centre, to transform the use of real time data processing, computer simulation and data analytics to deliver more effective research at our national facilities.

Computational Modelling to help Super-Microscopes



SCD is helping researchers carrying out experiments with the ISIS Neutron and Muon Source to understand what is happening with samples at a microscopic level. Our scientists have developed computer simulations to help interpret the results of these experiments and reduce the number of uncertainties, and have established ways to automate the work. For example, when performing muon spectroscopy analysis, the so-called "stopping site" for the muon traditionally had to be guessed or searched by trial and error. (Understanding precisely the way that the muons - which are short-lived elementary particles fired at the sample - interact with the sample's crystal cell structure, is key to interpreting the outcome.) For this task a new method has been developed, combining machine learning and numerical simulations. This method allows the researcher to sift through hundreds of crystal structures far more efficiently and quickly. These new computational techniques can be used for studies with a variety of materials – including molecular magnets (which have potential medical applications) and pharmaceuticals.

The STFC Rutherford Appleton Laboratory is a world-class centre for research that houses the ISIS Neutron and Muon Sources, which produce beams of neutrons and muons that can be used to study materials at the atomic level.

One of the projects that we are currently involved in is centered on the computational modelling of experiments performed with muons, which are subatomic particles produced by bombarding a graphite target with pulses of high-energy protons that originate in a synchrotron. The infrastructure required for these types of experiments includes a massive particle accelerator, tunnels for directing the particle beams and large buildings for housing the muon targets and all their corresponding scientific equipment. The name "Super Microscope" arises from the combination of this massive infrastructure with muon beams that can probe the local magnetic and physical structures of materials with an unprecedented level of accuracy and resolution.

In our project, we are using computer simulation to model muon-spin relaxation (μ+ SR) experiments. These experiments involve implanting spin-polarized positive muons in samples of materials that range from crystalline solids to organic molecules. μ+SR can, for instance, be used to study hydrogen defects in a sample, probe a sample's local magnetic structure or study the organic radicals that may result from adding muons to an organic sample.

The muon is an extremely sensitive probe, but it also has some significant limitations. One of them is that the site of implantation of the muon in a sample is usually unknown, which is something that may, for instance, hinder the measurement of magnetic moments using μ+SR. Moreover, as the mass of the muon is almost ten times smaller than that of hydrogen, the quantum effects associated to the muon have a massive effect on some of the properties measured using μ+SR, such as the hyperfine coupling constant (HFCC) of the muon.

In this project we are using a combination of first principles calculations, machine learning techniques, ab initio random structure searching (AIRSS) and other computational tools to address the limitations of μ+SR that we mentioned above. In the sections below we describe in some detail two concrete examples where these computational techniques have been applied.

Temperature Dependence of the HFCC of Muonium and its Corresponding ipso-Hydrogen in Crystalline Benzene[1].

Longitudinal field Avoided Level Crossing (ALC) is an especially suitable technique for the study of organic radicals[2]. The ALC spectrum can be used to estimate the values of HFCCs and, in particular, the ALC spectrum of crystalline benzene is found to exhibit multiple complex features that strongly depend on temperature.

Benzene, (C_6H_6), is solid under 278.6 K. It forms a crystal with four C_6H_6 molecules per unit cell with space group *Pbca*. Inside the crystal muonium will bond with a benzene ring, breaking one of the double bonds between carbons and thus leaving one unpaired electron, which

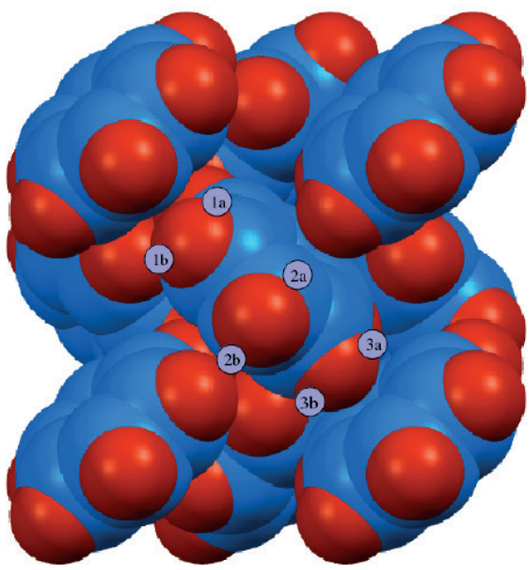


Figure 1. Crystalline structure of benzene with the 6 possible muonium bonding sites labeled.

couples via hyperfine interaction to both the muon and the Hydrogens in the molecule. The hydrogen that is bonded to the same carbon as the muon is called the 'ipso hydrogen'. Its role in determining the final ALC spectrum is especially important, as two transitions can give rise to ALC lines - the 1-quantum muon spin flip (called the Δ1 line) and the 0-quantum muon-ipso hydrogen flip-flop (or Δ0 line).

While in an isolated C_6H_6 molecule all possible bonding sites would be equivalent, in a crystal there are six inequivalent positions, which can be seen in Figure 1. These sites have been labeled with numbers and letters to remark the symmetry relationships between them.

In this work we focused on the experimental ALC spectrum for benzene at 12 K and at 210 K, which we considered as the low and high temperature limits for the main dynamical processes affecting the ALC spectrum. The experimental data was acquired at the muon beamline in ISIS, RAL.

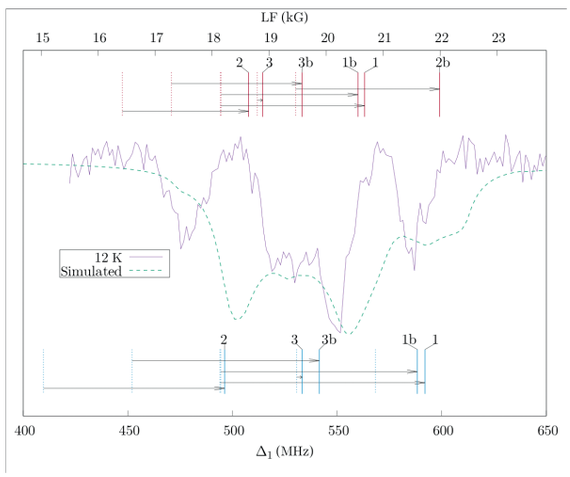


Figure 2. Experimental ALC spectrum for solid benzene at 12 K (solid line) and static simulated spectrum (dashed). The red lines on the top of the plot represent the predicted positions for the Δ1 peak at the various sites, before (dashed) and after (solid) the quantum corrections. The ones at the bottom are the same for the Δ0 peaks.

At 12 K, the ALC spectrum of benzene displays a number of broad peaks, partially overlapping across the explored range of frequencies. Figure 2 displays the spectrum, together with the prediction of the simulation. The simulation was performed by taking the six corrected hyperfine tensors for the six possible muon and ipso-Hydrogen sites, calculated using our computational techniques.

The resulting spectrum matches the experimental one closely enough except for the leftmost peak, related to site 2a, which seems to be too shifted to the right by an excessive quantum correction. Furthermore, the peaks appear to be broader than the experimental ones.

Regarding the high-temperature value of 210K, it was chosen because the most relevant temperature-activated process for the ALC spectrum in benzene is the axial rotation of benzene rings, which only needs a relatively small amount of thermal energy to be activated. The temperature at which this rotational process becomes relevant on the time scale of a muon's lifetime is around 130 K. Hence, the 210 K spectrum represents a suitable case to study the high temperature limit.

The spectra simulated with our computational methodology is compared with the experimental one in Figure 3.

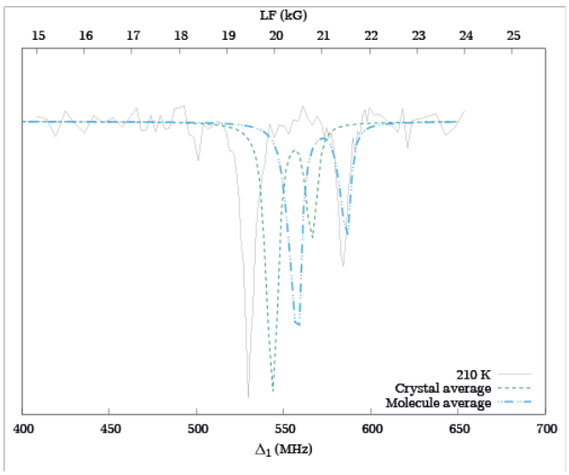


Figure 3. Experimental ALC spectrum for solid benzene at 210 K (solid line), and simulated spectra with averaged crystal tensors (dashed line) and single molecule tensors (dash and dot).

Two peaks are visible in each spectrum; the large ones on the left are the Δ1 peaks, whereas the smaller ones on the right are the Δ0. The simulations capture the narrowing process but seem to miss the proper peak position; only the Δ0 peak from the single molecule seems to match the experiment perfectly. This suggests that the role of the crystal is more complex than either of these simplifying assumptions takes it to be.

Overall, our methodology can be used to calculate the ALC spectrum at different temperatures for an organic radical. The method gives predictions that are reasonably close to the experimental results, taking into account corrections for quantum and classical dynamical effects. Ongoing work is aimed towards refining this method and eventually delivering tools that would enable a larger audience of users to apply it in a general way as an aid to the interpretation of measured ALC spectra for this class of compounds.

Computational Prediction of Muon Stopping Sites in Semiconductors[3].

There are some experimental techniques that can determine the muon stopping sites, but these techniques are limited to certain specific cases. For instance, in some materials, the determination of the muon stopping site was possible by using accurate experimental studies of the muon frequency shift in an applied magnetic field[4]. Nonetheless, the number of examples where the muon site can be determined by experimental means alone is limited, and combinations of experiments and calculations have been applied to this problem with mixed success[5].

Hence, we believe that an alternative method is needed to find the muon stopping site, and we have developed a method that combines machine learning with a computational technique known as ab initio random structure searching (AIRSS)[6]. This method focuses on the paramagnetic states formed by muons in semiconductors, and we have applied it to the case of muons in pure Si, Ge and Diamond.

The method first applies AIRSS to generate a number of Si, Diamond and Ge super-cell structures where the muon is placed in random positions, and then uses machine learning techniques to classify these structures combining energetic and geometric parameters. This classification is motivated by our interest in all potential muon stopping sites and not just the lowest energy one. Intuitively, if suitable parameters are chosen for the classification, all final configurations that represent random fluctuations around a specific stopping site will look far more similar to each other than to those around a different site, and will form a cluster. This intuition is what our method automates in a way that allow us to sift through hundreds of candidate structures far more efficiently and quickly than any human would. Our technique is implemented in the Python library Soprano, which has been developed with funding from the Computational Collaborative Project for NMR Crystallography and is licensed under the GNU LGPL[7].

Figure (4) shows the clusters that our method obtained for Si, Diamond and Germanium. The clusters are represented by circles whose diameter indicates the number of structures contained in each cluster. Muons cluster due to the actions of the attraction basins formed around the potential energy surfaces' local minima of each cluster. These attraction basins tend to have a hypervolume, in the space of the potential energy surface, whose size is related to the size of the energy minima. A small standard deviation for the average energy of a cluster is an indication of a consistent cluster, with structures that are more similar to each other than the structures around a site with a larger standard deviation. Hence, sites with small standard deviations are likely to be representing a local energy minimum and, consequently, a muon stopping site. Furthermore, as our methodology is based on DFT calculations that are performed at 0K, these stopping sites are likely to be low-temperature stopping sites.

In Figure (4), the largest standard deviation is ≈0.05 eV, which indicates that all of the clustering sites are potential muon stopping. Experimental evidence suggests that, at low temperatures (T≤100K), the system cannot be in thermal equilibrium between all possible sites and charge states and, as muonium centres are never formed in sufficient concentration, the muonium centres never equilibrate among themselves. Under these conditions, our purely theoretical method predicted the experimentally well-known "bond centered" (MuBC) and "tetragonal" (MuT) stopping sites of muonium in Si, Diamond and Ge.

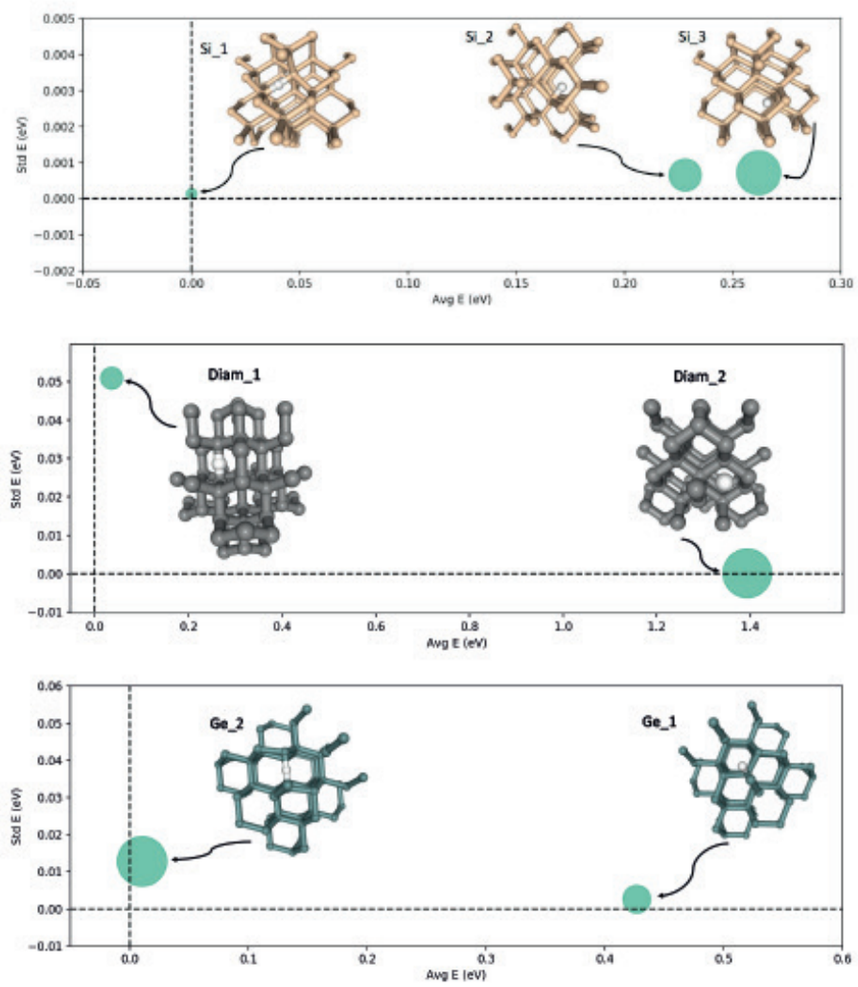


Figure 4. Clusters obtained for muonated Si, Diamond and Ge. The x coordinate of the centre of each one of the circles indicates the relative average energy of the corresponding cluster, while the y coordinate of the centre indicates the standard deviation of the average energy of that cluster.

Authors
L. Liborio, STFC Rutherford Appleton Laboratory

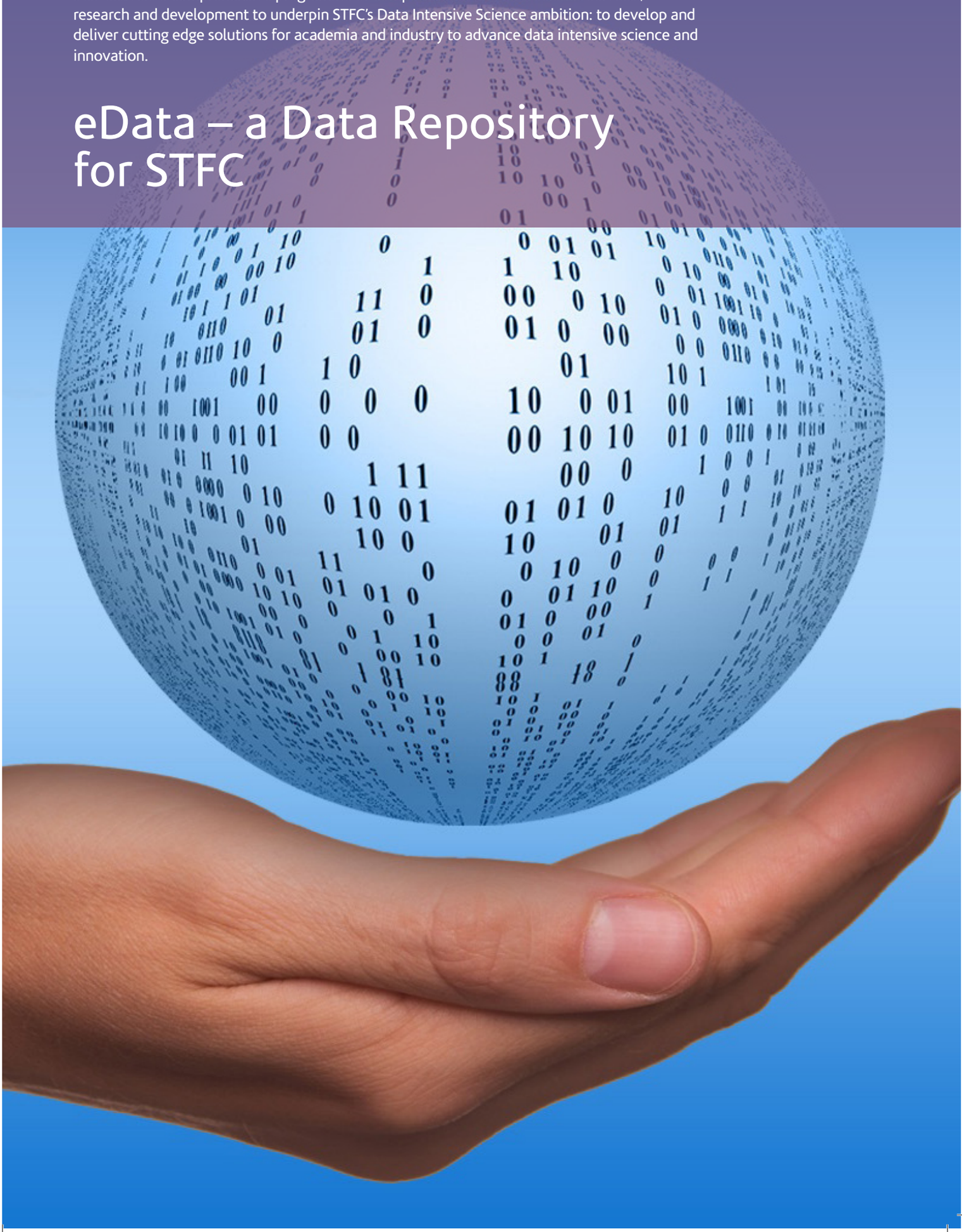
References

[1] Exploring the Temperature Dependent Solid-State ALC Spectrum of the C6H6Mu• Radical with Ab-Initio Simulation Techniques; S. Sturniolo, L. Liborio et.al. This work was presented at the 14th International Conference on Muon Spin Rotation (http://jmeson.org/musr2017/index.html). A short paper based on this work has been accepted for publication at the Journal of Physical Society of Japan: Conference Proceedings.
[2] S.F.J. Cox, G.H. Eaton and J.E. Magraw, Hyperfine Interactions 65, 773, (1991).
[3] Computational Prediction of Muon Stopping Sites, L. Liborio, S. Sturniolo, et.al. Part of this work was presented at the 14th International Conference on Muon Spin Rotation (http://jmeson.org/musr2017/index.html). A paper based on this work is in preparation.
[4] R. De Renzi, et. al., Phys. Rev. B, 30(1):186, 1984.
[5] M. Bimbi, et. al., Phys. Rev. B, 77, 045115, 2008.
[6] A.J.Morris, et. al., Phys. Rev. B, 78, 184102, 2008.
[7] S. Sturniolo. Soprano - a library to crack crystals. https://ccpforge.cse.rl.ac.uk/gf/ project/soprano/.

Strategic Theme 3

SCD delivers a comprehensive programme of computational and data science services, research and development to underpin STFC's Data Intensive Science ambition: to develop and deliver cutting edge solutions for academia and industry to advance data intensive science and innovation.

eData – a Data Repository for STFC



In line with its goal to define and implement open science, SCD has developed eData, an institutional repository where STFC staff can deposit data and software that underpin journal articles and other published research.

Open Science

SCD has national and international influence in defining and implementing UK, EU and Global policy towards open science – "the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods."[1]

SCD also supports the dissemination of scientific outputs for STFC, including literature, data and software, making them discoverable, accessible and useable worldwide, in the spirit of open science.

Open Data

Research data are the evidence that underpins the answer to the research question, and can be used to validate findings regardless of its form. The primary purpose of research data is to provide the information necessary to support or validate a research project's observations, findings or outputs. Open research data are those research data that can be freely accessed, used, modified, and shared, provided that there is appropriate acknowledgement if required.[2]

eData

In line with its goal to define and implement open science, SCD has developed eData, an institutional repository where STFC staff can deposit data and software that underpin journal articles and other published research. It offers a service for the deposit, discovery, sharing, preservation, and citation of such data.

eData offers a range of open access licences, links to authors' ORCID accounts, related publications, funding information, embargo management, and provides citations for datasets. A suite of policy and guidance documents is available to help creators deposit their datasets.

eData is managed by a cross-department team comprised of information management specialists from the Chadwick & RAL Library service and developers from the Software Engineering Group, who use their expertise to provide a 'one-stop shop' research data service to STFC staff.

We have chosen the DSpace platform for eData, which provides customisable, free open source software with a large community of users and developers so we can learn from and share our experiences with our peers in academia. We use this network of peers, and engagement with eData users, for ongoing review and development of our research data services in the spirit of open science.

Authors
D. Franks, Chadwick & RAL Libraries, C. Jones. Software Engineering Group Leader, STFC Rutherford Appleton Laboratory

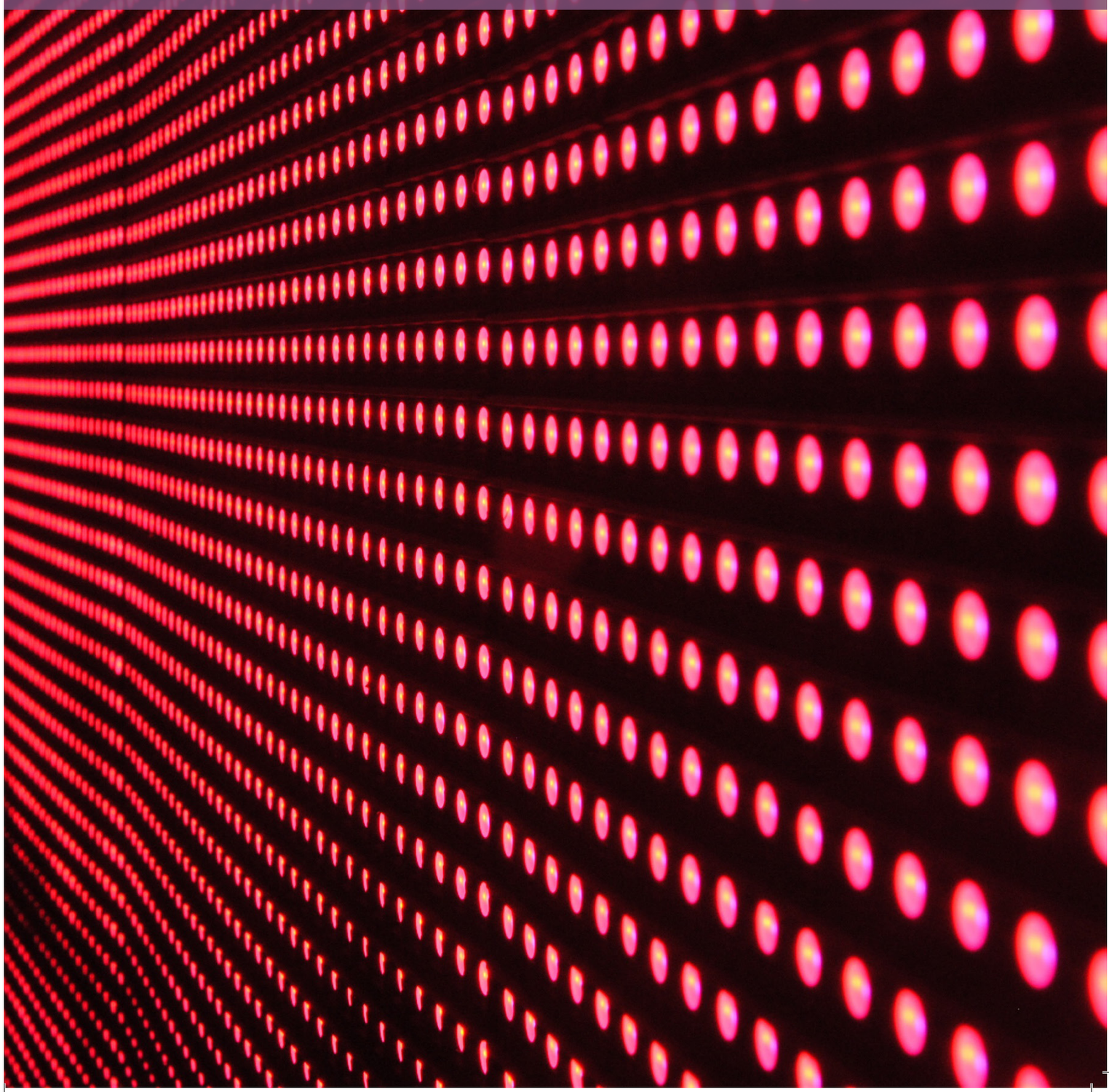
References

[1] Open Science definition. FOSTER www.fosteropenscience.eu/foster-taxonomy/open-science-definition
[2] Definitions. Concordat on Open Research Data. RCUK www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf/

Strategic Theme 3

SCD delivers a comprehensive programme of computational and data science services, research and development to underpin STFC's Data Intensive Science ambition: to develop and deliver cutting edge solutions for academia and industry to advance data intensive science and innovation.

ADDoPT – Advanced Digital Design of Pharmaceutical Therapeutics



SCD provides vital science and software expertise to enable the STFC Hartree Centre to fulfil its mission to: "Transform the competitiveness of UK industry by accelerating the adoption of data-centric computing, big data and cognitive technologies". SCD's computational expertise and support helps to reduce laboratory time and costs spent on research and development.

The STFC Hartree Centre is a partner in ADDoPT (Advanced Digital Design of Pharmaceutical Therapeutics), a four-year £20.4m project which aims to secure the UK's position at the forefront of pharmaceutical development and manufacture by establishing robust manufacturing for current and next generation medicines built on digital design.

This collaboration between government, industry and academia is tackling the challenge of delivering new, higher quality medicines to patients, faster and more cost effectively.

This will be achieved through the development of advanced digital design techniques that eliminate non-viable drug candidate formulations as early as possible, streamlining design, development and manufacturing processes.

Industrial pharmaceutical project partners Pfizer, GSK, AstraZeneca and Bristol-Myers Squibb have provided data to which we can apply STFC's expertise in supercomputing, big data analytics and computational chemistry. The aim is to develop more accurate digital design techniques that will enable more of the early R&D process to be carried out virtually, saving time and improving cost efficiency.

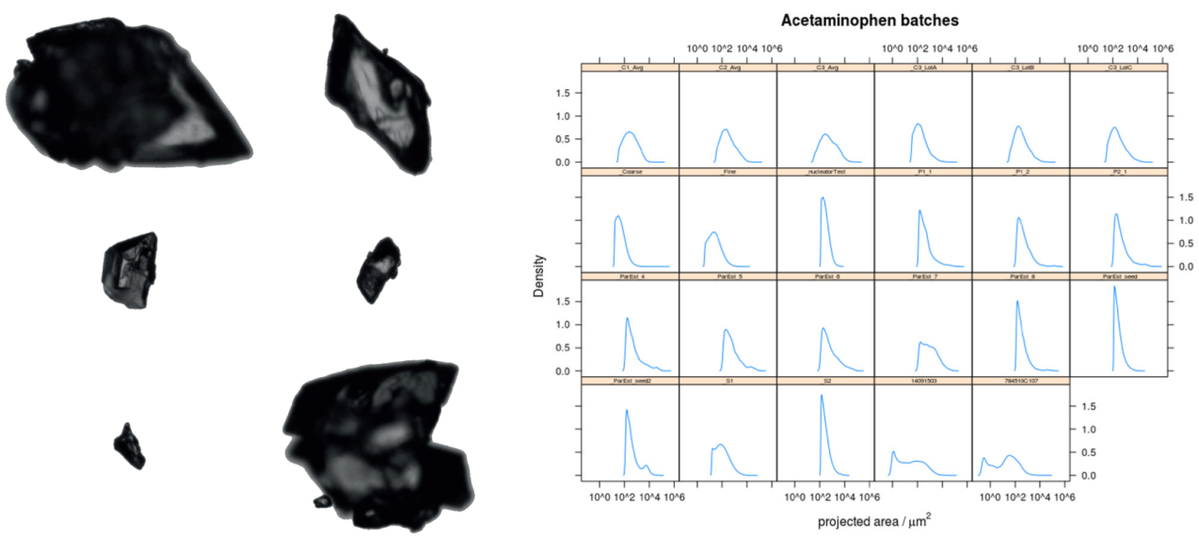


Figure 1. Left: Six illustrative optical microscopy images of powder particles of acetaminophen (paracetamol) from one batch, showed at the same scale. Right: density plots of particle projected area (closely related to particle size) for 23 batches of acetaminophen, showing the range of particle size distributions obtained in pharmaceutical development laboratories.

ADDOPT
ADVANCED DIGITAL DESIGN OF PHARMACEUTICAL THERAPEUTICS

Hervé Barjat in the Computational Biology Group is working on statistical modelling of active pharmaceutical ingredients powders in order to predict their suitability for producing tablets. The standard practice in industry is to measure size and shape descriptors for particles in each powder sample using a range of analytical methods, in order to summarise the distributions of these descriptors using median and first and last deciles. Using the same descriptors, but a more comprehensive summary of their distributions for each powder sample, Hervé is testing the value of kernel methods to see if they can lead to better prediction of bulk powder properties, such as the ability of the powder to flow during the manufacturing process. Particle size and shape data on many powder samples, together with their associated bulk properties, are being gathered from our industrial partners. While we will receive more data in the coming year, we now have enough to proceed to the modelling stage.

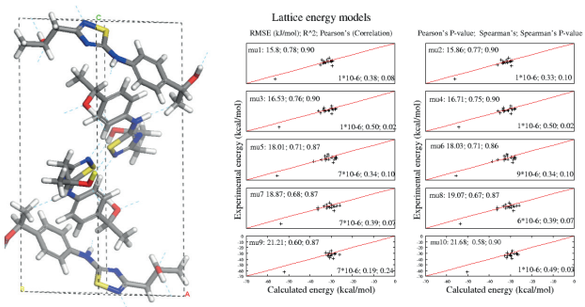


Figure 2. A molecular crystal of a derivative of the drug molecule thiadiazole; lattice energy calculations using force fields where mu1 to mu10 represent different models accounting for variations in parameter settings. The red line represents a perfect match between calculated and experimentally-derived lattice energies.

Dawn Geatches and Rebecca Mackenzie in SCD's Computational Chemistry and Computational Biology Group respectively, are working with ADDoPT partners at the University of Leeds to predict the lattice energy of crystals. The computational chemistry work focuses on calculating lattice energies of atomistic structures where the molecular crystals are known, and the life sciences work builds statistical models to predict lattice energies where no crystal structure is available.

Statistically calculated lattice energies are an important analogue for enthalpy of sublimation values where crystal structures are unavailable. Statistical models are built using lattice energy values obtained from experimental enthalpies of sublimation, or, if they are unknown, from force field calculations of a set of molecular crystals. The experimental enthalpies of sublimation or force field values form a 'training set' for the statistical model, which is then applied to a new set of molecules whose crystal structures are unknown. The new set of molecules must be of the same or similar classification as the training set, i.e. 'drug' or 'drug-like'.

The statistical model is only as good as its training set, whose values originate from either experiment or atomistic calculations. Where quantum mechanical methods such as Density Functional Theory (DFT) might be assumed to be the Gold Standard in atomistic calculations, and more accurate than force field calculations, DFT proved itself inadequate for large, molecular organic crystals, due to its inability to correctly model long-range dispersion forces that dominate these pharmaceutically-relevant structures. To address this deficit, Dawn coordinated a successful application to ARCHER's eCSE programme (embedded Computational Science & Engineering programme), together with colleagues at STFC (Dominik Jochym), and the Universities of York (Phil Hasnip, Matt Probert), Leeds (Ian Rosbottom), and Dublin City (Anthony Reilly). Peter Byrne (STFC) has now been employed to carry out dispersion force enhancements to CASTEP[1].

Switching to classical force fields, atomistic, lattice energy calculations have continued, the quality of which was checked by exploring several parameters - see Figure 2. The final output is approximately 60,000 lattice energy calculations, ready to be fed upstream to the statistical modelling work.

Chin Yong, also of the Computational Chemistry Group, is meanwhile working with ADDoPT researchers at Leeds on enhancing DL_FIELD and DL_ANALYSER. These codes were written by Chin to interface with DL_POLY[2], a popular molecular dynamics code, and are now being adapted to capture analytical detail pertinent to the pharmaceutical industry.

Rebecca Mackenzie is building on this work to develop a statistical model which predicts lattice energy even before a crystal structure is available.The training set currently consists of a number of drug-like molecules (3000), for which each molecule's probability of being 'drug-like' was determined using a Naïve-Bayes classifier (Figure 3).

Using Quantitative Structure Property Relationship analysis, in which around 200 molecular descriptors are calculated for each crystal, a combined model using a kernel (similarity) function and a gradient boosted regression tree has resulted in predicted crystal lattice energies being close to those determined via calculated methods when compared to experimental values (Figure 4). This work is ongoing, with models being updated as new training lattice energy values are calculated.

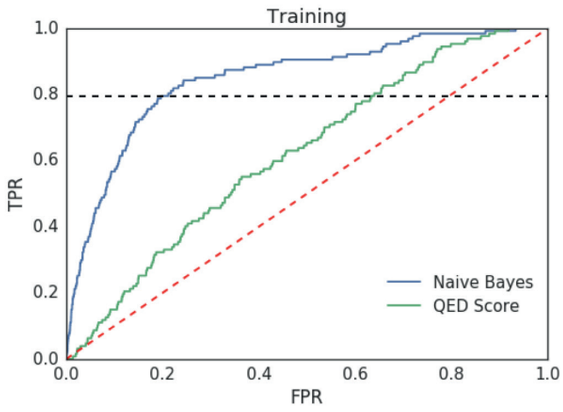


Figure 3. Naive Bayes Classifier for drug-likeness is able to predict 'drugs to be drugs' more often, outperforming a published quantitative estimate of drug-likeness (QED) model and a random model. TPR – true positive rate, FPR – false positive rate.

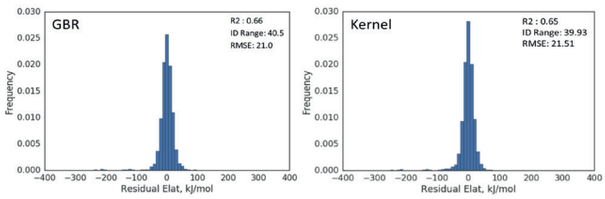


Figure 4. A combination of a GBR and Kernel model can be used to predict calculated lattice energy. On an individual basis, the root mean squared error (RMSE) is comparable to the error between calculated and experimental lattice energies. When combined in a linear model, these scores are improved. GBR – Gradient boosted regression tree using molecular descriptors; Kernel – similarity function assessing the similarity of training and test molecules according to their molecular fingerprint; R2 – R squared coefficient of determination; ID Range – interdecile range; elat – lattice energy.

Rebecca is also working with Andrew Gargett of the Hartree Centre on text mining of chemical papers, with the intent of extracting relevant data (descriptive or numerical) to a query using a web application programming interface (Figure 5). Using STFC's Web of Science (WoS) licence, users submit a WoS query, with the interface returning the most relevant papers, whilst also giving an insight into the returned information as a whole, through analysis of the most frequent terms in the available abstracts. Work is currently ongoing to model and thus identify articles containing relevant data, such that reports of physical constants may also be generated without the user spending excess time performing a search.

The Hartree Centre has provided computational resource to ADDoPT partners including lead partner Process Systems Enterprise (PSE), the Cambridge Crystallographic Data Centre, and the University of Leeds. We have submitted another grant application with PSE which would fund the APE group (Application Performance Engineering) to parallelise some simulation code, and have begun discussing a follow on grant for ADDoPT.

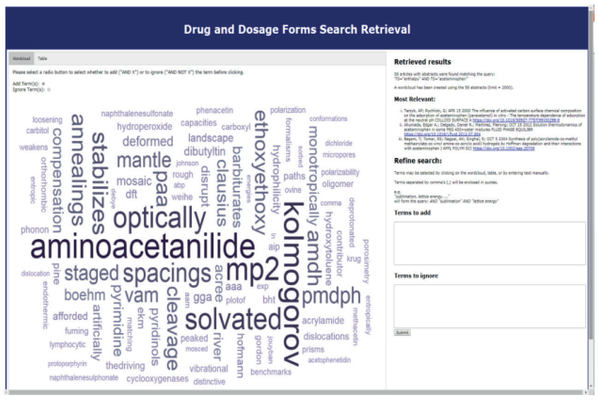


Figure 5. Results retrieved searching WoS for 'TS=Acetaminophen'. The web application programming interface assesses the term-frequency, inverse document frequency (tf-idf) of all terms present in the abstracts of articles returned from the query. Those with large tf-idf are deemed informative and are displayed in a word cloud figure to give insight into the returned information as a whole.

In early 2018, we will organise a training course in Machine Learning for Cheminformatics at Daresbury.

ADDoPT is funded under the Advanced Manufacturing Supply Chain Initiative (AMSCI). Adrian Toland, Hartree Centre, participated in writing the grant application. Chris Morris is STFC's PI for the grant.

Authors
H. Barjat, D. Geatches, R. Mackenzie, and C. Morris, STFC Daresbury Laboratory

References

[1] Clark, S. J.; Segall, M. D.; Pickard, C. J.; Hasnip, P. J.; Probert, M. J.; Refson, K.; Payne, M. C. Zeitschrift für Kristallographie 2005, 220(5-):567-570
[2] I.T. Todorov, W. Smith, K. Trachenko & M.T. Dove, Journal of Materials Chemistry, (2006) 16, 1911-1918 http://www.ccp5.ac.uk/DL_POLY/

Strategic Theme 3

SCD delivers a comprehensive programme of computational and data science services, research and development to underpin STFC's Data Intensive Science ambition: to develop and deliver cutting edge solutions for academia and industry to advance data intensive science and innovation.

ECHO: A Ceph Cluster to Meet Future LHC Data Requirements

The new Ceph-backed storage service, "Echo", is now in production at the Rutherford Appleton Laboratory (RAL) Tier-1, the system which will meet our future Large Hadron Collider (LHC) data storage commitments. The LHC has been producing greater volumes of data for processing on the Worldwide LHC Computing Grid, of which the RAL Tier-1 is a partner, and the previous storage system, CASTOR, could not scale to meet future storage and computing needs. Echo provides the additional storage required for the increased data needs for the LHC, whilst also being economical and easier to maintain.

As part of the Tier-1 computing service for GridPP, the Data Services Group has been running the CASTOR (CERN Advanced STORage) system for the storage of LHC data since before the start of LHC Run 1. CASTOR had also been used at other sites, including CERN and the Academia Sinica Grid Computing Centre, however in the past five years both had switched to alternative systems for their disk storage to meet the increased data taking requirements for LHC Run 2 and beyond. In 2013 the RAL Tier-1 team began their project to identify an alternative to meet the need to scale to the multi-petabyte level. Alternative systems run at other sites were considered and assessed and the result of this investigation was to take a more in-depth view of Ceph[1]. The key requirements for the new storage cluster were that the hardware cost per usable TB of storage should be kept in line with CASTOR and it must scale to meet the demands of the LHC Run 3. In 2014 a prototype Ceph cluster was installed as a potential replacement for CASTOR disk storage and test systems installed in 2015. The outcome of these tests was promising and, in July 2016, the Ceph cluster, Echo, was brought into existence to provide the new storage and has now been running as a production service since March, 2017.

The Echo Cluster

The name Echo comes from the key design principles behind the cluster: Erasure coded, Ceph, High throughput, Object store. Ceph provides an open-source, distributed, object storage solution which can be run on economical commodity hardware.

A Ceph cluster stores data as objects on storage nodes made up of Object Storage Daemons (OSDs), typically one OSD per disk, and these OSDs send information about their state to the Ceph monitors. The Ceph monitors maintain maps of the state of the cluster, including information on state changes. Where an object is stored in the cluster is calculated using the CRUSH (Controlled Replication Under Scalable Hashing) data distribution algorithm. The Ceph object store, RADOS (Reliable Autonomic Distributed Object Store), can be accessed via the Ceph RADOS gateway in addition to industry-standard S3 and Swift compatible API.

The Echo cluster at RAL is made up of 5 physical monitors with 62 storage nodes each with 36 x 6TB disks. The overall raw storage capacity is currently 13.4 PB. There are also five gateway servers providing access to the storage cluster for external traffic. In order to meet the needs of the LHC experiments, access to the storage via gateways supporting the XRootD and GridFTP protocols is also in place although there is significant interest in supporting S3/Swift and considerable development work is being put towards achieving this.

Erasure Coding

One of the key specifications when building Echo was that the hardware cost per TB of usable storage needed to be kept in line with CASTOR.

The objects in a Ceph cluster are stored in logical groups, referred to as pools and these may be replicated pools or erasure coded pools. In a replicated pool more than one copy, or replica, of an object is stored. Erasure coded pools may be used instead of replicated pools to save space. Erasure coding splits an object into ***k*** chunks and then writes a total of *n* chunks (where $k < n$) of data to the storage. The object can be re-created from any k chunks. n is known as the set size. The difference, $m = n - k$, is the number of chunks that can be lost before data integrity is compromised. The n disks that a particular piece of data is written to is referred to as its placement group. Ceph storage pools are made up of large numbers of placement groups so as to spread the data evenly across the storage nodes. It should be noted that any particular disk will be part of many different placement groups. For the ~13PB of raw storage in Echo the available storage for a 3-way replicated cluster would be around 3.3PB usable space given there is some additional overhead. It is immediately apparent that using replicated pools would not provide anywhere near the same cost per TB of usable storage. Erasure coding was therefore the only choice.

In CASTOR storage nodes were normally 24 or 36 bay machines with a RAID6 or RAID60 configuration. The RAID configuration normally used on storage nodes in CASTOR would be equivalent to $k = 16$, $m = 2$. Table 1 shows what percentage of raw storage capacity is actually usable under different erasure coding profiles:

	k=8	k=10	k=12	k=14	k=16
m=2	80	83	86	88	89
m=3	73	77	80	83	84
m=4	67	71	75	78	80

Table 1. Table showing the percentage of usable storage with different Erasure Coding profiles.

One of the specifications for Echo was that it needed to be easier to maintain than CASTOR. With RAID6, single disk failures are easily dealt with, however double disk failures, which happen on average a little over once a month, will trigger a call-out to the support team to minimise the risk of data loss. The number of placement groups in Echo is already an order of magnitude higher than the number of RAID arrays in CASTOR so the chance of two random disk failures being part of the same placement groups is significantly higher. By increasing *m* to 3, even if two disks in an acting set fail in quick succession, there is still resilience against a further disk failure and the need for call-outs is reduced. Initially a *k = 16*, *m = 3* erasure coding profile was chosen, however it quickly became obvious that this wouldn't work as erasure coded pools require more resources than replicated pools. The disks in each placement group found it very difficult to remain synced with each other causing the cluster to become unstable; the number of messages sent is proportional to the square of the set size. The vast majority of known production Ceph clusters use 3 replicas of the data which has a set size of three. There is ongoing work by the Ceph developers to improve this, however it was decided to reduce k to 8 in the Echo configuration. This was primarily chosen because another use case was known to be running in production with these settings; Yahoo run a modified version of Ceph to store their data and use a *k = 8*, *m = 3* erasure coded profile [2]. The default jerasure erasure code plugin is being used with the Reed, Solomon, Vandermonde algorithm[3].

CRUSH

To access data stored in a Ceph cluster, all that is required is access to the CRUSH map as well as a keyring. The CRUSH Map describes the layout of the storage cluster. The CRUSH ruleset is a method to select which disks make up a placement group. The CRUSH ruleset that has been chosen for Echo is to require every disk in a placement group to be on a separate host. This means that the loss of an entire storage node will not cause any data to be lost. In future, this should mean that a hardware problem on a single storage node will no longer need to trigger a call out. It also means that rolling upgrades of the cluster can happen in a manner that is transparent to the users.

Conclusion

Bringing the new Ceph-backed storage service, Echo, into production has been an exciting challenge for the group. In addition to deploying Echo with a reliable, supportable configuration the group have developed, in collaboration with the other teams providing Tier-1 services, monitoring and alerting tools and operating procedures to maintain the cluster in a healthy state and ensure the integrity and availability of the data.

Authors
A. Packer, A. Dewhurst, I. Johnson, G. Vasilakakos, T. Byrne, B. Canning, STFC Rutherford Appleton Laboratory

References

[1] R. Appleyard, S. de Witt, J, Adams First Experiences with Ceph on the WLCG ISGC, 2013
[2] Yahoo Ceph Storage https://www.nextplatform.com/2015/04/16/inside-the-ceph-exascale-storage-at-yahoo/ accessed Oct 2017.
[3] J.Plank, K.Greenan 2014 Jerasure: A Library in C Facilitating Erasure Coding for Storage Applications Technical Report UT-EECS-14-721 University of Tennessee

Strategic Theme 4

In order to ensure that the UK's e-infrastructure supports the country's leading international research status and delivers the data capabilities essential for academia, industry and STFC's science program , SCD will collaborate with our Research Council UK partners and stakeholders, and with Innovate UK, and JISC, as part of the emerging UK Research and Innovation (UKRI), to shape and build an efficient and effective UK e-infrastructure for UK science and industry.

From Project to Production: How the Research Projects we participate in Enhance the SCD Infrastructure

As scientists produce larger and larger volumes of data it is essential to be able to see the impact of that data, and be able to decide how best to curate it. Our Distributed Computing Infrastructure Group contributes to several large European collaborations which will expand the computational, communication and skills capabilities of science communities, as well as developing ways to realise the benefits offered by cloud computing technologies, storage and networks. Here we look at some of that in more detail.

Scientific Computing on the International Stage

Over the last year the Distributed Computing Infrastructure Group DCIG) has contributed a large amount of effort towards the EGI-Engage and INDIGO DataCloud projects. Both of these European collaborations receive their funding through the Horizon 2020 programme and work towards developing the European computing infrastructure. DCIG has contributed development effort to both EGI-Engage and INDIGO through the production of the GOCBD and APEL services, as well as through work testing deployments of services produced as part of INDIGO. The outputs from these projects can then be utilised and fed back into the department to improve our own services and systems.

A summary of the aims of the two projects is provided below.

- The EGI-Engage project (Engaging the Research Community towards an Open Science Commons) aims to accelerate the implementation of the Open Science Commons by expanding the capabilities of a European backbone of federated services for compute, storage, data, communication, knowledge and expertise, complementing community-specific capabilities.
- INDIGO DataCloud intends to develop an open source data and computing platform targeted at scientific communities, deployable on multiple hardware and provisioned over hybrid, private or public, e-infrastructures. By filling existing gaps in PaaS and SaaS levels, INDIGO-DataCloud will help developers, resources providers, e-infrastructures and scientific communities to overcome current challenges in the Cloud computing, storage and network areas.

Research Involvement (INDIGO/Engage) – deployment and development work.

APEL

APEL is an accounting tool that collects computing resource usage data from over 300 sites participating in the EGI infrastructure – Figure 1 shows a map of these sites. The central Accounting Repository receives around 3 million records a day, which it then processes to generate statistical summaries that are available through the EGI Accounting Portal. Over the last year, these records have provided data on over 600 million individual grid jobs consuming a total of almost 3 billion CPU hours - Figure 2 displays the upwards trend of usage from February 2009 through to August 2017.

Figure 1: A map from the GOCDB showing, some of the sites participating in the EGI infrastructure

As part of EGI-Engage, a prototype APEL system has been created to account for the usage of datasets. A dataset in this context is a logical set of files which may exist in several places at once and to which it is possible to assign some form of persistent unique identifier. Datasets are an increasingly important resource to account for – as scientists produce larger and larger volumes of data, it is essential to be able to see the impact of that data and to be able to decide how best to curate it. This prototype aims to assist with this, presenting site and experiment administrators with information about the location and storage of datasets so that the use of infrastructures may be more efficient, as well as enabling scientists to assess the impact of their work. This work will be developed further during the next main H2020 funded project EOSC-Hub and will be part of a contribution to SeaDataCloud.

As part of this work, the APEL team added a new method of interacting with usage accounting data by enabling the software to retrieve usage metrics via an HTTP based API.

The APEL team also developed a REST interface for the Indigo DataCloud project.

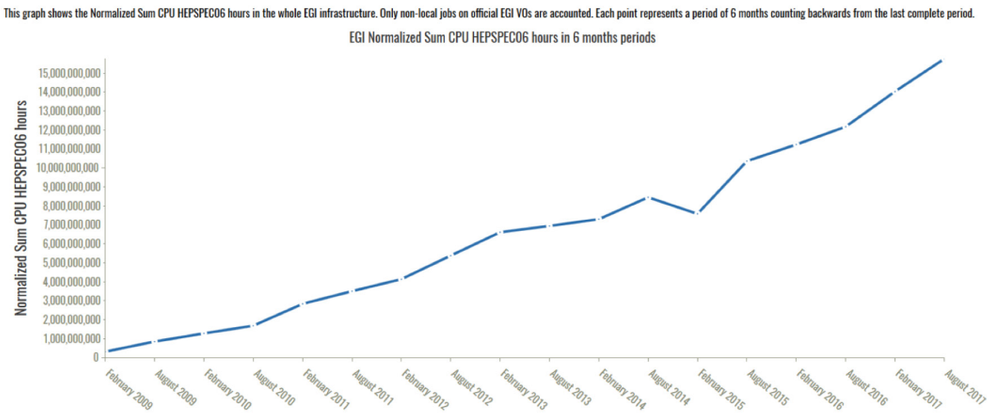


Figure 2: A graph from the EGI Accounting Portal showing the number of CPU hours consumed continues to increase

The development of a REST interface will provide the opportunity for accounting information to be sent directly to an APEL Server. This could later be used for an in-house APEL instance used by other services in R89, without the need for involving the EGI message brokers and for the accounting of other infrastructures separate to EGI.

The REST experience has also been useful for dataset accounting, which interacts with OneData instances via REST interfaces, and the development of the APEL messaging software to interact with the new HTTP based message broker.

Docker images of the APEL REST interface and APEL server were also created. Doing this highlighted a number of potential future issues in the code base early enough to not affect the instance ran as a service for EGI.

INDIGO DataCloud Testing

Alongside the development of APEL (and GOCDB undertaken as part of INDIGO and EGI, effort has also contributed towards pilot testing of some of the core INDIGO services. The Infrastructure Manager (IM) and the Identity and Access Manager (IAM) are two of the larger INDIGO software components and both offered promising use cases within SCD. These components were both deployed and tested using the SCD Cloud system and their functionality tested. These deployments were used to establish whether they would be useful within the established SCD infrastructure.

The Infrastructure Manager aims to provide a set of functions to enable the configuration of virtual infrastructures to enable the deployment of an application or service within a cloud environment. An instance of the IM was deployed for the SCD Cloud OpenStack and its configuration allows a user to be able to pass a VM infrastructure definition to the IM. The IM then provisions virtual machines to match the configuration requested. Whilst testing of the IM's functionality found it to be unsuitable for its intended use case, details about issues found and suggestions for improvements to the software and its documentation were communicated to its development team.

The IAM provides authentication layer managing where it identities and manages access to distributed resources and services in a homogenous way. The IAM service provides user identity and policy information to connected services to make consistent authorization decisions across distributed services. The configuration of a test instance of the IAM is currently underway, with the intention of potentially using this within the systems division to provide identity and access control across the range of services provided.

Alongside testing of two of the core Indigo components, other smaller INDIGO services with potential use cases – primarily for implementation to the SCD cloud – were passed to the relevant department members to investigate their potential.

Enhancing the SCD Infrastructure

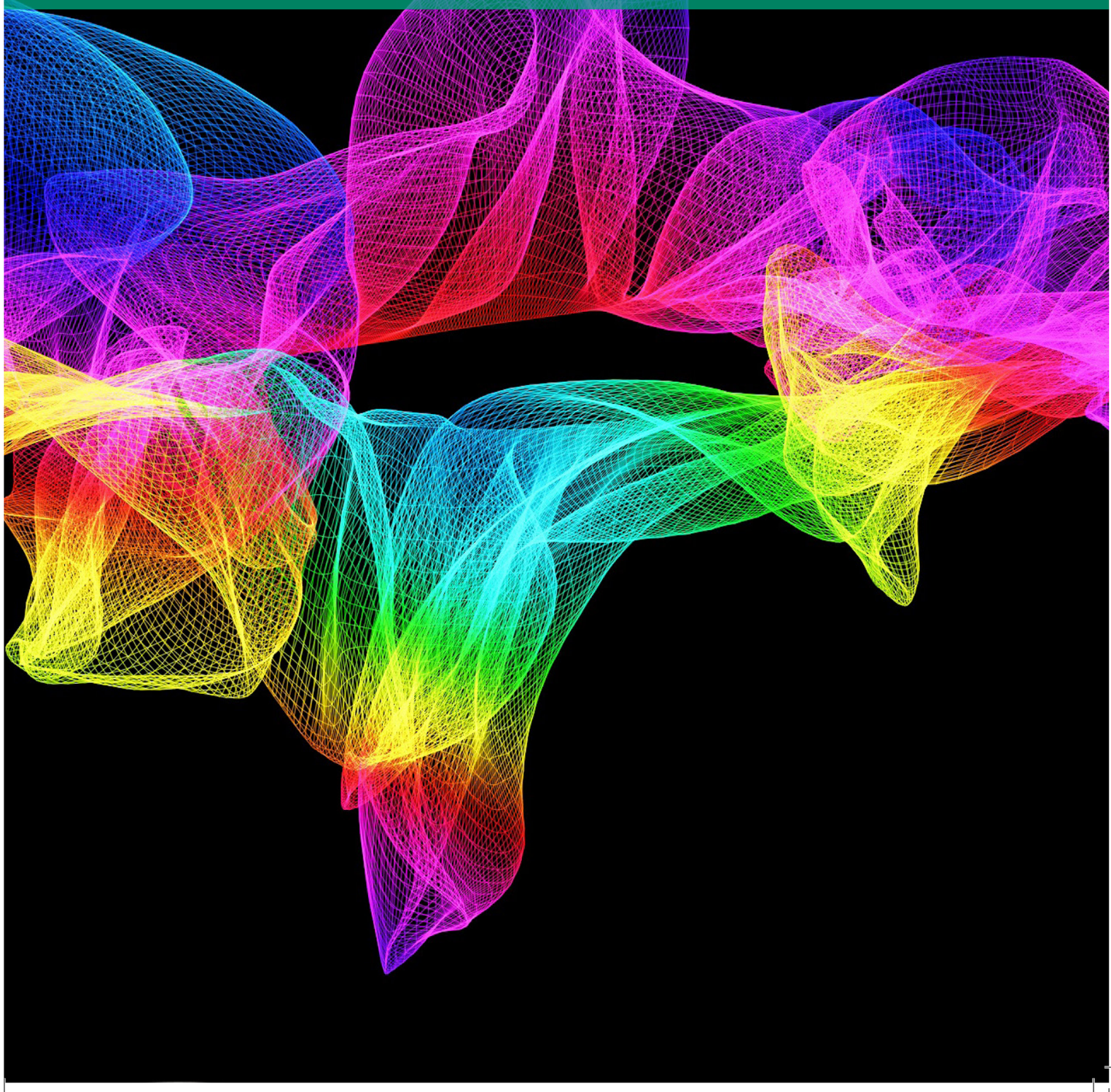
Collaborative projects such as EGI Engage and Indigo DataCloud enable the enhancement of the SCD infrastructure through the development of applicable software products. Large components developed by other institutions, such as the IAM or IM, can provide potential implementation solutions within the department, whilst the work undertaken within the department on projects such as APEL and GOCDB can find use by other services or projects within the department – such as EOSC-Hub and SeaDataCloud.

Author
I. Collier, T. Dack, STFC Rutherford Appleton Laboratory

Strategic Theme 4

In order to ensure that the UK's e-infrastructure supports the country's leading international research status and delivers the data capabilities essential for academia, industry and STFC's science program, SCD will collaborate with our Research Council UK partners and stakeholders, and with Innovate UK, and JISC, as part of the emerging UK Research and Innovation (UKRI), to shape and build an efficient and effective UK e-infrastructure for UK science and industry.

Towards the European Open Science Cloud



Scientists in Europe are, together, the largest producers of scientific data in the world. However, insufficient and fragmented support for sharing and studying that data means this 'big data' is not being exploited to its full potential. The European Commission has a vision to create a new European Open Science Cloud (EOSC) that will offer Europe's 1.7 million researchers and 70 million science and technology professionals a place to store, share and re-use their data across disciplines and borders. The Scientific Computing Department is leading the pilot project to set the future shape of the EOSC, how it might work and how it might be run. The project will explore how the EOSC can provide the basis for better scientific research, make more efficient use of vital resources and improve every area of science and innovation.

STFC's Scientific Computing Department has a long history of working across Europe to support international science. The long standing collaboration with CERN and other international partners to provide the World-Wide LHC Computing Grid (WLCG) has been a prime example of how computing is delivered across borders so that scientists can work together to access the scientific data which has been generated at the experiments at CERN. SCD also works within the European projects: the European Grid Initiative (EGI); and the EUDAT collaborative data infrastructure. These initiatives provide access to data and computers to support researchers to collaborate across a variety of disciplines.

These "e-infrastructure" projects are working to support research in particular areas. However, there is no generally available computing infrastructure to support any research collaboration. And this is a missed opportunity; if easy-to-use and economic tools and services were widely available to researchers across Europe and beyond, then they would be able to find valuable data from other researchers that they could use in their work. Also, they need to easily access compute resources and programs in order to process that data. If this were possible, science would become Open; that is accessible to all, without the barriers which can block scientific advances, and realise its full potential.

However, there are significant obstacles in the way of open science. Existing computing support has been largely developed within a specific discipline or within particular projects, resulting in a fragmentation of resources and duplication of tools. The tools and data supplied by these projects are typically incompatible, not using common standards to make them work together. And perhaps most importantly, researchers do not always see the benefits from making their data, and thus their science, more openly available.

Recognising the potential impact of open science, the European Commission has proposed that there should be a *European Open Science Cloud (EOSC)*, as a common platform to allow researchers to share and access their data, to access computing resources, and to collaborate together. This European Open Science Cloud will offer 1.7 million European researchers and 70 million professionals in science and technology a virtual environment with open and seamless services for storage, management, analysis and re-use of research data. This will work across borders and scientific disciplines by federating existing scientific data infrastructures, today scattered across disciplines and Member States. This is a major initiative which could transform the way we do science; however, working out how the EOSC should work is not straightforward!

STFC's Scientific Computing Department is leading the first project within the European Open Science Cloud programme. This project, called the EOSC Pilot[1] is a 10M€ project which started in January 2017, and will last for two years. It brings together over 50 partners from across Europe which represent a wide range of disciplines and computing providers, from particle physics and materials science, to biology, linguistics and archaeology. The project is focussed on consulting the research community as widely as possible to form a "bottom-up" view of what the EOSC should be, how it could benefit scientific progress, and how it should be established and run. Its main activities are to:

- Propose a governance frameworks for the EOSC and contribute to the development of European open science policy and best practice;
- Develop an architectural framework for tools and services to work together in the EOSC and to recommend how they might interoperate together;

- Consider how skills and capabilities can be developed so that the scientific community can take the best advantage of the EOSC;
- Develop a number of demonstrators as high-profile pilots that integrate services and infrastructures to show how the EOSC benefits science in a number of domains; and
- Engage with a broad range of stakeholders, crossing borders and communities, to build the consensus required for adoption of an open approach to scientific research.

Many of STFC science programmes and partners are now involved in demonstrator activities, including: CERN in the particle physics community; DESY, EU-XFEL and ESS within the analytic facilities; the Cryo-Electron Microscopy community; and the LOFAR radio astronomy project. The EOSC represents an exciting vision of how science could be transformed by open collaboration and access to computing resources. SCD is playing a leading role in this effort, to further its strategic aim of ensuring that the UK's e-infrastructure supports leading international research and delivers the data capabilities essential for academia, industry and STFC's science.

Authors
B.Matthews, STFC Rutherford Appleton Laboratory

References

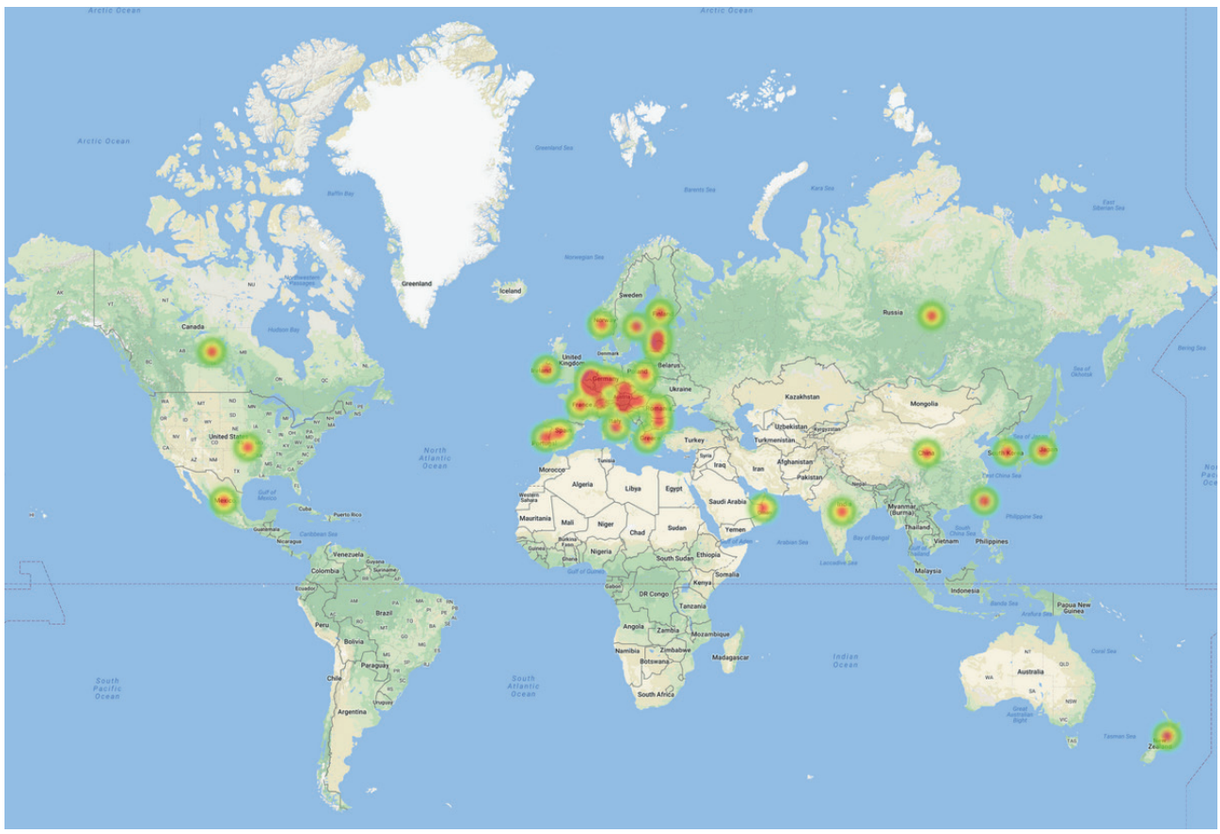
[1] https://www.eoscpilot.eu

Strategic Theme 4

In order to ensure that the UK's e-infrastructure supports the country's leading international research status and delivers the data capabilities essential for academia, industry and STFC's science program, SCD will collaborate with our Research Council UK partners and stakeholders, and with Innovate UK, and JISC, as part of the emerging UK Research and Innovation (UKRI), to shape and build an efficient and effective UK e-infrastructure for UK science and industry.

International Presence

SCD aims to form and strengthen partnerships with institutes around the world. During 2016-17 we presented, demonstrated and discussed our research and services in over 36 countries across the globe. This included participation in some major conferences and exhibitions. Here are just a few of these event highlights.



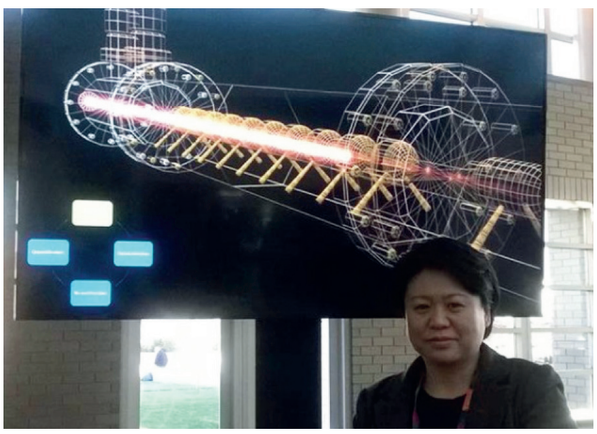
Super Computing Conference 2016

Staff from Scientific Computing and the Hartree Centre were in Salt Lake City in November for the world's largest supercomputing conference. Visitors to the STFC booth were able to find out more about SCD's expertise in computational science and professional large-scale scientific, data management and computing systems; and the Hartree Centre's work with industry and the research community to address real life challenges and to accelerate the adoption of high performance technologies.

At the conference, details of ULTRA, a new platform that enables scientists to have hundreds of digital replicas of a sample of interest, inside and out within tens of minutes, depicting layers and segments of the internal details at a resolution not achievable before was unveiled.

Our teams will once again be at Supercomputing 2017 in Denver to welcome current colleagues and partners and to develop new collaborations.

A team from Scientific Computing Department (SCD) and Hartree had a stand in the exhibition hall



Erica Yang explaining the concepts of ULTRA

SCD and Hartree at ISC High Performance 2017

The Hartree Centre and Scientific Computing Department exhibited at the ISC High Performance exhibition in Frankfurt, Germany. Europe's largest HPC (High Performance Computing) exhibition, returned to Frankfurt this year attracting a record 3,200 attendees from 60 countries internationally and bringing together industry and research professionals in HPC, networking and storage.

Our own experts were looking at product updates and new technologies from vendors at the conference, as well as exchanging knowledge and networking with like minded people from other organisations.

Alison Kennedy, Director of the Hartree Centre, spoke at a workshop on Factors affecting the Adoption of Machine Learning Technologies in Industrial Collaborations.

The Hartree Centre and Scientific Computing Department were pleased to have a strong presence at a stand throughout the conference exhibition with 147 other exhibitors. It was great to be able to meet with our existing partners and network with new connections in the HPC community, we can't wait to see you all at next year's event!

A team from Scientific Computing Department (SCD) and Hartree attending ISC High Performance 2017

Data-Intensive Science and Technologies

In September 2016 the Scientific Computing Department, which is based at the Harwell Campus in the UK, hosted a meeting of international experts to discuss data-intensive science and the new software technologies now being used to analyse scientific Big Data. Tony Hey, STFC's Chief Data Scientist welcomed colleagues from the USA and Europe, and gave a brief overview of the emerging 'Fourth Paradigm' of data-intensive scientific discovery. This was followed by some fascinating talks on data and compute-intensive science problems – ranging from how to estimate snow cover in the Hindu Kush mountains (which straddle the borders between Pakistan, Afghanistan, and India) using only satellite measurements, and how database technologies can be used to save and explore the results of numerical fluid dynamics simulations, to how revolutions in both the light source brilliance and detector technology were resulting in new data analysis and computational challenges.

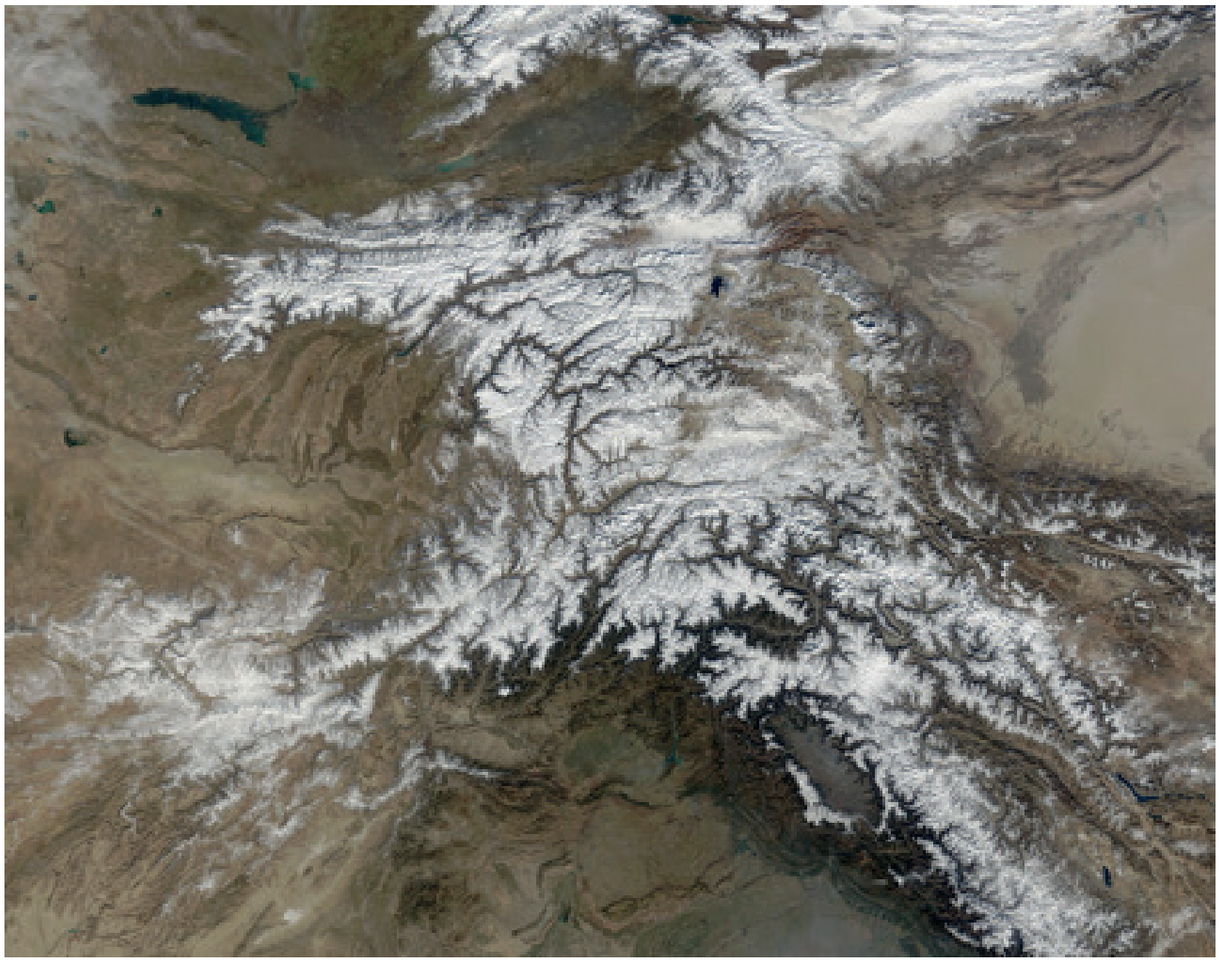


Image: Satellite image of the Hindu Kush
Credit: Jeff Schmaltz, Modis Land Rapid Response Team at NASA GSFC

Notes

Notes

Notes

Contacts and Further Information

Mr David Corney	Director of the Scientific Computing Department	david.corney@stfc.ac.uk
Dr Peter Oliver	Head of Operations	peter.oliver@stfc.ac.uk
Prof Tony Hey	Chief Data Scientist	tony.hey@stfc.ac.uk
Mrs Shirley Miller	Personal Assistant	shirley.miller@stfc.ac.uk
Mrs Jean Pearce	Personal Assistant	jean.pearce@stfc.ac.uk
Mr Damian Jones	CoSeC Project Support Manager	damian.jones@stfc.ac.uk
Mrs Karen McIntyre	Administrative Assistant	karen.mcIntyre@stfc.ac.uk
Miss Helen Walker	Administrative Assistant	h.walker@stfc.ac.uk
Ms Esme Williams	Administrative Assistant	esme.williams@stfc.ac.uk
Ms Clare Johnson	Administrative Assistant	clare.johnson@stfc.ac.uk

Division Heads

Dr Juan Bicarregui	Head of Data	juan.bicarregui@stfc.ac.uk
Mr Gordon Brown	Acting Head of Ada Lovelace	gordon.brown@stfc.ac.uk
Dr Andrew Sansum	Head of Systems	andrew.sansum@stfc.ac.uk
Dr Barbara Montanari	Acting Head of Computational Science and Engineering	barbara.montanari@stfc.ac.uk

Group Leaders

Mr Ian Collier	Distributed Computing Infrastructure	ian.collier@stfc.ac.uk
Prof Dave Emerson	Engineering and Environment	david.emerson@stfc.ac.uk
Mrs Debbie Franks	Library Services	debbie.franks@stfc.ac.uk
Mr Nick Hill	Research Infrastructure	nick.hill@stfc.ac.uk
Mrs Catherine Jones	Software Engineering Group	catherine.jones@stfc.ac.uk
Dr Brian Matthews	Data Science and Technology	brian.matthews@stfc.ac.uk
Dr Barbara Montanari	Theoretical and Computational Physics	barbara.montanari@stfc.ac.uk
Ms Alison Packer	Data Services	alison.packer@stfc.ac.uk
Dr Jennifer Scott	Computational Mathematics	jennifer.scott@stfc.ac.uk
Dr Ilian Todorov	Computational Chemistry	ilian.todorov@stfc.ac.uk
Dr Erica Yang	Visual Analytics and Imaging Systems	erica.yang@stfc.ac.uk
Dr Martyn Winn	Biology and Life Sciences	martyn.winn@stfc.ac.uk

Websites

Scientific Computing Department	www.stfc.ac.uk/scd
Science and Technology Facilities Council	www.stfc.ac.uk

Science and Technology Facilities Council
Rutherford Appleton Laboratory, Harwell Oxford, Didcot, Oxfordshire OX11 0QX, UK
T: +44 (0) 1235 445000 F: +44 (0) 1235 445808

Sci-Tech Daresbury, Daresbury, Warrington, Cheshire WA4 4AD, UK
T: +44 (0) 1925 603000 F: +44 (0) 1925 603100

www.stfc.ac.uk

Head office, Science and Technology Facilities Council, Polaris House, North Star Avenue, Swindon SN2 1SZ, UK
Establishments at: Rutherford Appleton Laboratory, Oxfordshire; Daresbury Laboratory, Cheshire:
UK Astronomy Centre, Edinburgh; Chilbolton Observatory, Hampshire: Boulby Underground Science Facility
Boulby Mine, Cleveland.

Science & Technology
Facilities Council